Scientific Theory in Informatics A1N

UNIVERSITY OF SKÖVDE

Discrete Probability

Alberto Montebelli
School of Informatics
University of Skövde

alberto.montebelli@his.se

Lecture Overview

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

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

Motivation

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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
  - **A way of combining different sources of uncertain information to make rational decisions**
  - ...

Background

UNIVERSITY OF SKÖVDE

- 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

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

Number of times the event is observed

Number of independent experiments

Background

UNIVERSITY OF SKÖVDE

- 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
    - » Doesn't capture idea of probability as internal knowledge of cognitive systems

Background

UNIVERSITY OF SKÖVDE

- Three major interpretations of probability

  - **Bayesian or subjectivist**: probability as uncertain knowledge - a mental phenomenon in continuous re-negotiation

    » "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"

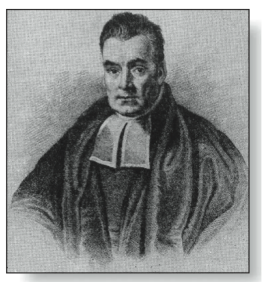
    » This notion of probability is cognitive and does not need to be grounded in empirical frequencies

    "I will probably die poor"
    …not able to repeat that experiment many times and count the number of lives in which I die poor – but I can still imagine the associated probability

Background

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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)}
  - **Event** E: subset of the sample space
    - » Sets of outcomes
    - » E.g. Rolling an even number on a die: E = {2, 4, 6}

Finite Probability

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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 $E_1, E_2, \ldots\ E_n$ are **disjoint** (i.e. no elements in common)

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

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

$$p(\{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\}) = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$$

Finite Probability

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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) …
    - » $p(E_1) = p(\{2\} \cup \{4\} \cup \{6\}) = 3/6 = 1/2$
    - » $p(E_2) = p(\{4\} \cup \{5\} \cup \{6\}) = 3/6 = 1/2$
    - » $p(E_1 \cap E_2) = p(E_1, E_2) = p(\{4\} \cup \{6\}) = 2/6$
    - » Thus the joint probability of E_1 and E_2, $p(E_1, E_2)$, is 1/3

Finite Probability

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

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

$p(E, F)$ joint prob.

Probability Theory

UNIVERSITY OF SKÖVDE

- 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

$S = \{\text{GG}, BB, BG, GB\}$

$E = \{BB\}$

$F = \{BB, BG, GB\}$

$E \cap F = \{BB\}$

$p(F) = ¾$ 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- Independence

Let E be the event that the family with two children has 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- Random Variables
  - "A **random variable** is a **function** from the sample space of an experiment to the real numbers", i.e.:
    - » 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, therefore:
      - It is not a variable
      - It is not random!

Probability Theory

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- 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

UNIVERSITY OF SKÖVDE

- Shows how to revise probability of events in the light of new data
- Derivation Bayes' Th. – from the definition of conditional probability:

$$P(A \cap B) = P(A|B)\,P(B)$$
$$P(B \cap A) = P(B|A)\,P(A)$$

$P(A \cap B) = P(B \cap A)$

divide both terms by P(B)

$$\Rightarrow P(A|B)\,P(B) = P(B|A)\,P(A) \Rightarrow$$

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Bayes' Theorem

UNIVERSITY OF SKÖVDE

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$$

- Interpretation of Bayes' Theorem:
  - *H: patient with a certain pathology*
  - *E: measurement on patient (e.g. blood pressure)*

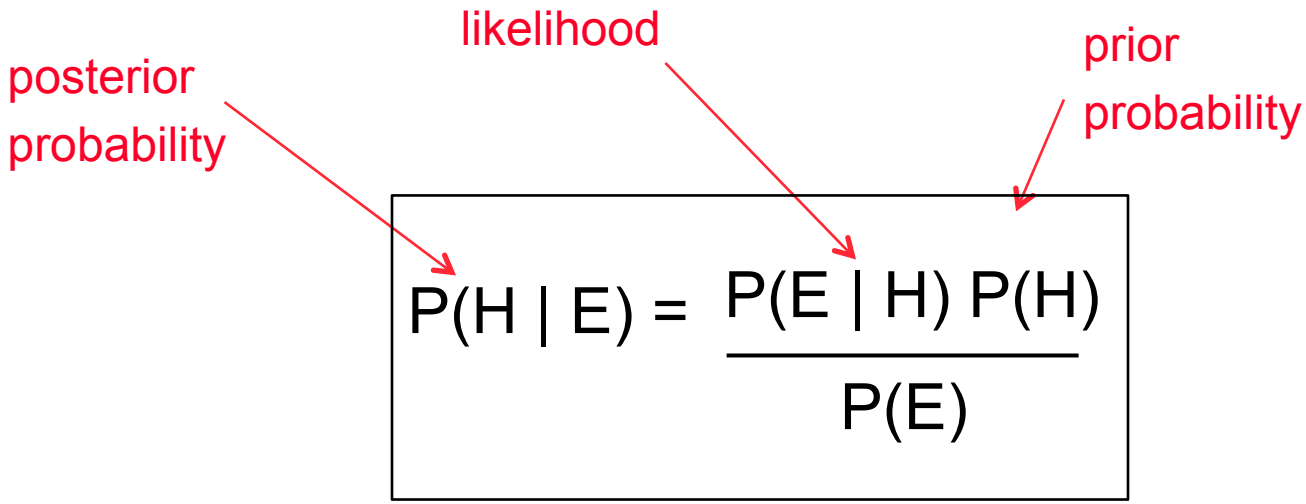
P(E|H), probability that pathological patients produce measure E

P(H), probability of pathological patients in given population

P(E), probability of measurement E in population

Such probabilities constitute the body of medical knowledge, and from such *a priori* knowledge we can calculate the posterior probability (after we have measured!)

Bayes' Theorem



We can also read p(H | E) as:

"The probability that hypothesis H is true given evidence E"

Which is computed

- from the **prior probability** p(H) that the hypothesis is true,
- and the probability (**likelihood**) of that evidence occurring in the case of the hypothesis p(E | H) and of the evidence itself p(E)

Bayes' Theorem

We can also write Bayes' Theorem in a different form. Let:

- E be an event from a sample space S
- $F_1, F_2, \ldots F_n$, are mutually exclusive events such that $F_1 \cup F_2, \ldots \cup F_n = S$
- Be $p(E) \neq 0$ and $p(F_i) \neq 0$

$$p(F_j \mid E) = \frac{p(E \mid F_j)\,p(F_j)}{p(E)}$$

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

$$p(E) = \sum_{i=1}^{n} p(E \cap F_i) = (from\ def.\ of\ conditional\ prob.)$$

$$= \sum_{i=1}^{n} p(E|F_i)p(F_i)$$

Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Bayesian Spam Filter

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

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, we can start by assuming that the two cases are equally likely:

$$p(F_1) = p(F_2)$$

Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Bayesian Spam Filter

Suppose we have a set of $\boldsymbol{B}$ of messages known to be spam
and a set of $\boldsymbol{G}$ 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

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

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

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 on a continuous scale, and different feature values will have different probabilities associated with them … a **probability distribution**

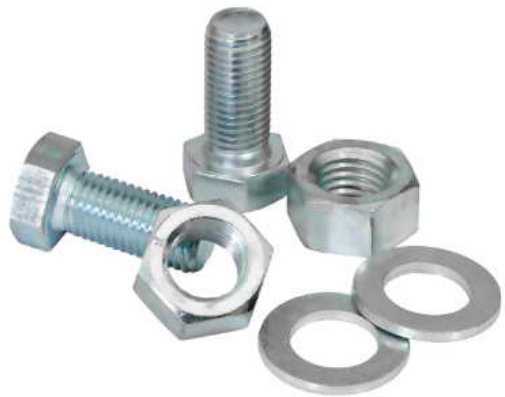
  » We have already met the concept of a distribution … the number of heads that can occur when we toss a coin three times (the same idea can be extended to the continuous case)

Bayes' Theorem

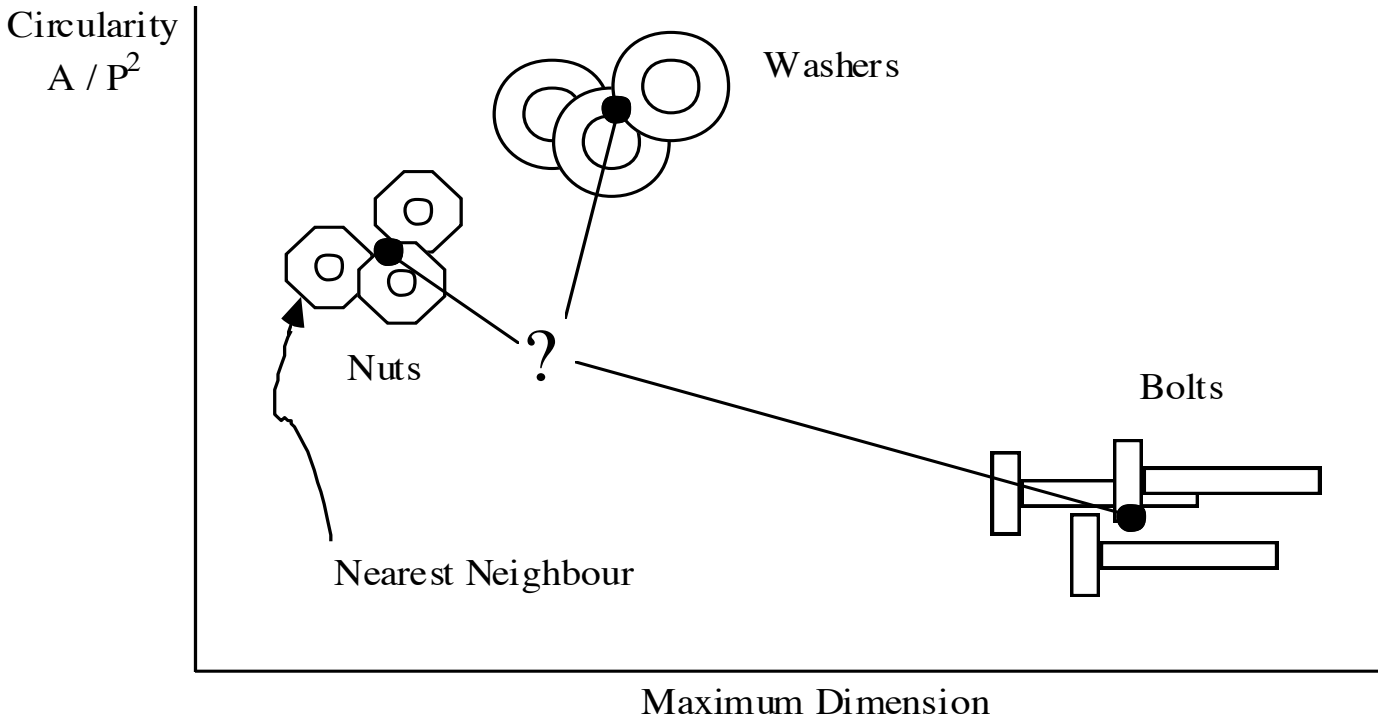
UNIVERSITY OF SKÖVDE

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 neighbor' classifier) by employing probability theory



Bayes' Theorem



Nearest Neighbor Classification
In 2-D feature space

Bayes' Theorem

UNIVERSITY OF SKÖVDE

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

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- We want the probability that an object belongs to a particular class, *given that a particular value of* x *has occurred* $\boldsymbol{P(C_i|x)}$
- 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 estimate (i.e. prior) to the ones we need (posterior)

Bayes' Theorem

UNIVERSITY OF SKÖVDE

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

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

By applying Bayes' Theorem, we can see that to compare posterior probabilities we need to estimate prior probabilities and likelihoods for the two classes!

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- One information is needed: estimate the *prior*, i.e. 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 :

$$P(C_n) = 0.666 \text{ and } P(C_b) = 0.333$$

  - » 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

Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- Then we have to estimate the conditional probability (likelihood) for each of these two classes, *i.e.* $P(x|C_i)$, 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 already know these, we will have to **estimate** them (see next slide)
- The circularity value is going to vary continuously (i.e. it won't have a finite set of values), we should use a continuous random variable
  - » The probability distribution for a continuous random variable is called **Probability Density Function (PDF)**

Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- The **PDF for nuts** (similarly for **bolts**) can be (roughly) **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
  - » normalizing the values so that the total area under the histogram equals 1
- The normalization 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

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- The PDFs that we have just estimated (likelihoods – visualized in the next slide) 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, *i.e.* $P(x|C_i)$

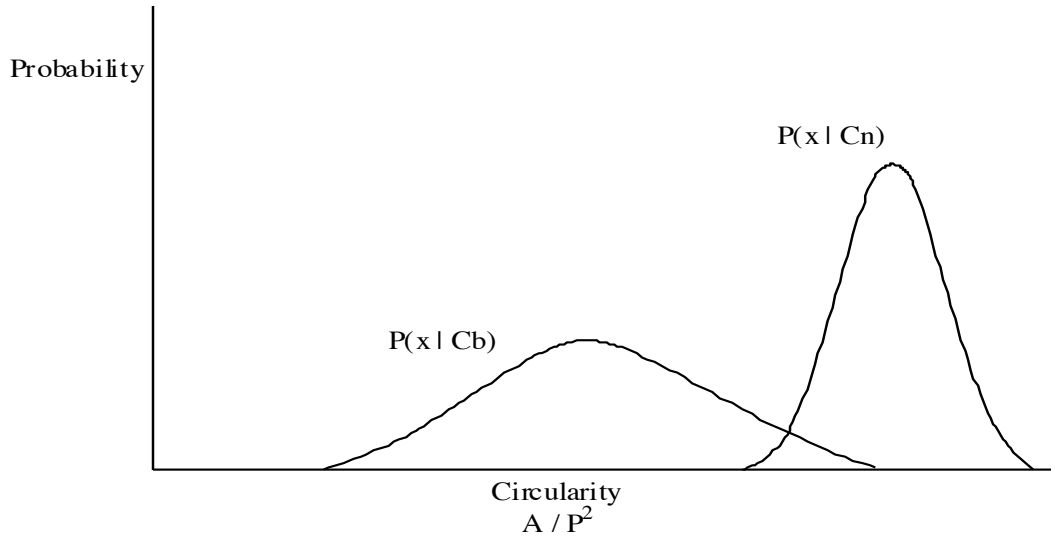
- As we know, this is termed the conditional probability of an object having a certain feature value, given that we know that it belongs to a particular class

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

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

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- Again, 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

UNIVERSITY OF SKÖVDE

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

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

where

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

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- $p(x)$ is a normalization 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

UNIVERSITY OF SKÖVDE

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 PDF 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

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

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

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

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

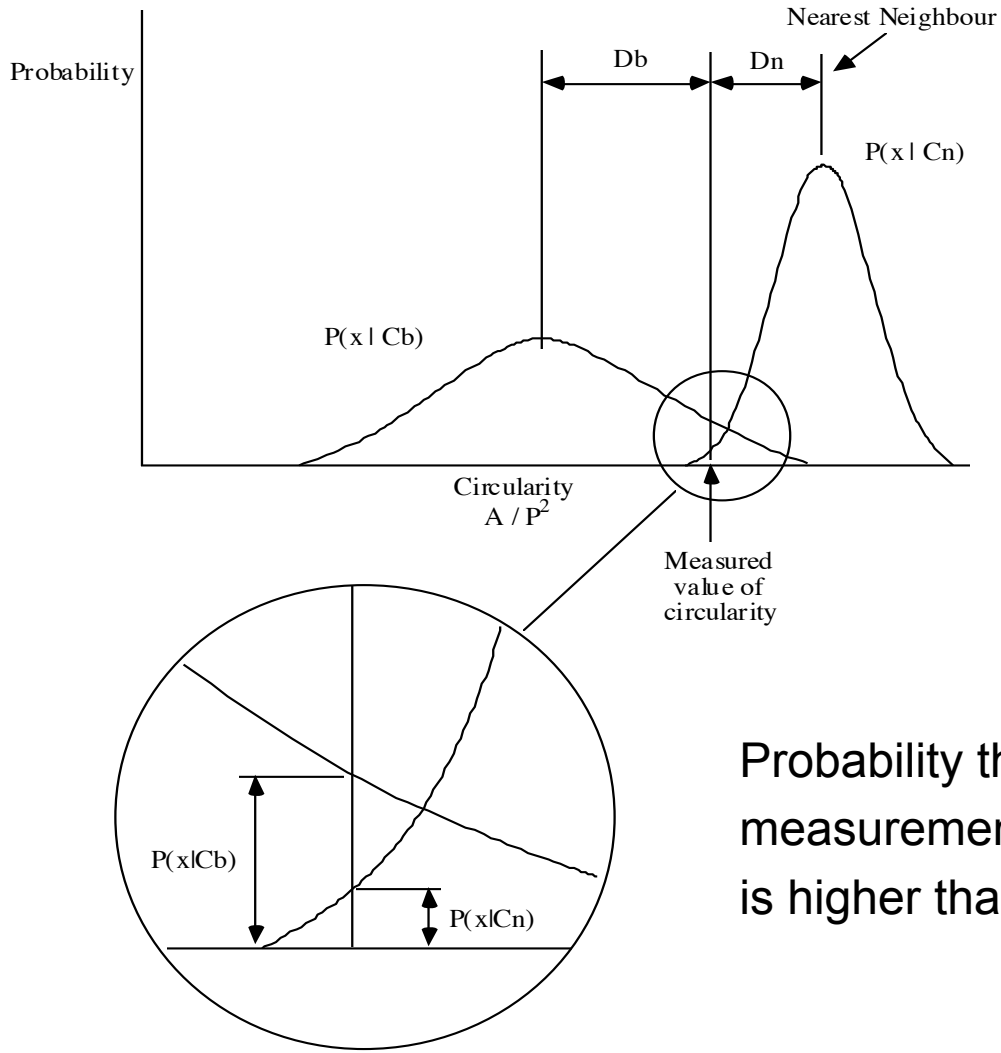
- Simplification #1: using Bayes' Theorem (see slide 53), and noting that the normalizing factor $p(x)$ **is the same** for both expressions, we can rewrite this test on posterior probabilities in terms of priors and likelihood and simplify the denominator:

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

- Simplification #2: 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 can introduce a further simplification and classify the unknown object as a bolt if :

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

Bayes' Theorem



Probability that measurement x comes from C_b is higher than it comes from C_n

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

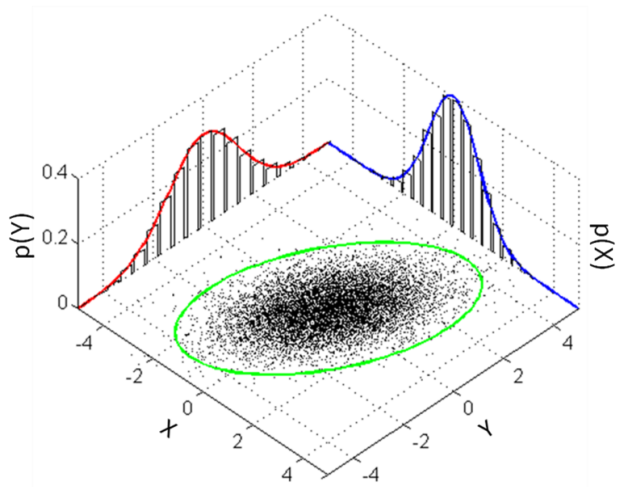
- For the example shown $p(x|C_b)$ is indeed greater than $p(x|C_n)$ for the measured value of circularity and we classify the object **as a bolt**
- If, on the other hand, we were to use the nearest neighbor classification technique, we would choose the class whose **mean value** "is closer to" the measured value
  - » In fact, in the case shown in the previous slide, 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 (do you have comments about this "erroneously"??) classify the object as a nut

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

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



Example of PDF
on a 2-D feature space

Classification using Bayes' Theorem

UNIVERSITY OF SKÖVDE

Example: Maximum Likelihood Classifier

- In the 2-D case, if we assume that the features are independent, then we can use the theory we've just outlined, multiplying together the conditional PDF for each class to calculate the joint PDF
- This approach is known as **Naïve Bayes Classifier**
  - » It may be naïve, but it works surprisingly well
  - » Even when features are not actually independent!
- If we don't (or can't) assume independence, then we need a more complex theory!