Information in Spoken Language

A quantitative approach

Rob van Son

Chair of Phonetic Sciences
ACLC
University of Amsterdam

LOT winterschool 2006

Introduction

Statistics is the bookkeeping of information

- Language is about communication
- Communication implies a message
- A message is only useful if it is "surprising" to some extend
- That is, the receiver must be uncertain about the content of the message
- Information and probability quantify uncertainty
- Information is the more fundamental concept

Introduction

Statistics is the bookkeeping of information

- Language is about communication
- Communication implies a message
- A message is only useful if it is "surprising" to some extend
- That is, the receiver must be uncertain about the content of the message
- Information and probability quantify uncertainty
- Information is the more fundamental concept

Introduction

Statistics is the bookkeeping of information

- Language is about communication
- Communication implies a message
- A message is only useful if it is "surprising" to some extend
- That is, the receiver must be uncertain about the content of the message
- Information and probability quantify uncertainty
- Information is the more fundamental concept

Introduction

Statistics is the bookkeeping of information

- Language is about communication
- Communication implies a message
- A message is only useful if it is "surprising" to some extend
- That is, the receiver must be uncertain about the content of the message
- Information and probability quantify uncertainty
- Information is the more fundamental concept

Introduction

Statistics is the bookkeeping of information

- Language is about communication
- Communication implies a message
- A message is only useful if it is “surprising” to some extend
- That is, the receiver must be uncertain about the content of the message
- Information and probability quantify uncertainty
- Information is the more fundamental concept

Introduction

Statistics is the bookkeeping of information

- Language is about communication
- Communication implies a message
- A message is only useful if it is "surprising" to some extend
- That is, the receiver must be uncertain about the content of the message
- Information and probability quantify uncertainty
- Information is the more fundamental concept

Introduction

Probability is

- A measure of the frequency of outcomes
- A measure of chance given what is known
- A number between 0 and 1 (inclusive)
- A measure of our knowledge (or ignorance)
- Boring?

[Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas] [Schneider(1999)] [MacKay(2003)]

Introduction

Probability is

- A measure of the frequency of outcomes
- A measure of chance given what is known
- A number between 0 and 1 (inclusive)
- A measure of our knowledge (or ignorance)
- Boring?

[Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas] [Schneider(1999)] [MacKay(2003)]

Introduction

Probability is

- A measure of the frequency of outcomes
- A measure of chance given what is known
- A number between 0 and 1 (inclusive)
- A measure of our knowledge (or ignorance)
- Boring?

[Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas] [Schneider(1999)] [MacKay(2003)]

Introduction

Probability is

- A measure of the frequency of outcomes
- A measure of chance given what is known
- A number between 0 and 1 (inclusive)
- A measure of our knowledge (or ignorance)
- Boring?

[Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas] [Schneider(1999)] [MacKay(2003)]

Introduction

Probability is

- A measure of the frequency of outcomes
- A measure of chance given what is known
- A number between 0 and 1 (inclusive)
- A measure of our knowledge (or ignorance)
- Boring?

[Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas] [Schneider(1999)] [MacKay(2003)]

Introduction: Axioms

Probability: if $E_1, \ldots, E_n$ are possible outcomes of an observation, then $P(E_i)$ is the probability of outcome E_i iff

1. $0 \leq P(E_i) \leq 1$
2. $P(E_1 \vee \cdots \vee E_i \vee \cdots \vee E_n) = 1$
3. Additivity: $P(E_1 \vee E_2) = P(E_1) + P(E_2)$ where E_1 and E_2 are mutually exclusive.
4. Countable additivity: $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ for $n = 1, 2, \ldots, N$ where $E_1, E_2, \ldots$ are mutually exclusive.

Introduction: Axioms

Probability: if $E_1, \ldots, E_n$ are possible outcomes of an observation, then $P(E_i)$ is the probability of outcome E_i iff

1. $0 \leq P(E_i) \leq 1$
2. $P(E_1 \vee \cdots \vee E_i \vee \cdots \vee E_n) = 1$
3. Additivity: $P(E_1 \vee E_2) = P(E_1) + P(E_2)$ where E_1 and E_2 are mutually exclusive.
4. Countable additivity: $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ for $n = 1, 2, \ldots, N$ where $E_1, E_2, \ldots$ are mutually exclusive.

Introduction: Axioms

Probability: if $E_1, \ldots, E_n$ are possible outcomes of an observation, then $P(E_i)$ is the probability of outcome E_i iff

1. $0 \leq P(E_i) \leq 1$
2. $P(E_1 \vee \cdots \vee E_i \vee \cdots \vee E_n) = 1$
3. Additivity: $P(E_1 \vee E_2) = P(E_1) + P(E_2)$ where E_1 and E_2 are mutually exclusive.
4. Countable additivity: $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ for $n = 1, 2, \ldots, N$ where $E_1, E_2, \ldots$ are mutually exclusive.

Introduction: Axioms

Probability: if $E_1, \dots, E_n$ are possible outcomes of an observation, then $P(E_i)$ is the probability of outcome E_i iff

1. $0 \leq P(E_i) \leq 1$
2. $P(E_1 \vee \dots \vee E_i \vee \dots \vee E_n) = 1$
3. Additivity: $P(E_1 \vee E_2) = P(E_1) + P(E_2)$ where E_1 and E_2 are mutually exclusive.
4. Countable additivity: $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ for $n = 1, 2, \dots, N$ where $E_1, E_2, \dots$ are mutually exclusive.

Introduction: Conventions

Take three observables: *Color*, *Flower*, and *Place*

The following conventions will be used:

1. $P(C = Red)$: the probability of seeing a *Red flower*
2. $P(C = Red, F = Rose)$: the probability of seeing a *Red Rose*
3. $P(C = Red|F = Rose)$: the probability of seeing a *Red Flower*, given that the flower is a *Rose*
4. $P(C = Red, F = Rose|P = Flower\ Shop) \leq 1$
5. $P(Red \vee Blue) = P(Red) + P(Blue)$
   Basic sum rule for probabilities
6. $P(C, F|P) = P(C|F, P) \cdot P(F|P) = P(F|C, P) \cdot P(C|P)$
   The basic product rule for probabilities

Introduction: Conventions

Take three observables: *Color*, *Flower*, and *Place*

The following conventions will be used:

1. $P(C = Red)$: the probability of seeing a *Red flower*
2. $P(C = Red, F = Rose)$: the probability of seeing a *Red Rose*
3. $P(C = Red|F = Rose)$: the probability of seeing a *Red Flower*, given that the flower is a *Rose*
4. $P(C = Red, F = Rose|P = Flower\ Shop) \leq 1$
5. $P(Red \vee Blue) = P(Red) + P(Blue)$
   Basic sum rule for probabilities
6. $P(C, F|P) = P(C|F, P) \cdot P(F|P) = P(F|C, P) \cdot P(C|P)$
   The basic product rule for probabilities

Introduction: Conventions

Take three observables: *Color*, *Flower*, and *Place*

The following conventions will be used:

1. $P(C = Red)$: the probability of seeing a *Red flower*
2. $P(C = Red, F = Rose)$: the probability of seeing a *Red Rose*
3. $P(C = Red|F = Rose)$: the probability of seeing a *Red Flower*, given that the flower is a *Rose*
4. $P(C = Red, F = Rose|P = Flower\ Shop) \leq 1$
5. $P(Red \vee Blue) = P(Red) + P(Blue)$
   Basic sum rule for probabilities
6. $P(C, F|P) = P(C|F, P) \cdot P(F|P) = P(F|C, P) \cdot P(C|P)$
   The basic product rule for probabilities

Introduction: Conventions

Take three observables: *Color*, *Flower*, and *Place*

The following conventions will be used:

1. $P(C = Red)$: the probability of seeing a *Red flower*
2. $P(C = Red, F = Rose)$: the probability of seeing a *Red Rose*
3. $P(C = Red|F = Rose)$: the probability of seeing a *Red Flower*, given that the flower is a *Rose*
4. $P(C = Red, F = Rose|P = Flower\ Shop) \leq 1$
5. $P(Red \vee Blue) = P(Red) + P(Blue)$
   Basic sum rule for probabilities
6. $P(C, F|P) = P(C|F, P) \cdot P(F|P) = P(F|C, P) \cdot P(C|P)$
   The basic product rule for probabilities

Introduction: Conventions

Take three observables: *Color*, *Flower*, and *Place*

The following conventions will be used:

1. $P(C = Red)$: the probability of seeing a *Red flower*
2. $P(C = Red, F = Rose)$: the probability of seeing a *Red Rose*
3. $P(C = Red|F = Rose)$: the probability of seeing a *Red Flower*, given that the flower is a *Rose*
4. $P(C = Red, F = Rose|P = Flower\ Shop) \leq 1$
5. $P(\color{red}{Red} \vee \color{blue}{Blue}) = P(\color{red}{Red}) + P(\color{blue}{Blue})$
   Basic sum rule for probabilities
6. $P(C, F|P) = P(C|F, P) \cdot P(F|P) = P(F|C, P) \cdot P(C|P)$
   The basic product rule for probabilities

Introduction: Conventions

Take three observables: *<u>C</u>olor*, *<u>F</u>lower*, and *<u>P</u>lace*

The following conventions will be used:

1. $P(C = Red)$: the probability of seeing a *Red flower*
2. $P(C = Red, F = Rose)$: the probability of seeing a *Red Rose*
3. $P(C = Red|F = Rose)$: the probability of seeing a *Red Flower*, given that the flower is a *Rose*
4. $P(C = Red, F = Rose|P = Flower\ Shop) \leq 1$
5. $P(Red \vee Blue) = P(Red) + P(Blue)$
   Basic sum rule for probabilities
6. $P(C, F|P) = P(C|F, P) \cdot P(F|P) = P(F|C, P) \cdot P(C|P)$
   The basic product rule for probabilities

Probability distributions

Useful distributions, called Probability Density Functions (pdf)

- Uniform distribution, discrete and uniform
- Poisson distribution
- Normal (Gaussian) distribution
- Zipf distribution
- Mean value, μ, is called Expected value
  $\mu = E[x] = \int_{-\infty}^{+\infty} x \cdot P(x) dx$
- Distribution width is called *Standard Deviation* which is defined as $\sigma = \sqrt{E[(x - E(x))^2]}$

Probability distributions

Useful distributions, called Probability Density Functions (pdf)

- Uniform distribution, discrete and uniform
- Poisson distribution
- Normal (Gaussian) distribution
- Zipf distribution
- Mean value, μ, is called Expected value
  $\mu = E[x] = \int_{-\infty}^{+\infty} x \cdot P(x)dx$
- Distribution width is called *Standard Deviation* which is defined as $\sigma = \sqrt{E[(x - E(x))^2]}$

Probability distributions

Useful distributions, called Probability Density Functions (pdf)

- Uniform distribution, discrete and uniform
- Poisson distribution
- Normal (Gaussian) distribution
- Zipf distribution
- Mean value, μ, is called Expected value $\mu = E[x] = \int_{-\infty}^{+\infty} x \cdot P(x)dx$
- Distribution width is called *Standard Deviation* which is defined as $\sigma = \sqrt{E[(x - E(x))^2]}$

Probability distributions

Useful distributions, called Probability Density Functions (pdf)

- Uniform distribution, discrete and uniform
- Poisson distribution
- Normal (Gaussian) distribution
- Zipf distribution
- Mean value, μ, is called Expected value $\mu = E[x] = \int_{-\infty}^{+\infty} x \cdot P(x)dx$
- Distribution width is called *Standard Deviation* which is defined as $\sigma = \sqrt{E[(x - E(x))^2]}$

Probability distributions

Useful distributions, called Probability Density Functions (pdf)

- Uniform distribution, discrete and uniform
- Poisson distribution
- Normal (Gaussian) distribution
- Zipf distribution
- Mean value, μ, is called Expected value
  $\mu = E[x] = \int_{-\infty}^{+\infty} x \cdot P(x) dx$
- Distribution width is called *Standard Deviation* which is defined as $\sigma = \sqrt{E[(x - E(x))^2]}$

Probability distributions

Useful distributions, called Probability Density Functions (pdf)

- Uniform distribution, discrete and uniform
- Poisson distribution
- Normal (Gaussian) distribution
- Zipf distribution
- Mean value, μ, is called Expected value
  $\mu = E[x] = \int_{-\infty}^{+\infty} x \cdot P(x)dx$
- Distribution width is called *Standard Deviation* which is defined as $\sigma = \sqrt{E[(x - E(x))^2]}$

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N}\sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29,
http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N} \sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29,
http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N} \sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29, http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N} \sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29, http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N}\sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29,
http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N} \sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29, http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Discrete

N, equally probable and equally spaced values $\{E_1, \ldots, E_n\}$ (possibly if $N \to \infty$)

- Each category, E_i, has the same probability
- $P(E_i) = 1/N$
- Example: Dice $\{1, \ldots, 6\}$ and coins $\{Head, Tail\}$
- Most basic distribution
- Default if only the number of values is known
- Mean $\mu = \frac{1}{N} \sum_{i=1}^{N} E_i = \frac{1}{2}(E_1 + E_N)$
- Variance $\sigma^2 = \frac{1}{12}(E_N - E_1)^2$

http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29, http://gwydir.demon.co.uk/jo/probability/diceinfo.htm

Probability distributions: Uniform Continuous

Equally probable values in interval $[a, b]$

- $Pdf(x) = \dfrac{1}{b - a}$
- Most basic distribution (continuous case)
- Default if only the range is known
- Mean $\mu = \dfrac{a + b}{2}$
- Variance $\sigma^2 = \dfrac{(b - a)^2}{12}$

http://en.wikipedia.org/wiki/Uniform_distribution_%28continuous%29

Probability distributions: Uniform Continuous

Equally probable values in interval $[a, b]$

- $Pdf(x) = \dfrac{1}{b - a}$
- Most basic distribution (continuous case)
- Default if only the range is known
- Mean $\mu = \dfrac{a + b}{2}$
- Variance $\sigma^2 = \dfrac{(b - a)^2}{12}$

http://en.wikipedia.org/wiki/Uniform_distribution_%28continuous%29

Probability distributions: Uniform Continuous

Equally probable values in interval $[a, b]$

- $Pdf(x) = \dfrac{1}{b - a}$
- Most basic distribution (continuous case)
- Default if only the range is known
- Mean $\mu = \dfrac{a + b}{2}$
- Variance $\sigma^2 = \dfrac{(b - a)^2}{12}$

http://en.wikipedia.org/wiki/Uniform_distribution_%28continuous%29

Probability distributions: Uniform Continuous

Equally probable values in interval $[a, b]$

- $Pdf(x) = \dfrac{1}{b-a}$
- Most basic distribution (continuous case)
- Default if only the range is known
- Mean $\mu = \dfrac{a+b}{2}$
- Variance $\sigma^2 = \dfrac{(b-a)^2}{12}$

http://en.wikipedia.org/wiki/Uniform_distribution_%28continuous%29

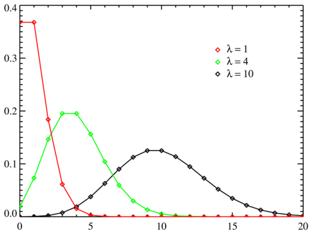
Probability distributions: Uniform Continuous

Equally probable values in interval $[a, b]$

- $Pdf(x) = \dfrac{1}{b-a}$
- Most basic distribution (continuous case)
- Default if only the range is known
- Mean $\mu = \dfrac{a+b}{2}$
- Variance $\sigma^2 = \dfrac{(b-a)^2}{12}$

`http://en.wikipedia.org/wiki/Uniform_distribution_%28continuous%29`

Probability distributions: Poisson



$$Pdf(k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

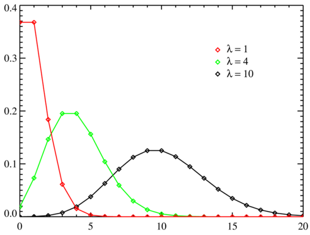
k: count, λ: rate

Rare events occuring with a fixed rate λ

- Mushrooms per meter of forest, typing errors per page, radio-active decay
- Average and variance are identical $\mu = \sigma^2 = \lambda$
- Default if only an average is known

http://en.wikipedia.org/wiki/Poisson_distribution

Probability distributions: Poisson



$$Pdf(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

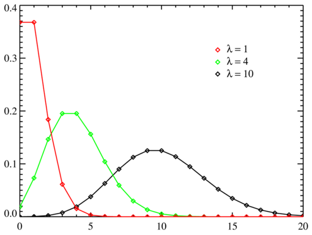
k: count, λ: rate

Rare events occuring with a fixed rate λ

- Mushrooms per meter of forest, typing errors per page, radio-active decay
- Average and variance are identical $\mu = \sigma^2 = \lambda$
- Default if only an average is known

http://en.wikipedia.org/wiki/Poisson_distribution

Probability distributions: Poisson



$$Pdf(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

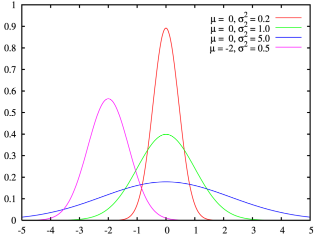
k: count, λ: rate

Rare events occuring with a fixed rate λ

- Mushrooms per meter of forest, typing errors per page, radio-active decay
- Average and variance are identical $\mu = \sigma^2 = \lambda$
- Default if only an average is known

http://en.wikipedia.org/wiki/Poisson_distribution

Probability distributions: Normal or Gaussian



$$Pdf(x;\mu,\sigma) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

x: observable

μ: Average

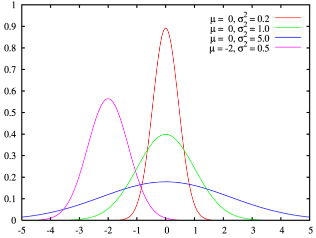
σ^2: variance

General measurements

- Many physical and physiological measurements, counting
- Default if both an average and a variance are known
- A sum of a large number of independent variables is approximately normal (under certain conditions)

http://en.wikipedia.org/wiki/Normal_distribution

Probability distributions: Normal or Gaussian



$$Pdf(x;\mu,\sigma) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

x: observable

μ: Average

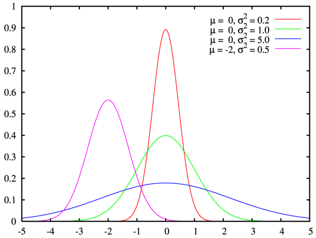
σ^2: variance

General measurements

- Many physical and physiological measurements, counting
- Default if both an average and a variance are known
- A sum of a large number of independent variables is approximately normal (under certain conditions)

http://en.wikipedia.org/wiki/Normal_distribution

Probability distributions: Normal or Gaussian



$$Pdf(x;\mu,\sigma) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

x: observable

μ: Average

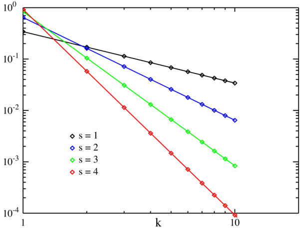
σ^2: variance

General measurements

- Many physical and physiological measurements, counting
- Default if both an average and a variance are known
- A sum of a large number of independent variables is approximately normal (under certain conditions)

http://en.wikipedia.org/wiki/Normal_distribution

Probability distributions: Zipf



$$Pdf(k; s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

k: rank; s: exponent

N: number of elements

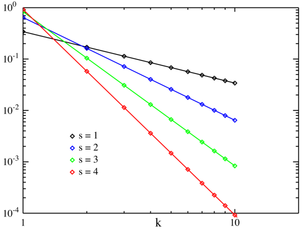
note logarithmic scales

Product of frequency and rank is constant: $f_i \approx C \cdot \frac{1}{r_i}$

- Word frequencies, city sizes, high incomes, earthquake sizes
- Default with power laws
- For word frequencies, $s \approx 1$

http://en.wikipedia.org/wiki/Zipf_distribution

Probability distributions: Zipf



$$Pdf(k; s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

k: rank; s: exponent

N: number of elements

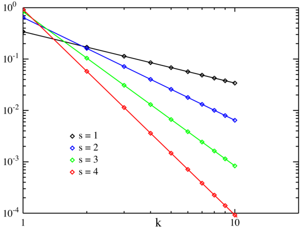
note logarithmic scales

Product of frequency and rank is constant: $f_i \approx C \cdot \frac{1}{r_i}$

- Word frequencies, city sizes, high incomes, earthquake sizes
- Default with power laws
- For word frequencies, $s \approx 1$

http://en.wikipedia.org/wiki/Zipf_distribution

Probability distributions: Zipf



$$Pdf(k; s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

k: rank; s: exponent

N: number of elements

note logarithmic scales

Product of frequency and rank is constant: $f_i \approx C \cdot \frac{1}{r_i}$

- Word frequencies, city sizes, high incomes, earthquake sizes
- Default with power laws
- For word frequencies, $s \approx 1$

http://en.wikipedia.org/wiki/Zipf_distribution

Bayesian probabilities

Incorporating knowledge

- Probability is predicting outcomes from knowledge
- Explicitely formulate knowledge as probabilities
- Formalize the probability of a hypothesis
- Destinguish a priori (knowledge) and a posteriori (observations) probabilities
- Determine the information content of a single observation

Bayesian probabilities

Incorporating knowledge

- Probability is predicting outcomes from knowledge
- Explicitely formulate knowledge as probabilities
- Formalize the probability of a hypothesis
- Destinguish a priori (knowledge) and a posteriori (observations) probabilities
- Determine the information content of a single observation

Bayesian probabilities

Incorporating knowledge

- Probability is predicting outcomes from knowledge
- Explicitely formulate knowledge as probabilities
- Formalize the probability of a hypothesis
- Destinguish a priori (knowledge) and a posteriori (observations) probabilities
- Determine the information content of a single observation

Bayesian probabilities

Incorporating knowledge

- Probability is predicting outcomes from knowledge
- Explicitely formulate knowledge as probabilities
- Formalize the *probability of a hypothesis*
- Destinguish a priori (knowledge) and a posteriori (observations) probabilities
- Determine the information content of a single observation

Bayesian probabilities

Incorporating knowledge

- Probability is predicting outcomes from knowledge
- Explicitely formulate knowledge as probabilities
- Formalize the *probability of a hypothesis*
- Destinguish *a priori* (knowledge) and *a posteriori* (observations) probabilities
- Determine the information content of a single observation

Bayesian probabilities

$$
\begin{aligned}
P(Data, Hypothesis) &= P(Hypothesis|Data) \cdot P(Data) \\
&= P(Data|Hypothesis) \cdot P(Hypothesis) \\
&\Leftrightarrow \\
P(Hypothesis|Data) &= \frac{P(Data|Hypothesis) \cdot P(Hypothesis)}{P(Data)}
\end{aligned}
$$

Express $P(Hypothesis|Data)$:

- As a function of the measurements
- And the a priori probability of the hypothesis
- Normalized by the a priori probability of the data
- The normalization probability can often be ignored, as it will be identical for all hypotheses

Bayesian probabilities

$$P(Data, Hypothesis) = P(Hypothesis|Data) \cdot P(Data)$$
$$= P(Data|Hypothesis) \cdot P(Hypothesis)$$
$$\Leftrightarrow$$
$$P(Hypothesis|Data) = \frac{\mathbf{P}(\mathbf{Data}|\mathbf{Hypothesis}) \cdot P(Hypothesis)}{P(Data)}$$

Express $P(Hypothesis|Data)$:

- As a function of the measurements
- And the a priori probability of the hypothesis
- Normalized by the a priori probability of the data
- The normalization probability can often be ignored, as it will be identical for all hypotheses

Bayesian probabilities

$$
\begin{aligned}
P(Data, Hypothesis) &= P(Hypothesis|Data) \cdot P(Data) \\
&= P(Data|Hypothesis) \cdot P(Hypothesis) \\
&\Leftrightarrow \\
P(Hypothesis|Data) &= \frac{\mathbf{P(Data|Hypothesis) \cdot P(Hypothesis)}}{P(Data)}
\end{aligned}
$$

Express $P(Hypothesis|Data)$:

- As a function of the measurements
- And the *a priori* probability of the hypothesis
- Normalized by the a priori probability of the data
- The normalization probability can often be ignored, as it will be identical for all hypotheses

Bayesian probabilities

$$
\begin{aligned}
P(Data, Hypothesis) &= P(Hypothesis|Data) \cdot P(Data) \\
&= P(Data|Hypothesis) \cdot P(Hypothesis) \\
&\Leftrightarrow \\
P(Hypothesis|Data) &= \frac{\mathbf{P(Data|Hypothesis) \cdot P(Hypothesis)}}{\mathbf{P(Data)}}
\end{aligned}
$$

Express $P(Hypothesis|Data)$:

- As a function of the measurements
- And the *a priori* probability of the hypothesis
- Normalized by the *a priori* probability of the data
- The normalization probability can often be ignored, as it will be identical for all hypotheses

Bayesian probabilities: Toy example

Where has Watson most likely been: *Market*, *Garden*, *Meadow*, *Park*?

- Watson caries a Yellow *Buttercup*
- He divides his walks equally along these "*places*" (uniform prior)
- Which is most likely, obtaining a Yellow *Buttercup* in a *Market*, a *Garden*, a *Meadow*, or a *Park*?
- In formula:

$$\underset{p}{argmax} P(p|Y,B) = \underset{p}{argmax} P(Y,B|p) \cdot P(p)$$

$$where\ \ p \in \{Market, Garden, Meadow, Park\}$$

Bayesian probabilities: Toy example

Where has Watson most likely been: *Market*, *Garden*, *Meadow*, *Park*?

- Watson caries a Yellow *Buttercup*
- He divides his walks equally along these "*places*" (uniform prior)
- Which is most likely, obtaining a Yellow *Buttercup* in a *Market*, a *Garden*, a *Meadow*, or a *Park*?
- In formula:

$$\underset{p}{argmax} P(p|Y,B) = \underset{p}{argmax} P(Y,B|p) \cdot P(p)$$

$$where\ \ p \in \{Market, Garden, Meadow, Park\}$$

Bayesian probabilities: Toy example

Where has Watson most likely been: *Market*, *Garden*, *Meadow*, *Park*?

- Watson caries a Yellow *Buttercup*
- He divides his walks equally along these "*places*" (uniform prior)
- Which is most likely, obtaining a Yellow *Buttercup* in a *Market*, a *Garden*, a *Meadow*, or a *Park*?
- In formula:

$$\underset{p}{argmax} P(p|Y,B) = \underset{p}{argmax} P(Y,B|p) \cdot P(p)$$

$$where \; p \in \{Market, Garden, Meadow, Park\}$$

Bayesian probabilities: Toy example

Where has Watson most likely been: *Market*, *Garden*, *Meadow*, *Park*?

- Watson caries a Yellow *Buttercup*
- He divides his walks equally along these "*places*" (uniform prior)
- Which is most likely, obtaining a Yellow *Buttercup* in a *Market*, a *Garden*, a *Meadow*, or a *Park*?
- In formula:

$$\underset{p}{argmax} P(p|Y, B) = \underset{p}{argmax} P(Y, B|p) \cdot P(p)$$
$$where\ p \in \{Market, Garden, Meadow, Park\}$$

Information and probabilities: Surprise!

Information is a quantification of *surprise*

- Information depends on probability p_i
- A more surprising observation, ie, a lower p_i, caries more information
- Information should be additive, two CD's can carry twice the information of one CD
- Define information in observation O_i with probability p_i as $h(p_i) = -\log_2 p_i$

Information and probabilities: Surprise!

Information is a quantification of *surprise*

- Information depends on probability p_i
- A more surprising observation, ie, a lower p_i, caries more information
- Information should be additive, two CD's can carry twice the information of one CD
- Define information in observation O_i with probability p_i as $h(p_i) = -\log_2 p_i$

Information and probabilities: Surprise!

Information is a quantification of *surprise*

- Information depends on probability p_i
- A more surprising observation, ie, a lower p_i, caries more information
- Information should be additive, two CD's can carry twice the information of one CD
- Define information in observation O_i with probability p_i as $h(p_i) = -\log_2 p_i$

Information and probabilities: Surprise!

Information is a quantification of *surprise*

- Information depends on probability p_i
- A more surprising observation, ie, a lower p_i, caries more information
- Information should be additive, two CD's can carry twice the information of one CD
- Define information in observation O_i with probability p_i as $h(p_i) = -\log_2 p_i$

Information and probabilities: Uncertainty

The uncertainty is the average information content and is called *Entropy*, $H(p_1, p_2, \ldots, p_n)$. *Entropy* should be:

- Independent of the labeling, ie, numbering, of p_i
- Decomposable, splitting a category in two gives: $H'(p'_1, p''_1, \ldots) = H(p_1, \ldots) + p_1 \cdot H(\frac{p'_1}{p_1}, \frac{p''_1}{p_1})$
- Continuous, a small change in the probabilities should result in a small change in *entropy*
- Monotonic, for a uniform distibution of n items, entropy increases monotonically with the number of categories $n \geq 1$
- $\Rightarrow H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Uncertainty

The uncertainty is the average information content and is called *Entropy*, $H(p_1, p_2, \ldots, p_n)$. *Entropy* should be:

- *Independent* of the labeling, ie, numbering, of p_i
- Decomposable, splitting a category in two gives: $H'(p_1', p_1'', \ldots) = H(p_1, \ldots) + p_1 \cdot H(\frac{p_1'}{p_1}, \frac{p_1''}{p_1})$
- Continuous, a small change in the probabilities should result in a small change in *entropy*
- Monotonic, for a uniform distibution of n items, entropy increases monotonically with the number of categories $n \geq 1$
- $\Rightarrow H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Uncertainty

The uncertainty is the average information content and is called *Entropy*, $H(p_1, p_2, \ldots, p_n)$. *Entropy* should be:

- *Independent* of the labeling, ie, numbering, of p_i
- *Decomposable*, splitting a category in two gives: $H'(p_1', p_1'', \ldots) = H(p_1, \ldots) + p_1 \cdot H(\frac{p_1'}{p_1}, \frac{p_1''}{p_1})$
- Continuous, a small change in the probabilities should result in a small change in *entropy*
- Monotonic, for a uniform distibution of n items, entropy increases monotonically with the number of categories $n \geq 1$
- $\Rightarrow H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Uncertainty

The uncertainty is the average information content and is called *Entropy*, $H(p_1, p_2, \ldots, p_n)$. *Entropy* should be:

- *Independent* of the labeling, ie, numbering, of p_i
- *Decomposable*, splitting a category in two gives: $H'(p'_1, p''_1, \ldots) = H(p_1, \ldots) + p_1 \cdot H(\frac{p'_1}{p_1}, \frac{p''_1}{p_1})$
- *Continuous*, a small change in the probabilities should result in a small change in *entropy*
- Monotonic, for a uniform distibution of n items, entropy increases monotonically with the number of categories $n \geq 1$
- $\Rightarrow H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Uncertainty

The uncertainty is the average information content and is called *Entropy*, $H(p_1, p_2, \ldots, p_n)$. *Entropy* should be:

- *Independent* of the labeling, ie, numbering, of p_i
- *Decomposable*, splitting a category in two gives: $H'(p'_1, p''_1, \ldots) = H(p_1, \ldots) + p_1 \cdot H(\frac{p'_1}{p_1}, \frac{p''_1}{p_1})$
- *Continuous*, a small change in the probabilities should result in a small change in *entropy*
- *Monotonic*, for a uniform distibution of n items, entropy increases monotonically with the number of categories $n \geq 1$
- $\Rightarrow H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities

Probability distributions have *entropies*: Examples

- Discrete Uniform distribution: $H(\frac{1}{N}) = \log_2(N)$
- Continuous Uniform distribution $[a, b]$: $H(\frac{1}{b-a}) = \log_2(b-a)$
- Poisson distribution: $H(k; \lambda) = \lambda[\frac{1}{\ln(2)} - \log_2(\lambda)] + e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k \log_2(k!)}{k!}$
- Normal (Gaussian) distribution: $H(x; \mu, \sigma) = \log_2(\sigma\sqrt{2\pi e})$
- Zipf distribution ($C_{N,s} = \sum_{k=1}^{N} \frac{1}{k^s}$): $H(k; s, N) = \frac{s}{C_{N,s}} \sum_{k=1}^{N} \frac{\log_2(k)}{k^s} + \log_2(C_{N,s})$

Information and probabilities

Probability distributions have *entropies*: Examples

- Discrete Uniform distribution: $H(\frac{1}{N}) = \log_2(N)$
- Continuous Uniform distribution $[a, b]$:
  $H(\frac{1}{b-a}) = \log_2(b-a)$
- Poisson distribution:
  $H(k;\lambda) = \lambda[\frac{1}{\ln(2)} - \log_2(\lambda)] + e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k \log_2(k!)}{k!}$
- Normal (Gaussian) distribution:
  $H(x;\mu,\sigma) = \log_2(\sigma\sqrt{2\pi e})$
- Zipf distribution ($C_{N,s} = \sum_{k=1}^{N}\frac{1}{k^s}$):
  $H(k;s,N) = \frac{s}{C_{N,s}}\sum_{k=1}^{N}\frac{\log_2(k)}{k^s} + \log_2(C_{N,s})$

Information and probabilities

Probability distributions have *entropies*: Examples

- Discrete Uniform distribution: $H(\frac{1}{N}) = \log_2(N)$
- Continuous Uniform distribution $[a, b]$:
  $H(\frac{1}{b-a}) = \log_2(b-a)$
- Poisson distribution:
  $H(k;\lambda) = \lambda[\frac{1}{\ln(2)} - \log_2(\lambda)] + e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k \log_2(k!)}{k!}$
- Normal (Gaussian) distribution:
  $H(x;\mu,\sigma) = \log_2(\sigma\sqrt{2\pi e})$
- Zipf distribution ($C_{N,s} = \sum_{k=1}^{N} \frac{1}{k^s}$):
  $H(k;s,N) = \frac{s}{C_{N,s}} \sum_{k=1}^{N} \frac{\log_2(k)}{k^s} + \log_2(C_{N,s})$

Information and probabilities

Probability distributions have *entropies*: Examples

- Discrete Uniform distribution: $H(\frac{1}{N}) = \log_2(N)$
- Continuous Uniform distribution $[a, b]$:
  $H(\frac{1}{b-a}) = \log_2(b-a)$
- Poisson distribution:
  $H(k;\lambda) = \lambda[\frac{1}{\ln(2)} - \log_2(\lambda)] + e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k\log_2(k!)}{k!}$
- Normal (Gaussian) distribution:
  $H(x;\mu,\sigma) = \log_2(\sigma\sqrt{2\pi e})$
- Zipf distribution ($C_{N,s} = \sum_{k=1}^{N}\frac{1}{k^s}$):
  $H(k;s,N) = \frac{s}{C_{N,s}}\sum_{k=1}^{N}\frac{\log_2(k)}{k^s} + \log_2(C_{N,s})$

Information and probabilities

Probability distributions have *entropies*: Examples

- Discrete Uniform distribution: $H(\frac{1}{N}) = \log_2(N)$
- Continuous Uniform distribution $[a, b]$: $H(\frac{1}{b-a}) = \log_2(b-a)$
- Poisson distribution: $H(k;\lambda) = \lambda[\frac{1}{\ln(2)} - \log_2(\lambda)] + e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k \log_2(k!)}{k!}$
- Normal (Gaussian) distribution: $H(x;\mu,\sigma) = \log_2(\sigma\sqrt{2\pi e})$
- Zipf distribution ($C_{N,s} = \sum_{k=1}^{N}\frac{1}{k^s}$): $H(k;s,N) = \frac{s}{C_{N,s}}\sum_{k=1}^{N}\frac{\log_2(k)}{k^s} + \log_2(C_{N,s})$

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{parameter}{X} \overset{channel}{\Rightarrow} \underset{observation}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{parameter}{X} \overset{channel}{\Rightarrow} \underset{observation}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{\textit{parameter}}{X} \overset{\textit{channel}}{\Rightarrow} \underset{\textit{observation}}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{parameter}{X} \overset{channel}{\Rightarrow} \underset{observation}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{parameter}{X} \overset{channel}{\Rightarrow} \underset{observation}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{parameter}{X} \overset{channel}{\Rightarrow} \underset{observation}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Information and probabilities: Measuring

Information is the reduction of uncertainty

$$\underset{parameter}{X} \overset{channel}{\Rightarrow} \underset{observation}{Y}$$

- Entropy in X before the observation: $H(X)$
- Entropy after the observation of value of Y: $H(X|Y)$
- Average information gained through observing Y: $I(X|Y) = H(X) - H(X|Y)$
- If there is *no* uncertainty left after observing Y, ie, $H(X|Y) = 0$: $I(X|Y) = H(X)$
- If X and Y are independent, ie, $H(X|Y) = H(X)$, then $I(X|Y) = 0$
- Always, $H(X|Y) \leq H(X) \Rightarrow I(X|Y) \leq H(X)$
- It is common to use $H(\cdot)$ as a synonym of $I(\cdot)$

See chapter 1 of [Bavaud et al.(2005)Bavaud, Chappelier, and Kohlas]

Relative entropy: $KL(p:q) = H(p,q) - H(p)$

$$KL(p:q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \quad \bigvee \quad \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx$$

discontinuous *continuous*

$H(p,q) = \sum_i p_i \log_2 q_i$: Cross Entropy

Kullback-Leibler distance

- A non-symmetric divergence: $KL(p:q) \neq KL(q:p)$
- Measures "distance" between prob. distributions
- Information gain between Prior and Posterior distribution
- Example: Word distributions as a distance between document types

http://en.wikipedia.org/wiki/Kullback-Leibler_distance

Relative entropy: $KL(p:q) = H(p,q) - H(p)$

$$KL(p:q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \quad \bigvee \quad \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx$$

discontinuous *continuous*

$H(p,q) = \sum_i p_i \log_2 q_i$: Cross Entropy

Kullback-Leibler distance

- A non-symmetric divergence: $KL(p:q) \neq KL(q:p)$
- Measures "distance" between prob. distributions
- Information gain between Prior and Posterior distribution
- Example: Word distributions as a distance between document types

http://en.wikipedia.org/wiki/Kullback-Leibler_distance

Relative entropy: $KL(p:q) = H(p,q) - H(p)$

$$KL(p:q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \quad \vee \quad \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx$$

discontinuous *continuous*

$H(p,q) = \sum_i p_i \log_2 q_i$: Cross Entropy

Kullback-Leibler distance

- A non-symmetric divergence: $KL(p:q) \neq KL(q:p)$
- Measures "distance" between prob. distributions
- Information gain between Prior and Posterior distribution
- Example: Word distributions as a distance between document types

http://en.wikipedia.org/wiki/Kullback-Leibler_distance

Relative entropy: $KL(p:q) = H(p,q) - H(p)$

$$KL(p:q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \quad \bigvee \quad \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx$$

discontinuous *continuous*

$H(p,q) = \sum_i p_i \log_2 q_i$: Cross Entropy

Kullback-Leibler distance

- A non-symmetric divergence: $KL(p:q) \neq KL(q:p)$
- Measures "distance" between prob. distributions
- Information gain between *Prior* and *Posterior* distribution
- Example: Word distributions as a distance between document types

http://en.wikipedia.org/wiki/Kullback-Leibler_distance

Compression: Minimum size

Entropy, $H(A)$ can be understood as the minimal number of bits needed to fully *specify* A given a known production process

- In an unknown process, $K(A)$ replaces $H(A)$ as the information content
- $K(A)$: Minimum number of bits to reconstruct A
- $K(A)$ is the theoretical lower limit of compression size $C(A)$
- Practical (lossless) compression packages, $C(A)$, eg, ZIP, GZIP, BZIP2 etc. never reach this limit

$K(A)$ is called the Kolmogorov complexity [Vitanyi(2005)][Chater and Vitanyi(2001)]

Compression: Minimum size

Entropy, $H(A)$ can be understood as the minimal number of bits needed to fully *specify A given a known production process*

- In an *unknown process*, $K(A)$ replaces $H(A)$ as the information content
- $K(A)$: Minimum number of bits to reconstruct A
- $K(A)$ is the theoretical lower limit of compression size $C(A)$
- Practical (lossless) compression packages, $C(A)$, eg, ZIP, GZIP, BZIP2 etc. never reach this limit

$K(A)$ is called the Kolmogorov complexity [Vitanyi(2005)][Chater and Vitanyi(2001)]

Compression: Minimum size

Entropy, $H(A)$ can be understood as the minimal number of bits needed to fully *specify A given a known production process*

- In an *unknown process*, $K(A)$ replaces $H(A)$ as the information content
- $K(A)$: Minimum number of bits to *reconstruct* A
- $K(A)$ is the theoretical lower limit of compression size $C(A)$
- Practical (lossless) compression packages, $C(A)$, eg, ZIP, GZIP, BZIP2 etc. never reach this limit

$K(A)$ is called the Kolmogorov complexity [Vitanyi(2005)][Chater and Vitanyi(2001)]

Compression: Minimum size

Entropy, $H(A)$ can be understood as the minimal number of bits needed to fully *specify A given a known production process*

- In an *unknown process*, $K(A)$ replaces $H(A)$ as the information content
- $K(A)$: Minimum number of bits to *reconstruct* A
- $K(A)$ is the theoretical lower limit of *compression size* $C(A)$
- Practical (lossless) compression packages, $C(A)$, eg, ZIP, GZIP, BZIP2 etc. never reach this limit

$K(A)$ is called the Kolmogorov complexity [Vitanyi(2005)][Chater and Vitanyi(2001)]

Compression: Similarity metric

$$
\begin{aligned}
NCD(A,B) &= \frac{min\{C(A|B), C(B|A)\}}{max\{C(A), C(B)\}} \\
&= \frac{C(AB) - min\{C(A), C(B)\}}{max\{C(A), C(B)\}}
\end{aligned}
$$

NCD: Normalized Compression Distance

Similarity by compression

- Always $C(AB) \leq C(A) + C(B)$ $(+constant)$
- Estimate entropy by suitable "long range" compression
- $K(text) \leq C(text)$ in bits

`http://www.complearn.org/ncd.html`

[Chen et al.(2004)Chen, Li, Ma, and Vitányi][Vitanyi(2005)][Chater and Vitanyi(2001)]

Compression: Similarity metric

$$
\begin{aligned}
NCD(A,B) &= \frac{min\{C(A|B), C(B|A)\}}{max\{C(A), C(B)\}} \\
&= \frac{C(AB) - min\{C(A), C(B)\}}{max\{C(A), C(B)\}}
\end{aligned}
$$

NCD: Normalized Compression Distance

Similarity by compression

- Always $C(AB) \leq C(A) + C(B)$ $(+constant)$
- Estimate entropy by suitable "long range" compression
- $K(text) \leq C(text)$ in bits

`http://www.complearn.org/ncd.html`

[Chen et al.(2004)Chen, Li, Ma, and Vitányi][Vitanyi(2005)][Chater and Vitanyi(2001)]

Compression: Similarity metric

$$
\begin{aligned}
NCD(A,B) &= \frac{min\{C(A|B), C(B|A)\}}{max\{C(A), C(B)\}} \\
&= \frac{C(AB) - min\{C(A), C(B)\}}{max\{C(A), C(B)\}}
\end{aligned}
$$

NCD: Normalized Compression Distance

Similarity by compression

- Always $C(AB) \leq C(A) + C(B)$ $(+constant)$
- Estimate entropy by suitable "long range" compression
- $K(text) \leq C(text)$ in bits

`http://www.complearn.org/ncd.html`

[Chen et al.(2004)Chen, Li, Ma, and Vitányi][Vitanyi(2005)][Chater and Vitanyi(2001)]

Markov Chains

Words and letters never follow each other at random

- The simplest language "model" predicts the next word based on the previous word
- Markov chain: $P(w_{i+1}|w_i) = \dfrac{P(w_{i+1}, w_i)}{P(w_i)}$
- Can be extended to more words
- Large amounts of text are needed to determine $P(w_{i+1}, w_i)$ reliably
- Example Markov text:
  Step which one could go be grabbed. People to Do that my the former Netscape brand's fortunes that means indent command to The user visible displays a.

http://en.wikipedia.org/wiki/Markov_chain

Generate texts: http://www.jwz.org/dadadodo/

Markov Chains

Words and letters never follow each other at random

- The simplest language "model" predicts the next word based on the previous word
- Markov chain: $P(w_{i+1}|w_i) = \dfrac{P(w_{i+1}, w_i)}{P(w_i)}$
- Can be extended to more words
- Large amounts of text are needed to determine $P(w_{i+1}, w_i)$ reliably
- Example Markov text:
  Step which one could go be grabbed. People to Do that my the former Netscape brand's fortunes that means indent command to The user visible displays a.

http://en.wikipedia.org/wiki/Markov_chain

Generate texts: http://www.jwz.org/dadadodo/

Markov Chains

Words and letters never follow each other at random

- The simplest language "model" predicts the next word based on the previous word
- Markov chain: $P(w_{i+1}|w_i) = \dfrac{P(w_{i+1}, w_i)}{P(w_i)}$
- Can be extended to more words
- Large amounts of text are needed to determine $P(w_{i+1}, w_i)$ reliably
- Example Markov text:
  Step which one could go be grabbed. People to Do that my the former Netscape brand's fortunes that means indent command to The user visible displays a.

http://en.wikipedia.org/wiki/Markov_chain

Generate texts: http://www.jwz.org/dadadodo/

Markov Chains

Words and letters never follow each other at random

- The simplest language "model" predicts the next word based on the previous word
- Markov chain: $P(w_{i+1}|w_i) = \dfrac{P(w_{i+1}, w_i)}{P(w_i)}$
- Can be extended to more words
- Large amounts of text are needed to determine $P(w_{i+1}, w_i)$ reliably
- Example Markov text:
  Step which one could go be grabbed. People to Do that my the former Netscape brand's fortunes that means indent command to The user visible displays a.

http://en.wikipedia.org/wiki/Markov_chain

Generate texts: http://www.jwz.org/dadadodo/

Markov Chains

Words and letters never follow each other at random

- The simplest language “model” predicts the next word based on the previous word
- Markov chain: $P(w_{i+1}|w_i) = \dfrac{P(w_{i+1}, w_i)}{P(w_i)}$
- Can be extended to more words
- Large amounts of text are needed to determine $P(w_{i+1}, w_i)$ reliably
- Example Markov text:
  Step which one could go be grabbed. People to Do that my the former Netscape brand's fortunes that means indent command to The user visible displays a.

http://en.wikipedia.org/wiki/Markov_chain

Generate texts: http://www.jwz.org/dadadodo/

Markov Chains: Language models

With Markov chains, or N-grams, the probability of a sequence can be calculated

- What is the probability of encounting a sentence $(w_1, \ldots, w_n)$?
- A human style language model is not known
- Use *N-gram* Markov chains
- $P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, wi - 1)$ (exact)
- $P(w_1, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1})$ (*N-gram* approximation)

Markov Chains: Language models

With Markov chains, or N-grams, the probability of a sequence can be calculated

- What is the probability of encounting a sentence $(w_1, \ldots, w_n)$?
- A human style language model is not known
- Use *N-gram* Markov chains
- $P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, wi - 1)$ (exact)
- $P(w_1, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1})$ (*N-gram* approximation)

Markov Chains: Language models

With Markov chains, or N-grams, the probability of a sequence can be calculated

- What is the probability of encounting a sentence $(w_1, \ldots, w_n)$?
- A human style language model is not known
- Use *N-gram* Markov chains
- $P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, wi - 1)$ (exact)
- $P(w_1, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1})$ (*N-gram* approximation)

Markov Chains: Language models

With Markov chains, or N-grams, the probability of a sequence can be calculated

- What is the probability of encounting a sentence $(w_1, \ldots, w_n)$?
- A human style language model is not known
- Use *N-gram* Markov chains
- $P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, wi - 1)$ (exact)
- $P(w_1, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1})$ (*N-gram* approximation)

Markov Chains: Language models

With Markov chains, or N-grams, the probability of a sequence can be calculated

- What is the probability of encounting a sentence $(w_1, \ldots, w_n)$?
- A human style language model is not known
- Use *N-gram* Markov chains
- $P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, wi - 1)$ (exact)
- $P(w_1, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \ldots, w_{i-1})$ (*N-gram* approximation)

Markov Chains: Perplexity

$$P_X(Model) = 2^{H_X(w_i|W_{1\ldots i-1})}$$
$$H_X(w_i|W_{1\ldots i-1}) = -\sum_{\{W\}} P_{observed}(w_i|\ldots)\log P_{model}(w_i|\ldots)$$

$H_X(\cdot)$: Cross Entropy

Perplexity: "average" number of choices for the next word

- Matches observed with modelled word order
- A better language model has a lower perplexity
- For an *N-gram* Markov chain the perplexity is well defined
- Using the model entropy io. the cross entropy estimates the quality of the model on the training corpus

Markov Chains: Perplexity

$$
\begin{aligned}
P_X(Model) &= 2^{H_X(w_i|W_{1\ldots i-1})} \\
H_X(w_i|W_{1\ldots i-1}) &= -\sum_{\{W\}} P_{observed}(w_i|\ldots)\log P_{model}(w_i|\ldots)
\end{aligned}
$$

$H_X(\cdot)$: Cross Entropy

Perplexity: "average" number of choices for the next word

- Matches observed with modelled word order
- A better language model has a lower perplexity
- For an *N-gram* Markov chain the perplexity is well defined
- Using the model entropy io. the cross entropy estimates the quality of the model on the training corpus

Markov Chains: Perplexity

$$P_X(Model) = 2^{H_X(w_i|W_{1\ldots i-1})}$$
$$H_X(w_i|W_{1\ldots i-1}) = -\sum_{\{W\}} P_{observed}(w_i|\ldots)\log P_{model}(w_i|\ldots)$$

$H_X(\cdot)$: Cross Entropy

Perplexity: "average" number of choices for the next word

- Matches observed with modelled word order
- A better language model has a lower perplexity
- For an *N-gram* Markov chain the perplexity is well defined
- Using the model entropy io. the cross entropy estimates the quality of the model on the training corpus

Markov Chains: Perplexity

$$
\begin{aligned}
P_X(Model) &= 2^{H_X(w_i|W_{1\ldots i-1})} \\
H_X(w_i|W_{1\ldots i-1}) &= -\sum_{\{W\}} P_{observed}(w_i|\ldots)\log P_{model}(w_i|\ldots)
\end{aligned}
$$

$H_X(\cdot)$: Cross Entropy

Perplexity: "average" number of choices for the next word

- Matches observed with modelled word order
- A better language model has a lower perplexity
- For an *N-gram* Markov chain the perplexity is well defined
- Using the model entropy io. the cross entropy estimates the quality of the model on the training corpus

Maximum Entropy

$$
\begin{aligned}
findp^* &= \underset{p \in C}{argmax} H(p) \\
&= \underset{p \in C}{argmax} \left(-\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \right)
\end{aligned}
$$

Which model, p^*, fits my data best and by what criterium?

- Quantify all constraints (knowledge) and determine the set of possible distributions $p \in C$
- Determine the average entropy, $H(y|x)$, over the observed (measured) probabilities $\tilde{p}(x)$
- The best distribution, p^* has the highest entropy

[Berger()][Berger(1996)] [Berger et al.(1996)Berger, della Pietra, and della Pietra] [Maxent()]

[Roni Rosenfeld(1996)]

Maximum Entropy

$$
\begin{aligned}
findp^* &= \underset{p \in C}{argmax} H(p) \\
&= \underset{p \in C}{argmax} \left(-\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \right)
\end{aligned}
$$

Which model, p^*, fits my data best and by what criterium?

- Quantify all constraints (knowledge) and determine the set of possible distributions $p \in C$
- Determine the average entropy, $H(y|x)$, over the observed (measured) probabilities $\tilde{p}(x)$
- The best distribution, p^* has the highest entropy

[Berger()][Berger(1996)] [Berger et al.(1996)Berger, della Pietra, and della Pietra] [Maxent()]

[Roni Rosenfeld(1996)]

Maximum Entropy

$$\begin{aligned} find p^* &= \underset{p \in C}{argmax} H(p) \\ &= \underset{p \in C}{argmax} \left(-\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \right) \end{aligned}$$

Which model, p^*, fits my data best and by what criterium?

- Quantify all constraints (knowledge) and determine the set of possible distributions $p \in C$
- Determine the average entropy, $H(y|x)$, over the observed (measured) probabilities $\tilde{p}(x)$
- The best distribution, p^* has the highest entropy

[Berger()][Berger(1996)] [Berger et al.(1996)Berger, della Pietra, and della Pietra] [Maxent()]

[Roni Rosenfeld(1996)]

Maximum Entropy: Kangaroo example

$\frac{1}{3}$ of all kangaroos have blue eyes and $\frac{1}{3}$ are left handed

blue eyed	Left Handed true	Left Handed false	tot
true	x	$\frac{1}{3} - x$	$\frac{1}{3}$
false	$\frac{1}{3} - x$	$\frac{1}{3} + x$	$\frac{2}{3}$
tot	$\frac{1}{3}$	$\frac{2}{3}$	1

How many are both blue eyed and left handed?

- All $0 \leq x \leq \frac{1}{3}$ are possible
- $H(x = \frac{1}{9}) \approx 1.84$ has maximum entropy
- $x = \frac{1}{9}$ is the only solution with uncorrelated eye color and handedness

Bibliography I

Abelard.
cause, chance and Bayesian statistics a briefing document.
Web.
URL http://www.abelard.org/briefings/bayes.htm.

F. Bavaud, J.-C. Chappelier, and J. Kohlas.
An Introduction to Information Theory and Applications.
UniFr course, 3 September 2005.
URL http://diuf.unifr.ch/tcs/courses/it04-05/script/information-theory.pdf.

Adam L. Berger.
MaxEnt and Exponential Models.
Website.
URL http://www.cs.cmu.edu/~aberger/maxent.html.

Adam L. Berger.
A Brief Maxent Tutorial.
Web, 5 July 1996.
URL http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html.

Adam L. Berger, Stephen A. della Pietra, and Vincent J. della Pietra.
A maximum entropy approach to natural language processing.
Computational Linguistics, 1:22–1, March 1996.
URL http://www.cs.cmu.edu/afs/cs/user/aberger/www/ps/compling.ps.

Bibliography II

Tom Carter.
An introduction to information theory and entropy.
Web presentation, June 2005.
URL http://astarte.csustan.edu/~tom/SFI-CSSS/2005/info-lec.pdf.

Nick Chater and Paul Vitanyi.
The generalized universal law of generalization.
Arxiv.org Computer Science, 29 January 2001.
URL http://arxiv.org/pdf/cs.CV/0101036.

Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi.
The Similarity Metric.
IEEE TRANSACTIONS ON INFORMATION THEORY, 50(12):3250–32364, December 2004.
URL http://homepages.cwi.nl/~paulv/papers/similarity.pdf.

Minh N. Do.
Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models.
IEEE Signal Processing Letters, 10:115–118, April 2003.
URL http://www.ifp.uiuc.edu/~minhdo/publications/KLD_HMM.pdf.

Yaniv Dover.
A short account of a connection of power laws to the information entropy.
PHYSICA A, 334:591–597, 2004.
URL http://arxiv.org/pdf/cond-mat/0309383.

Bibliography III

FSF.
GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

E. T. Jaynes.
Probability Theory as Logic.
In P.F. Fougère, editor, *Maximum-Entropy and Bayesian Methods*, page 1. Kluwer, Dordrecht, 1990.
URL http://bayes.wustl.edu/etj/articles/prob.as.logic.pdf.

E. T. Jaynes.
Probability Theory: The logic of Science.
1995.
URL http://bayes.wustl.edu/etj/prob/book.pdf.
Web publication.

Kenji Kawamura and Naomichi Hatano.
Universality of zipf's law.
Preprint archive, 2002.
URL http://arxiv.org/pdf/cond-mat/0203455.

David J. C. MacKay.
Information Theory, Inference, and Learning Algorithms.
Cambridge University Press, September 2003.
URL http://www.inference.phy.cam.ac.uk/mackay/itila/.

Bibliography IV

David J.C. MacKay.
A Short Course in Information Theory.
Web, January 1995.
URL http://www.inference.phy.cam.ac.uk/mackay/info-theory/course.html.

Maxent.
Maximum Entropy Modeling.
Web.
URL http://homepages.inf.ed.ac.uk/s0450736/maxent.html.
Mainly links.

Roni Rosenfeld.
Adaptive Statistical Language Modeling: A Maximum Entropy Approach.
PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, 19 April 1994.
URL http://www.cs.cmu.edu/~roni/papers/me-thesis-TR-94-138.pdf.

Roni Rosenfeld.
A maximum entropy approach to adaptive statistical language modelling.
Computer, Speech and Language, 10:187–228, 1996.
URL http://www.cs.cmu.edu/~roni/me-csl-revised.ps.

Tom Schneider.
Information Theory Primer, With an Appendix on Logarithms.
Web, 9 February 1999.
URL http://www.lecb.ncifcrf.gov/~toms/paper/primer/.
updated 2005.

Bibliography V

D. J. Strom.
Introduction to Bayesian Statistics.
Web.
URL http://bidug.pnl.gov/presentations/PEP/.
Slides.

Nick Szabo.
Introduction to Algorithmic Information Theory.
Web, 1996.
URL http://szabo.best.vwh.net/kolmogorov.html.

H. Thornburg.
Introduction to Bayesian Statistics.
Web, 2003.
URL http://ccrma.stanford.edu/~jos/bayes/bayes.html.

Paul Vitanyi.
Universal Similarity.
In *Proceedings of ITW2005*, 29 August 2005.
URL http://arxiv.org/pdf/cs.IR/0504089.
arXiv:cs.IR/0504089.

Bibliography

Bibliography II

Bibliography III

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".
   Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.
1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.
   You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.
2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:
   1. You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.
   2. You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

Bibliography IV

Bibliography V

2. Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
3. Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.
If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.
6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.
7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.
   If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.
   It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to

decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.
This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.
9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.
   Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.
10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

No Warranty

Bibliography VIII

11. Because the program is licensed free of charge, there is no warranty for the program, to the extent permitted by applicable law. Except when otherwise stated in writing the copyright holders and/or other parties provide the program "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the program is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.
12. In no event unless required by applicable law or agreed to in writing will any copyright holder, or any other party who may modify and/or redistribute the program as permitted above, be liable to you for damages, including any general, special, incidental or consequential damages arising out of the use or inability to use the program (including but not limited to loss of data or data being rendered inaccurate or losses sustained by you or third parties or a failure of the program to operate with any other programs), even if such holder or other party has been advised of the possibility of such damages.

End of Terms and Conditions

Bibliography IX

Appendix: How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.
To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

> *one line to give the program's name and a brief idea of what it does.*
> *Copyright (C) yyyy name of author*
> *This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.*
>
> *This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.*
>
> *You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA.*

Also add information on how to contact you by electronic and paper mail.
If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

> *Gnomovision version 69, Copyright (C) yyyy name of author*
> *Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type 'show w'.*
> *This is free software, and you are welcome to redistribute it under certain conditions; type 'show c' for details.*

Bibliography X

The hypothetical commands `show w` and `show c` should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than `show w` and `show c`; they could even be mouse-clicks or menu items—whatever suits your program.
You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

> *Yoyodyne, Inc., hereby disclaims all copyright interest in the program*
> *'Gnomovision' (which makes passes at compilers) written by James Hacker.*
> *signature of Ty Coon, 1 April 1989*
> *Ty Coon, President of Vice*

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.