Scientific Theory in Informatics A1N

Lecture 06

Discrete Probability & Information Theory

David Vernon
Informatics Research Centre
University of Skövde

david.vernon@his.se

Lecture Overview

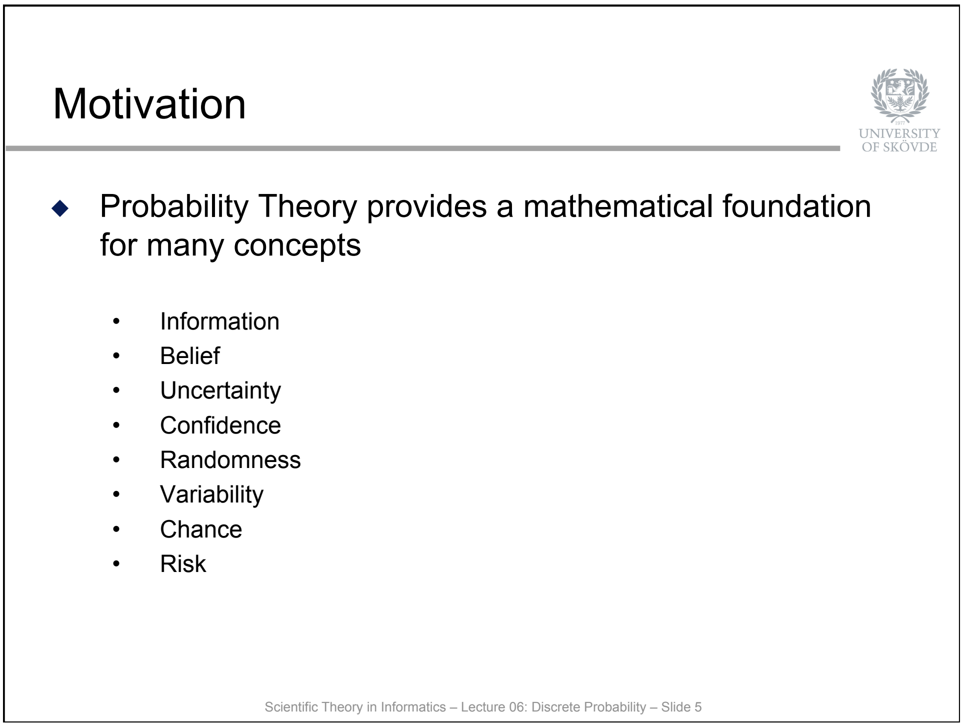
- Motivation & Background
- Probability
  - Experiments, sample spaces, and events
  - Definition of probability
  - Joint and Conditional Probability
  - Independence
  - Random variables
  - Bayes' Theorem
    - » Posterior probability
    - » Naïve Bayes classifier
    - » Probability density functions

Lecture Overview

- Information Theory
  - Uncertainty and surprise
  - Entropy
  - Mutual information
  - Information as reduction in uncertainty

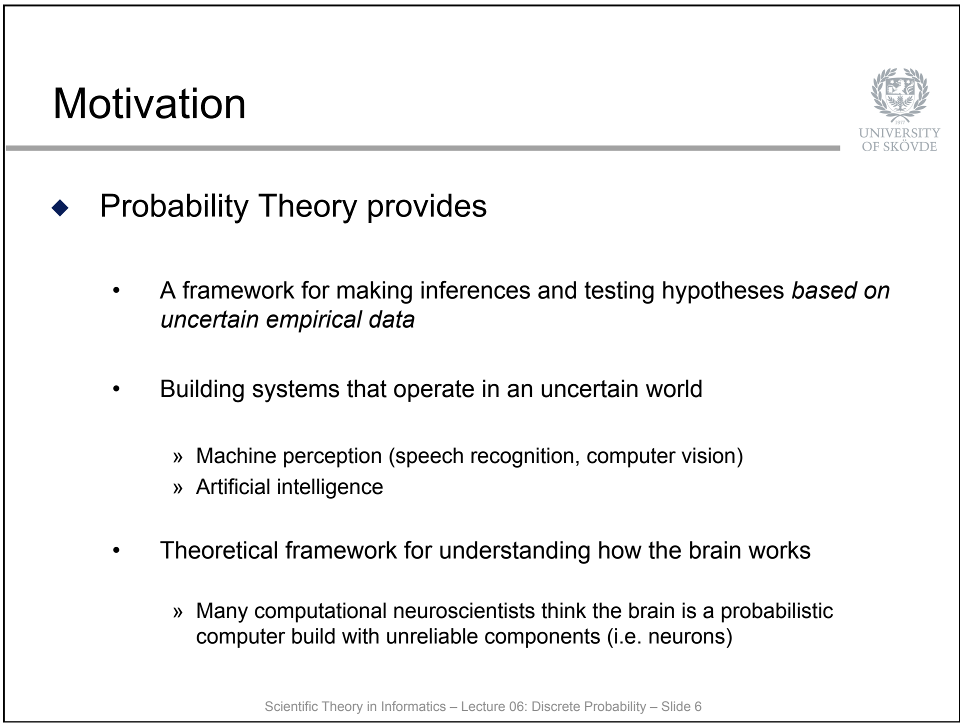
Lecture Overview

- Adapted from
  - K. H. Rosen, *Discrete Mathematics and Its Applications*, 2012.
  - J. R. Movellan, *Introduction to Probability Theory and Statistics*, 2008.
  - D. Vernon, *Machine Vision*, 1991.
  - T. Carter, *An Introduction to Information Theory and Entropy*, 2011.
  - T. Schneider, *Information Theory Primer*, 2012.



Motivation

- Probability Theory provides a mathematical foundation for many concepts
  - Information
  - Belief
  - Uncertainty
  - Confidence
  - Randomness
  - Variability
  - Chance
  - Risk



Motivation

- Probability Theory provides
  - A framework for making inferences and testing hypotheses *based on uncertain empirical data*
  - Building systems that operate in an uncertain world
    - » Machine perception (speech recognition, computer vision)
    - » Artificial intelligence
  - Theoretical framework for understanding how the brain works
    - » Many computational neuroscientists think the brain is a probabilistic computer build with unreliable components (i.e. neurons)

Motivation

- Probability Theory provides
  - A way of determining the **average-case** complexity of algorithms
  - A way of determining whether we should reject an incoming email message as spam based on the words that appear in the message

    (http://www.youtube.com/watch?v=M_eYSuPKP3Y)
  - **A way of combining different sources of uncertain information to make rational decisions**

Background

- Three major interpretations of probability
  - **Frequentist**: probability as a relative frequency
    - » Probability of an event as the proportion of times such an event is expected to happen in the long run.
    - » The probability of an event E would be the limit of the relative frequency of occurrence of that event as the number of observations grows large

Number of times the event is observed

$$P(\mathrm{E}) = \lim_{n\to\infty} \frac{n_{\mathrm{E}}}{n}$$

Number of independent experiments

Background

- Three major interpretations of probability
  - **Frequentist**: probability as a relative frequency
    - Appealing: objective, ties in with work on observation of physical events
    - Can’t perform an experiment an infinite number of times
    - Behaviourist approach: based on observable behaviour of physical systems
    - Doesn’t capture idea of probability as internal knowledge of cognitive systems

Background

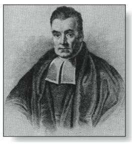
- Three major interpretations of probability
  - **Bayesian or subjectivist**: probability as uncertain knowledge
    - “I will probably get an A in this class”

      By which we mean, “based on what I know about myself and about this class, I would not be very surprised if I get an A. However, I wouldn’t bet my life on it, since there are a multitude of factors which are difficult to predict and that could make it impossible for me to get an A”
    - This notion of probability is cognitive and does not need to be grounded in empirical frequencies

      “I will probably die poor”
      …not able to repeat that experiment many times and count the number of lives in which I die poor

Background

- Three major interpretations of probability
  - **Bayesian or subjectivist**: probability as uncertain knowledge
    - » Useful in the field of machine intelligence
    - » Need to have knowledge systems capable of handling the uncertainty of the world
    - » Probabilists that are willing to represent internal knowledge using probability theory are called "Bayesian" (since he was the first mathematician to do so)



THOMAS BAYES (1702–1761)

Background

- Three major interpretations of probability
  - **Axiomatic or mathematical**: probability as a mathematical model
    - » Rigorous definition
    - » Traceable to first principles
    - » Avoid the frequentist vs. Bayesian debate
    - » Application of probability theory is not the main concern

Finite Probability

- Experiments with finitely many, equally likely, outcomes
  - **Experiment**: a procedure that yields one of a given set of possible outcomes
    - » E.g. rolling a die, tossing a coin, tossing a coin two times
  - **Sample space** S: set of possible outcomes
    - » E.g. Ω = {1,2,3,4,5,6}, Ω = {H, T}, Ω = {(H,H), (H, T), (T, H), (T, T)}
    - » Also called **outcome space** or **reference set**
  - **Event** E: subset of the sample space
    - » Sets of outcomes
    - » The set of all events is called the **event space**
    - » E.g. Rolling an even number on a die: E = {2, 4, 6}

Finite Probability

- Definition of probability

For an event E, and a sample space S

The probability of E is $p(E) = \frac{|E|}{|S|}$

The probability of an event is between 0 and 1

- Example: an box contains four blue balls and five red balls; what is the probability that a ball chosen at random for the box is blue?

9 possible outcomes, four produce a blue ball, so probability is 4/9

Finite Probability

- Probabilities of Complements and Unions of Events

For an event E, and a sample space S

The probability of the complement of E, $\overline{E} = S - E$, is given by

$$p(\overline{E}) = 1 - p(E)$$

- Example: A sequence of 10 bits is randomly generated. What is the probability that at least one of these bits is 0?

Let E be the event with at least one of the 10 bits is 0

Then $\overline{E}$ is the event that all the bits are 1.

The sample space S is the set of all strings of length 10,

Finite Probability

- Probabilities of Complements and Unions of Events

For an event E, and a sample space S

The probability of the complement of E, $\overline{E} = S - E$, is given by

$$p(\overline{E}) = 1 - p(E)$$

- Example: A sequence of 10 bits is randomly generated. What is the probability that at least one of these bits is 0?

$$p(E) = 1 - p(\overline{E}) = 1 - \frac{|\overline{E}|}{|S|} = 1 - \frac{1}{2^{10}}$$

$$= 1 - \frac{1}{1024} = \frac{1023}{1024}.$$

Finite Probability

- Probability measures

  - You can think of probability as a function that assigns a number to a set: probability ‘measures’ a set (hence probability measures)

  - If events $E_1, E_2, \dots\ E_n$ are disjoint (i.e. no elements in common)

$$p(E_1 \cup E_2 \dots \cup E_n) = p(E_1) + p(E_2) + \dots\ p(E_n)$$

  - Probability of rolling a die and getting a 1: p ({1}) = 1/6
    same for 2, 3, 4, 5, and 6.

p({1} ∪ {2} ∪ {3} ∪{4} ∪ {5} ∪ {6}) = 1/6 + 1/6 + 1/6+ 1/6+ 1/6+ 1/6
= 1

Finite Probability

- Probabilities of Intersection of Events: **Joint Probability**

  - For events E_1 and E_2 in a sample space S

$$p(E_1, E_2) = p(E_1 \cap E_2)$$

  - The joint probability of two or more events is the probability of the intersection of those events

Finite Probability

- Probabilities of Intersection of Events: **Joint Probability**
  - Consider the event E_1 = {2, 4, 6} when rolling a die (rolling an even number)
  - Consider the event E_2= {4, 5, 6} (rolling a number greater than 3)
  - The joint probability (rolling an even number greater than 3) …
    - » $p(E_1)$ = p ({2} ∪ {4} ∪ {6}) = 3/6
    - » $p(E_2)$ = p ({4} ∪ {5} ∪ {6}) = 3/6
    - » $p(E_1 \cap E_2)$ = $p(E_1, E_2)$ = p ({4} ∪ {6}) = 2/6
    - » Thus the joint probability of E_1 and E_2, $p(E_1, E_2)$, is 1/3

Finite Probability

- Probabilities of Complements and Unions of Events

  For events E_1 and E_2 in a sample space S

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

  - Example: What is the probability that a positive integer selected at random from the set of positive integers less than or equal to 100 is divisible by either 2 or 5?

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$
$$= \frac{50}{100} + \frac{20}{100} - \frac{10}{100} = \frac{3}{5}.$$

Probability Theory

- Probabilities of outcomes of experiments where outcomes may **not** be equally likely

  - Let S be a sample space of an experiment with a finite or countable number of outcomes

    $p(s)$ is the probability of each outcome s

$$0 \leq p(s) \leq 1 \text{ for each } s \in S$$

$$\sum_{s \in S} p(s) = 1.$$

Probability Theory

- When there are n possible outcomes

$$0 \leq p(x_i) \leq 1 \text{ for } i = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} p(x_i) = 1.$$

  - The function $p(s)$ or $p(x_i)$ from the set of all outcomes of sample space S is called a **probability distribution**
  - The *uniform distribution* assigns the probability $1/n$ to each element of S

Probability Theory

- Definition of the probability of an event

$$p(E) = \sum_{s\in E} p(s)$$

The probability of an event E is the sum of the probabilities of the outcomes in E

Probability Theory

- Conditional Probability

  - Let E and F be events with $p(F) > 0$

    The conditional probability of E given F, denoted $p(E \mid F)$, is defined as

$$p(E \mid F) = \frac{p(E \cap F)}{p(F)} \longleftarrow p(E, F)$$

Probability Theory

- Conditional Probability

  - What is the conditional probability that a family with two children has two boys, given that they have at least one boy?

    Assume that each of the possibilities BB, BG, GB, GG is equally likely.
    Let E be the event that the family with two children as two boys
    Let F be the event that a family with two children has at least one boy
    $E = \{BB\}$
    $F = \{BB, BG, GB\}$
    $E \cap F = \{BB\}$
    $p(F) = ¾$ and $p(E \cap F) = ¼$

$$p(E \mid F) = \frac{p(E \cap F)}{p(F)} = \frac{1/4}{3/4} = \frac{1}{3}$$

Probability Theory

- Independence

  - If $p(E \mid F) = p(E)$ it means F has no bearing on E

  - We say E and F are independent events

Definition: the events E and F are independent if and only if

$$p(E \cap F) = p(E)\, p(F)$$

Probability Theory

- Independence

Let E be the event that the family with two children as two boys
Let F be the event that a family with two children has at least one boy

Are the two events independent?

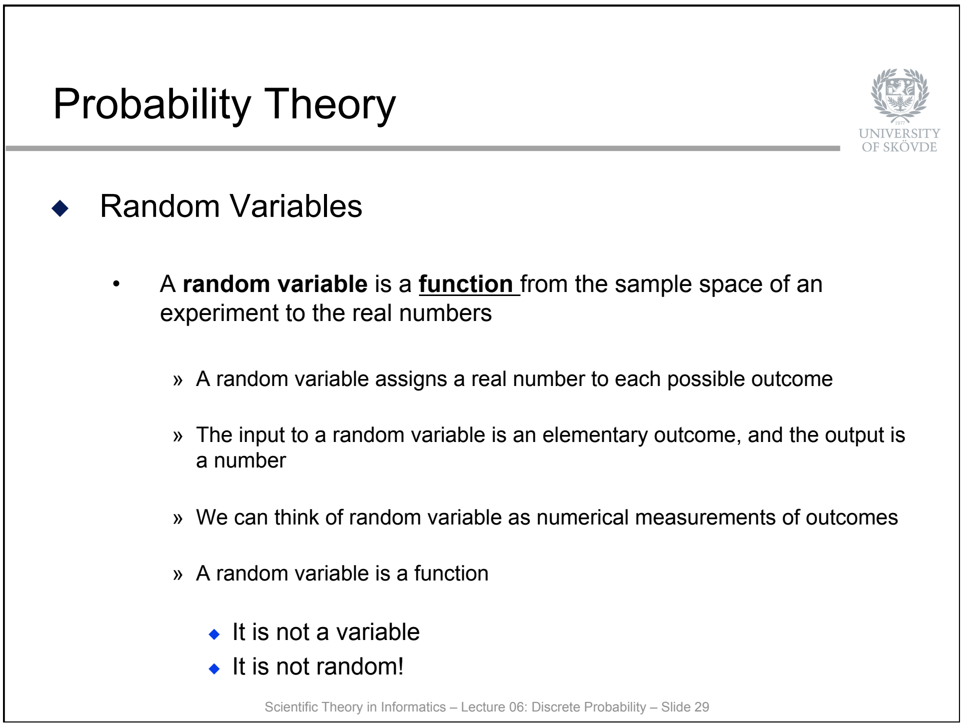
$E = \{BB\}$ so $p(E) = ¼$
$F = \{BB, BG, GB\}$ so $p(F) = ¾$
Thus, $p(E)\ P(F) = 3/16$
$p(E \cap F\ )\ ) = ¼$
Since $p(E \cap F\ ) \neq p(E)\ P(F)$ the events are not independent

Probability Theory

- Random Variables
  - Many problems are concerned with a numerical value associated with the outcome of an experiment
    - E.g. the number of 1 bits in a randomly generated string of 10 bits
    - E.g. the number of times a head comes up when you toss a coin 20 times
    - E.g. some feature of a manufactured part

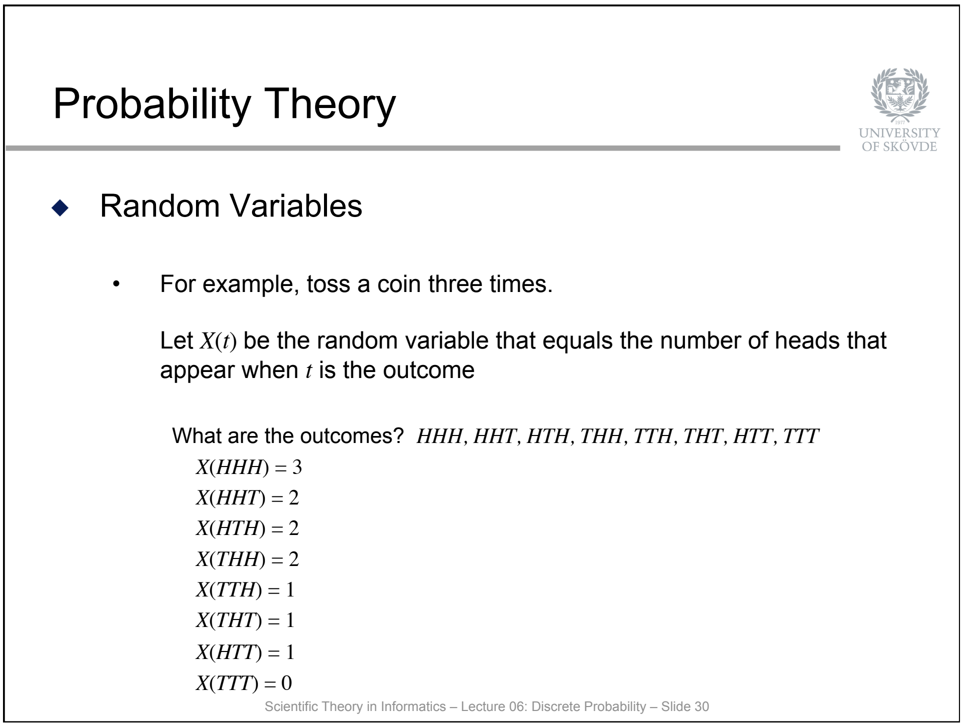
A **random variable** is a **function** from the sample space of an experiment to the real numbers

$f: S \rightarrow \Re$



Probability Theory

- Random Variables
  - A **random variable** is a **function** from the sample space of an experiment to the real numbers
    - » A random variable assigns a real number to each possible outcome
    - » The input to a random variable is an elementary outcome, and the output is a number
    - » We can think of random variable as numerical measurements of outcomes
    - » A random variable is a function
      - It is not a variable
      - It is not random!



Probability Theory

- Random Variables
  - For example, toss a coin three times.

    Let $X(t)$ be the random variable that equals the number of heads that appear when t is the outcome

    What are the outcomes? $HHH, HHT, HTH, THH, TTH, THT, HTT, TTT$

    $X(HHH) = 3$
    $X(HHT) = 2$
    $X(HTH) = 2$
    $X(THH) = 2$
    $X(TTH) = 1$
    $X(THT) = 1$
    $X(HTT) = 1$
    $X(TTT) = 0$

Probability Theory

- Random Variables

The distribution of a random variable X on a sample space S is the set of pairs

$$(r, p(X = r))$$

For all $r \in X(S)$, where $p(X = r)$ is the probability that X takes the value r

For the previous example,

$p(X = 3) = 1/8$
$p(X = 2) = 3/8$
$p(X = 1) = 3/8$
$p(X = 0) = 1/8$

Hence, the **distribution** of $X(t)$ is the set of pairs $(3, 1/8), (2, 3/8), (1, 3/8), (0, 1/8)$

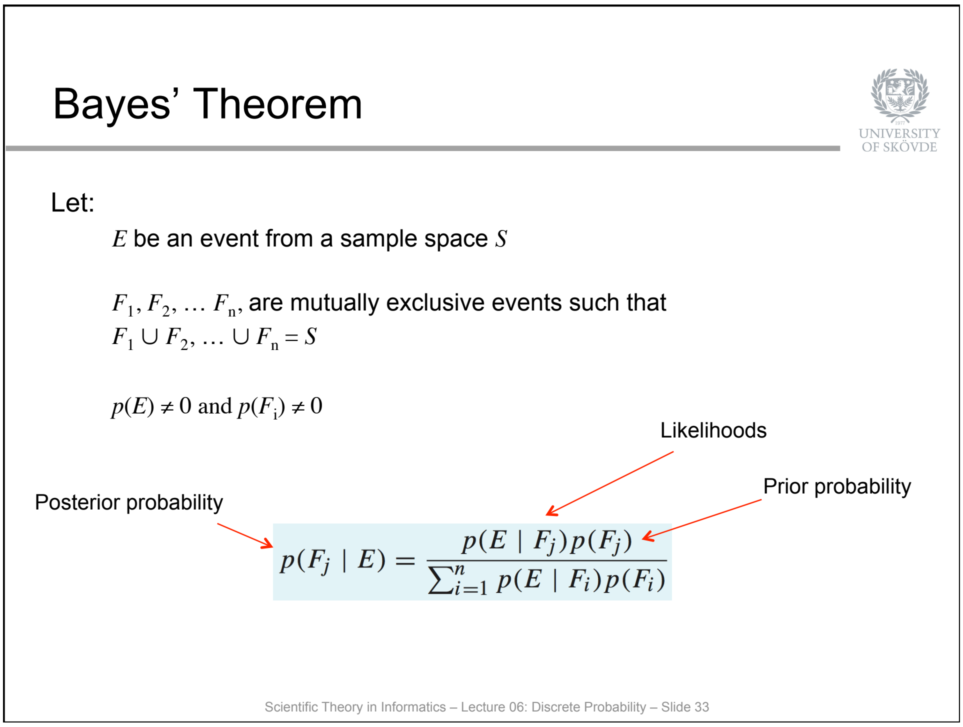
Bayes’ Theorem

- Shows how to revise probability of events in the light of new data

- For example

  - We can determine the probability that a particular incoming email is spam using the occurrence of words in the message

  - To do this we need to know
    - » The percentage of incoming emails that are spam
    - » The percentage of **spam** messages in which these words occur
    - » The percentage of messages that are **not spam** in which each of these words occur



Bayes’ Theorem

Let:

E be an event from a sample space S

$F_1, F_2, \dots F_n$, are mutually exclusive events such that
$F_1 \cup F_2, \dots \cup F_n = S$

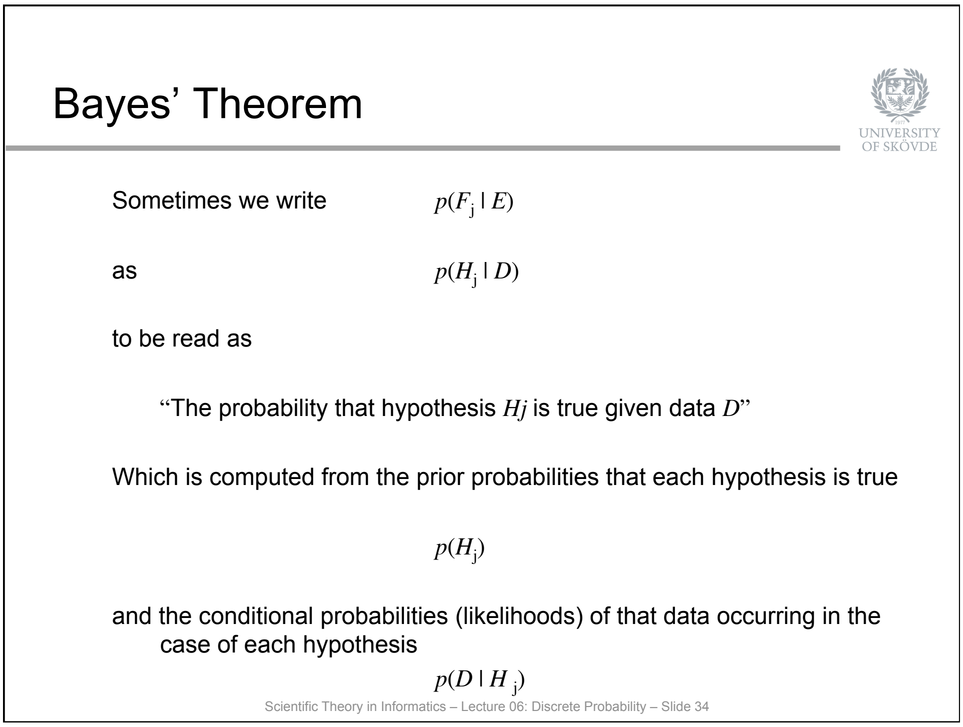
$p(E) \neq 0$ and $p(F_i) \neq 0$

Likelihoods

Prior probability

Posterior probability

$$p(F_j \mid E) = \frac{p(E \mid F_j)p(F_j)}{\sum_{i=1}^{n} p(E \mid F_i)p(F_i)}$$



Bayes’ Theorem

Sometimes we write $p(F_j \mid E)$

as $p(H_j \mid D)$

to be read as

“The probability that hypothesis Hj is true given data D”

Which is computed from the prior probabilities that each hypothesis is true

$$p(H_j)$$

and the conditional probabilities (likelihoods) of that data occurring in the case of each hypothesis

$$p(D \mid H_j)$$

Bayes' Theorem

Example: Bayesian Spam Filter

Let $p(F_1)$ be the prior probability that a message is spam

Let $p(F_2)$ be the prior probability that a message is valid

We can base this on historical data or, assuming maximum ignorance, assume the are both equally likely in which case $p(F_1) = p(F_2)$

Bayes' Theorem

Example: Bayesian Spam Filter

Suppose we have a set of B of messages known to be spam

Suppose we have a set of G of valid messages known not to be spam

Count the number of messages in B containing the word w: $n_B(w)$

The (empirical) probability that a spam message contains word w is $p(w \mid F_1) = n_B(w) / |B|$

Count the number of messages in G containing the word w: $n_G(w)$

The (empirical) probability that a valid message contains word w is $p(w \mid F_2) = n_B(w) / |G|$

Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Bayesian Spam Filter

Now, if we receive a new message with the word w,
the probability that it is spam is given by Bayes' Theorem

$$p(F_1 \mid w) = \frac{p(w \mid F_1)\, p(F_1)}{\Sigma\, p(w \mid F_i)\, p(F_i)}$$

$$= \frac{p(w \mid F_1)\, p(F_1)}{p(w \mid F_1)\, p(F_1) + p(w \mid F_2)\, p(F_2)}$$

To make a decision about whether to accept or reject the message,
we compare the computed probability $p(F_1 \mid w)$ with a threshold value, e.g. 0.9

So, if $p(F_1 \mid w) > 0.9$, we decide the message is spam and we reject it.

Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- Notice that when we said that the probability that a spam message contains word w is $p(w \mid F_1) = n_B(w) / |B|$

  we assumed that there was one unique probability value (which we estimated by $n_B(w) / |B|$)

- That may not always be the case

  » For example, when manufacturing a part, e.g. a nut or a bolt, some feature may vary and different feature values will have different probabilities associated with them … a probability distribution

  » We have already met the concept of a distribution … the number of heads that can occur when we toss a coin three times

Bayes' Theorem

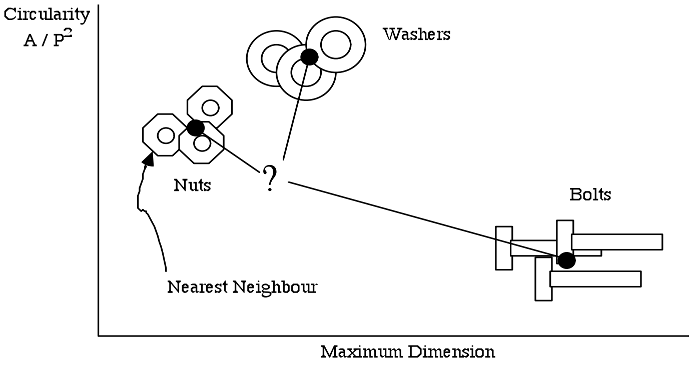
UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- Let's say we are designing a visual inspection system to count (or sort) different parts … this is a classification problem
- We can use Bayes' Theorem to create a good classifier (better than a simple 'nearest neighbour' classifier) by employing probability theory

Bayes' Theorem

UNIVERSITY OF SKÖVDE

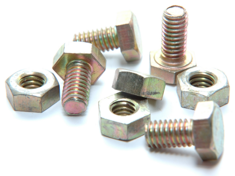


Nearest Neighbour Classification
In 2-D feature space

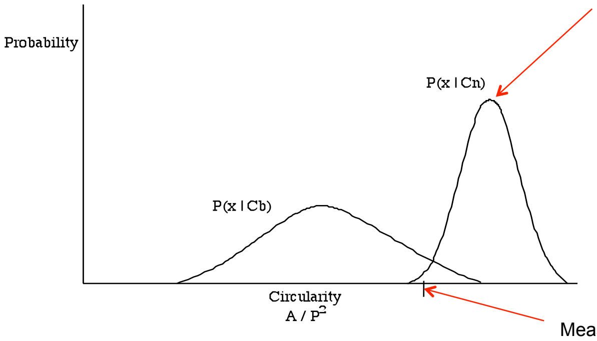
Bayes' Theorem

Example: Maximum Likelihood Classifier

- Let's design a system that can classify two parts: nuts and bolts
- Two classes $C_b C_n$
- Let's decide to use a feature 'circularity' x to distinguish nuts from bolts (nuts are more circular than bolts)



Bayes' Theorem



Nearest neighbour

Measured value

We can do better than nearest neighbour if we use knowledge about the probability of an object having some feature value $P(x|C_i)$ and the probability of that object being there at all $P(C_i)$

Bayes’ Theorem

Example: Maximum Likelihood Classifier

- Of course, neither of these probabilities are what we are interested in!
- We want the probability that an object belongs to a particular class*, given that a particular value of* x *has occurred* $P(C_i|x)$
- Thus, we classify the object as a bolt if

$$P(C_b|x) > P(C_n|x)$$

- We use Bayes’ Theorem to convert the probabilities we know (or can measure) to the ones we need

Bayes’ Theorem

Example: Maximum Likelihood Classifier

- The *posterior* probability, $P(C_i|x)$, that the object belongs to a particular class i and is given by Bayes’ Theorem:

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}$$

where

$$P(x) = \sum_{i=1}^{2} P(x|C_i)P(C_i)$$

Bayes’ Theorem

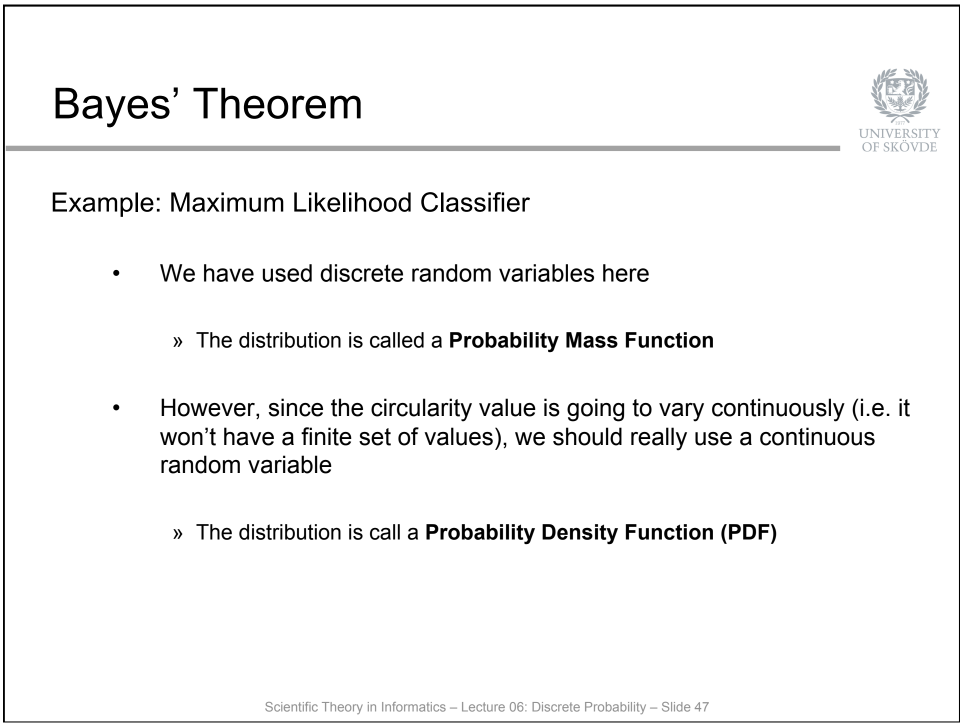
Example: Maximum Likelihood Classifier

- The first thing that is required is the probability for each of these two classes, *i.e.*, a measure of the probabilities that an object from a particular class will have a given feature value
- Since it is not likely that we will know these *a priori*, we will have to estimate them

Bayes’ Theorem

Example: Maximum Likelihood Classifier

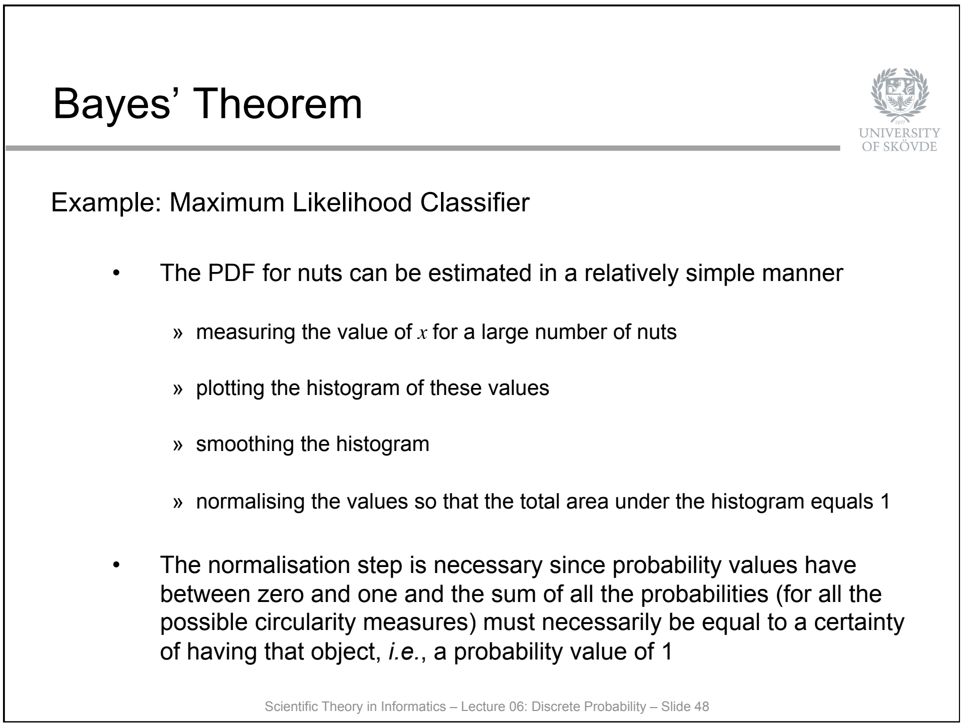
- Let S be the space of circularity values that we can measure with our computer vision system
- Let $X_n(x)$ be the random variable that equals the number of times a given circularity value appears that appear when x is the outcome (i.e. when the circularity of a nut is measured)
- Let $X_b(x)$ be the random variable that equals the number of times a given circularity value appears that appear when x is the outcome (i.e. when the circularity of a bolt is measured)
- We need the probability distribution of random variables X_n and X_b



Bayes’ Theorem

Example: Maximum Likelihood Classifier

- We have used discrete random variables here
  - » The distribution is called a **Probability Mass Function**
- However, since the circularity value is going to vary continuously (i.e. it won’t have a finite set of values), we should really use a continuous random variable
  - » The distribution is call a **Probability Density Function (PDF)**



Bayes’ Theorem

Example: Maximum Likelihood Classifier

- The PDF for nuts can be estimated in a relatively simple manner
  - » measuring the value of x for a large number of nuts
  - » plotting the histogram of these values
  - » smoothing the histogram
  - » normalising the values so that the total area under the histogram equals 1
- The normalisation step is necessary since probability values have between zero and one and the sum of all the probabilities (for all the possible circularity measures) must necessarily be equal to a certainty of having that object, *i.e.*, a probability value of 1

Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- The PDF for the bolts can be estimated in a similar manner.
- Next problem: the probability *of each class occurring*
  - » We may know, for instance, that the class of nuts is, in general, likely to occur twice as often as the class of bolts
  - » In this case we say that the prior (or *a priori*) probability of the two classes are :

    $P(C_n)$ = 0.666 and $P(C_b)$ = 0.333
  - » In fact, in this case, it is more likely that they will have the same *a priori* probabilities (0.5) since we usually have a nut for each bolt

Classification using Bayes’ Theorem

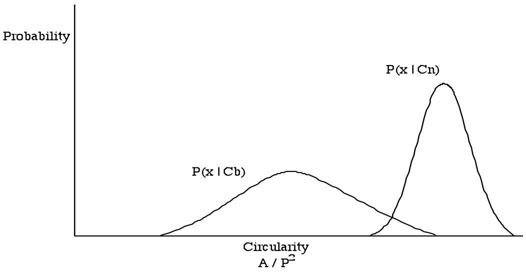
Example: Maximum Likelihood Classifier

- The PDFs tell us the probability that the circularity x will occur, given that the object belongs to the class of nuts C_n in the first instance and to the class of bolts C_b in the second instance
- As we know, this is termed the conditional probability of an object having a certain feature value, given that we know that it belongs to a particular class

Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- Thus, the conditional probability, $p(x|C_b)$ enumerates the probability that a circularity *x* will occur, given that the object is a bolt.
- The two conditional probabilities $p(x|C_b)$ and $p(x|C_n)$ are shown below



Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- This is not what are interested in …
- We want the probability that an object belongs to a particular class, given that a particular value of x has occurred (*i.e.* been measured), allowing us to establish its identity

Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- This is called the posterior (or *a posteriori*) probability, $p(C_i|x)$ that the object belongs to a particular class *i* and is given by Bayes’ Theorem

$$P\left(C_i|\,x\right)=\frac{P\left(x|\,C_i\right)P\left(C_i\right)}{P(x)}$$

where

$$P(x)=\sum_{i=1}^{2}P\left(x|\,C_i\right)P\left(C_i\right)$$

Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- $p(x)$ is a normalisation factor which is used to ensure that the sum of the *a posteriori* probabilities sum to one, for the same reasons as mentioned earlier

Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- In effect, Bayes’ theorem allows us to use
  - » the *a priori* probability of objects occurring in the first place
  - » the conditional probability of an object having a particular feature value given that it belongs to a particular class and …
  - » The actual measurement of a feature value (to be used as the parameter in the conditional probability) to estimate the probability that the measured object belongs to a given class
  - » Once we can estimate the probability that, for a given measurement, the object is a nut and the probability that it is a bolt, we can make a decision as to its identity, choosing the class with the higher probability

Classification using Bayes’ Theorem

Example: Maximum Likelihood Classifier

- This is why it is called the maximum likelihood classifier
- Thus, we classify the object as a bolt if :

$$P\left(C_b|x\right)> P\left(C_n|x\right)$$

Classification using Bayes' Theorem

Example: Maximum Likelihood Classifier

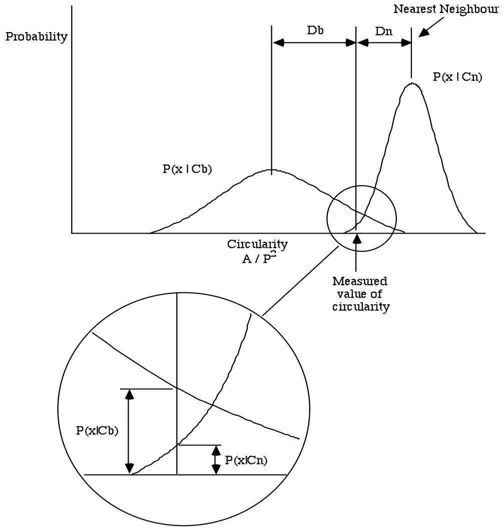
- Using Bayes' Theorem again, and noting that the normalising factor $p(x)$ is the same for both expressions, we can rewrite this test as

$$P(x|C_b)P(C_b) > P(x|C_n)P(C_n)$$

- If we assume that the chances of an unknown object being either a nut or a bolt are equally likely (*i.e.* $P(C_b) = P(C_n)$), then we classify the unknown object as a bolt if :

$$P(x|C_b) > P(x|C_n)$$

Bayes' Theorem



Classification using Bayes' Theorem

Example: Maximum Likelihood Classifier

- For the example shown $p(x|C_b)$ is indeed greater than $p(x|C_n)$ for the measured value of circularity and we classify the object as a bolt

- If, on the other hand, we were to use the nearest neighbour classification technique, we would choose the class whose mean value "is closer to" the measured value

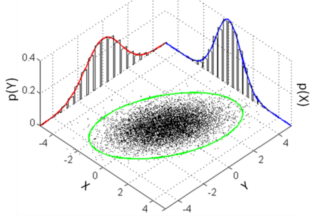
  » In this case, the distance D_n from the measured value to the mean of the PDF for nuts is less than D_b, the distance from the measured value to the mean of the PDF for bolts; we would erroneously classify the object as a nut

Classification using Bayes' Theorem

Example: Maximum Likelihood Classifier

- This was a simple example with just one feature and a 1-D PDF

- However, the argument generalizes directly to *n*-dimensions, where we have *n* features in which case the conditional probability density functions are also *n*-dimensional



Classification using Bayes' Theorem

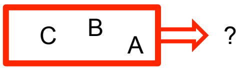
Example: Maximum Likelihood Classifier

- If we assume that the features are independent then we can use the theory we've just outlined, multiplying together the conditional probabilities for each class
  - » This is known a **Naïve Bayes Classifier**
  - » It may be naïve, but it works surprisingly well
- If we don't assume independence,
  then we need a more complex theory!

Information Theory

◆ Measures of uncertainty

- Information and uncertainty are technical terms that describe any process that selects one or more objects from a set of objects
- Suppose we have a device that can produce 3 symbols, A, B, or C
  - » Wait for the next symbol … uncertain about which symbol will be produced



  - » Once a symbol appears, and we see it, our uncertainty decreases … we have received some information
- **Information is a decrease in uncertainty**

Information Theory

- Measures of uncertainty
  - How should information be measured?
    - » For the three-symbol device: “uncertainty of 3 symbols”?
  - Consider a second device
    - » “uncertainty of 2 symbols”?

  - What happens when we combined them as one device:
    - » Six symbols: A1, A2, B1, B2, C1, C2 … uncertainty of 6 symbols
    - » Not good: **we would like our measure of information to be additive**

Information Theory

- Measures of uncertainty
  - We can do this by using logs
    - » $\log(3) + \log(2) = \log(3 \times 2) = \log(6)$
    - » The base of the log determines the units of uncertainty

      $\log_2$ units are bits (from ’binary’)
      $\log_3$ units are trits(from ’trinary’)
      $\log_e$ units are nats (from ’natural logarithm’) … usually use $\ln(x)$ for $\log_e(x)$
      $\log_{10}$ units are Hartleys, after an early worker in the field

Information Theory

- Measures of uncertainty
  - If a device produces one symbol, we are uncertain by $\log_2(1) = 0$ bits … because there is no uncertainty about what the device will do next
  - If a device produces two symbols (**with equal probability**), we are uncertain by $\log_2(1) = 1$ bit
  - Candidate formula for uncertainty is $\log_2(M)$, where M is the number of symbols

$$\begin{aligned}\log_2(M) &= -\log_2(M^{-1}) \\ &= -\log_2(1/M) \\ &= -\log_2(P)\end{aligned}$$

P is the probability that any symbol appears

Information Theory

- Measures of uncertainty
  - Now take the probability of the symbols appearing into account

    Generalize for the probabilities of the M individual symbols, P_i

$$\sum_{i=1}^{M} P_i = 1$$

  - The surprise we get send we see the i^{th} type of symbol is sometimes called “surprisal”: the degree of uncertainty about an outcome i

$$u_i = -\log_2(P_i) \qquad = \log_2(1/P_i)$$

Information Theory

- Measures of uncertainty
  - **Uncertainty** is the **average surprisal** for an infinite string of symbols produced by the device
  - Let's find the average for a string of length N that has an alphabet of M symbols
    - » Suppose the i^{th} kind of symbol appears N_i times, then $N = \sum_{i=1}^{M} N_i$

Information Theory

- Measures of uncertainty
  - The average surprisal for the N symbols is given by

$$\frac{\sum_{i=1}^{M} N_i u_i}{\sum_{i=1}^{M} N_i}$$

equivalently:

$$\sum_{i=1}^{M} \frac{N_i}{N} u_i$$

substituting in the term for probability

$$H = \sum_{i=1}^{M} P_i u_i$$

This is Shannon's famous 1948 definition of uncertainty (in bits per symbol)

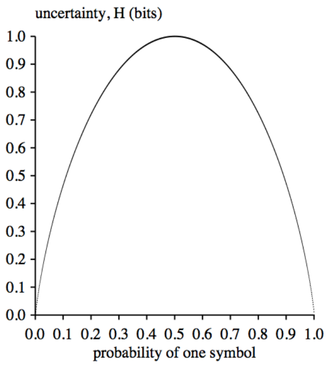
Hence:

$$\boxed{H = -\sum_{i=1}^{M} P_i \log_2 P_i}$$

H is called **Entropy**

Information Theory

- Measures of uncertainty



H for the case of two symbols

Information Theory

- Entropy
  - Suppose that we have n symbols $\{a_1, a_2, \ldots, a_n\}$
  - Some source is providing a stream of these symbols
  - Suppose that the source emits the symbols with probabilities $\{p_1, p_2, \ldots, p_n\}$ respectively
  - Assume that the symbols are emitted independently (successive symbols do not depend in any way on past symbols)
  - What is the average amount of information we get from each symbol we see in the stream?

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Information Theory

- Entropy
  - Entropy is defined over some probability distribution
  - **Entropy provides a lower bound for the efficiency of an encoding or communication scheme**, i.e. a lower bound on the possible compression of a data stream
  - **To improve a coding or compression technique, we need to improve the probability model of the data stream**

Information Theory

- Channel Capacity
  - So far, we have assumed that the storage mechanism or the communication channel is perfect, i.e. it is noise-free and doesn’t introduce errors
  - The more realistic situation is that the channel is noisy
    - » There is uncertainty about the data stream being transmitted through the channel
    - » There is uncertainty about the errors the channel introduces

Information Theory

- Information as decrease in uncertainty / entropy
  - In some books you will see information being equated with entropy
  - This is wrong: information is a decrease in uncertainty at the receiver

$R = H_{before} - H_{after}$

Where H is the Shannon uncertainty

Information Theory

- Information as decrease in uncertainty / entropy

“Imagine that we are in communication and that we have agreed on an alphabet.

Before I send you a bunch of characters, you are uncertain (H_{before}) as to what I'm about to send.

After you receive a character, your uncertainty goes down (to H_{after}).

H_{after} is never zero because of noise in the communication system.

Your decrease in uncertainty is the information (R) that you gain.”

T. Scheider 2013
http://schneider.ncifcrf.gov/bionet.info-theory.faq.html#Information.Equal.Entropy

Information Theory

- Information as decrease in uncertainty / entropy

> “Many of the statements in the early literature assumed a noiseless channel, so the uncertainty after receipt is zero (H_{after} = 0).
>
> This leads to the SPECIAL CASE where R = H_{before}.
>
> But H_{before} is NOT "the uncertainty", it is the uncertainty of the receiver BEFORE RECEIVING THE MESSAGE.”

T. Scheider 2013
http://schneider.ncifcrf.gov/bionet.info-theory.faq.html#Information.Equal.Entropy

Information Theory

- Channel Capacity
  - Given a probability models for the source AND for the channel
    - » We can talk about the **capacity** of the channel
  - The general model involves two sets of symbols

    Input symbols $A = \{a_1, a_2, \ldots , a_n\}$
    Output symbols $B = \{b_1, b_2, \ldots , b_m\}$

  - Given the noise in the channel, we cannot be certain which a_i was put in
    - » The channel is characterized by the set of conditional probabilities $\{P(a_i \mid b_j)\}$

Information Theory

- Mutual Information
  - Consider the information we get by observing a symbol b_j
  - Given a probability model of the source, we have an a priori estimate $P(a_i)$ that the symbol a_i will be sent next
  - Upon observing b_j we can revise our estimate to $P(a_i \mid b_j)$
  - The change in our information – **the mutual information** – is given by the decrease in uncertainty

$$
\begin{aligned}
I(a_i\,;\,b_j) &= \log(1 / P(a_i)) - \log\,((1 / P(a_i \mid b_j)) \\
&= \log(\,P(a_i \mid b_j)\ \ /\ P(a_i)\,)
\end{aligned}
$$

Information Theory

- Mutual Information
  - What we really want is the average **mutual information** over all the symbols

$$
\begin{aligned}
I(A;B) &= \sum_i \sum_j P(a_i,b_j) * \log\left(\frac{P(a_i,b_j)}{P(a_i)P(b_j)}\right) \\
&= I(B;A).
\end{aligned}
$$

Information Theory

$$
\begin{aligned}
H(A) &= \sum_{i=1}^{n} P(a_i) * \log(1/P(a_i)) \\
H(B) &= \sum_{j=1}^{m} P(b_j) * \log(1/P(b_j)) \\
H(A|B) &= \sum_{i=1}^{n} \sum_{j=1}^{m} P(a_i|b_j) * \log(1/P(a_i|b_j)) \\
H(A,B) &= \sum_{i=1}^{n} \sum_{j=1}^{m} P(a_i,b_j) * \log(1/P(a_i,b_j)) \\
H(A,B) &= H(A) + H(B|A) \\
&= H(B) + H(A|B),
\end{aligned}
$$

$$
\begin{aligned}
I(A;B) &= H(A) + H(B) - H(A,B) \\
&= H(A) - H(A|B) \\
&= H(B) - H(B|A) \\
&\geq 0
\end{aligned}
$$

Information Theory

- Shannon's Main Theorem
  - For any channel, there exist ways of encoding input symbols such that we can simultaneously utilize the channel as closely as we wish to capacity, and at the same time have an error rate as close to zero as we wish
  - Shannon's proof is non-constructive: it doesn't say how to build the coding system to optimize channel use, it just tells us that such a code exists
  - To use the capacity with low error rates, we may have to encode very large blocks of data … this might cause problems for real-time communication because of time lags while filling buffers
  - Lots of work still to be done in the search for efficient coding schemes, e.g turbo codes