DOC493: Data Analysis and Probabilistic Inference

Duncan Gillies & Marc Diesenroth

Department of Computing, Imperial College London

Course Outline

- Lectures 1 to 10
  - Probabilistic Inference and Bayesian Networks
  - Prerequisites - Familiarity with Probability and Matrix Methods
- Lectures 11 to 18
  - Data Modeling, sampling, re-sampling, Gaussian processes.
  - Prerequisites - Linear Algebra, Mathematics for Inference and Machine Learning.

Web Resources

The course material for the first part of the course can be found on the web (linked via CATE):

www.doc.ic.ac.uk/˜dfg

Tutorial solutions will be posted in .pdf format a few days after each tutorial.

Links to other supporting material are provided, both on the web page and through CATE.

Coursework

A practical exercise implementing data modeling algorithms in Python.

- Objectives
  - To aid understanding of the material taught in the course.
  - To introduce Python which is a powerful research tool for data analysis
- The coursework is divided into four parts with hand in dates distributed throughout the term
- Obtaining a grade B is very easy, obtaining grade A needs a bit of effort, getting above 85% is possible, but harder work!

Coursework

For the first two parts of the coursework you can work in groups of up to three. If you do so make sure that you all participate fully in the work.

You don't need to work in the same group for both parts. You might want to do the first exercise on your own and the second part with a friend.

There are a few bonus marks for working on your own.

Timetable

Tuesday 16 Jan	Lecture 1	Lecture 2
Thursday 18 Jan	Lecture 3	Tutorial 1
Tuesday 23 Jan	Lecture 4	Tutorial 2
Thursday 25 Jan	Lecture 5	Lecture 6
Tuesday 30 Jan	Lecture 7	Tutorial 3
Thursday 1 Feb	Lecture 8	Tutorial 4
Tuesday 6 Feb	Lecture 9	Lecture 10
Thursday 8 Feb	Lecture 11	Tutorial 5
Tuesday 13 Feb	Lectue 12	Tutorial 6
Thursday 15 Feb	Lecture 13	Lecture 14
Tuesday 20 Feb	tba	tba
Thursday 22 Feb	tba	tba
Tuesday 27 Feb	tba	tba
Thursday 1 Mar		tba
Tuesday 13 Mar	Revision session	Revision session

Lecture 1:

Bayes Theorem and Bayesian Inference

Probability and Statistics

Statistics emerged as an important mathematical discipline in the late nineteenth and early twentieth century.

Probability is much older and has been studied as long ago as man took an interest in games of chance.

Our story starts relatively recently with the famous theorem of the Rev. Thomas Bayes, published in 1763.

Independent Events

For independent events S and D:

$$P(D\&S) = P(D) \times P(S)$$

(read “disease” for D and “symptom” for S)

Dependent Events

However in cases where S and D are not independent we must write:

$$P(D\&S) = P(D) \times P(S|D)$$

where $P(S|D)$ is the probability of the symptom given that the disease has occurred.

Bayes' Theorem

Now since conjunction is commutative:

$$P(D\&S) = P(S) \times P(D|S) = P(D) \times P(S|D)$$

and re-arranging we get:

$$P(D|S) = P(D) \times P(S|D)/P(S)$$

(Bayes' Theorem)

Bayes' Theorem as an Inference Equation

$$P(D|S) = P(D) \times P(S|D)/P(S)$$

- $P(D|S)$: The probability of the disease given the symptom is what we wish to infer.
- $P(D)$ is the probability of the disease (within a population) this is a measurable quantity.
- $P(S|D)$ is the probability of the symptom given the disease. We can measure this from the case histories of the disease.
- $P(S)$ can also be measured, but fortunately does not need to be.

Notation

Note that:

$$P(D\&S) = P(D) \times P(S)$$

is a scalar equation with two variables: S and D

For much of this course we will use discrete variables. A discrete variable can only have one of a finite number of values (or states), which we denote by lower case letters: s_1, s_2, s_3 etc.

In the simplest case a variable may take just two values (sometimes thought of as true or false), eg:

$$d_t \text{ and } d_f.$$

Normalisation

Suppose that D can take two values (or states): d_t and d_f, and S can take more states: s_1, s_2 etc. Then for any state of S, say s_i we can write:

$$P(d_t|s_i) + P(d_f|s_i) = 1$$

and by applying Bayes' Theorem we can find an expression for $P(s_i)$

$$P(s_i|d_t)P(d_t)/P(s_i) + P(s_i|d_f)P(d_f)/P(s_i) = 1$$
$$P(s_i) = P(s_i|d_t)P(d_t) + P(s_i|d_f)P(d_f)$$

Thus given values for $P(S|D)$ and $P(D)$ we can calculate $P(S)$ for any state of S. This can be done regardless of the number of states that D and S can take.

Prior and Likelihood Information

We can write $1/P(S)$ as α to remind us it is just a normalising constant:

$$P(D|S) = \alpha \times P(D) \times P(S|D)$$

- $P(D)$ is prior information, since we knew it before we made any measurements.
- $P(S|D)$ is likelihood information, since we find its value from measurement of symptoms.

Bayesian Inference (for a single hypothesis variable: eg P(D))

Given any hypothesis variable we calculate a probability distribution over its states as follows:

- Convert the prior and likelihood information to probabilities;
- Multiply them together;
- Normalise the result to get the posterior probability of the hypothesis variable (ie the probability distribution over its states) given the evidence;
- Select the most probable state.

Prior Knowledge

- In simple cases we obtain the prior probability from data. For example, we can calculate the prior probability as the number of instances of the disease divided by the number of patients presenting for treatment.
- However in many cases this is not possible - since the data isn't there.
- There may also be prior knowledge in other forms.

In general turning measured data into probabilities is done in an heuristic way.

Example from Computer Vision

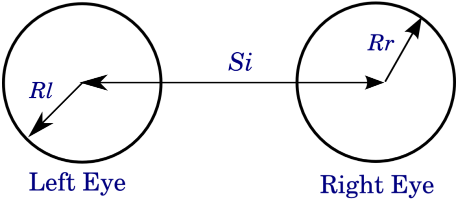
Consider a program to determine whether an image contains a picture of a cat.

Drawing by Kliban

Feature Extraction

- There are a lot of things that we could use to detect a cat, but lets start with something simple.
- We will write a program to extract circles from the image. If we find two adjacent circles of the same size we will assume that we have found cat's eyes.
- (I know that they sleep a lot, so our method isn't perfect!)

Representing prior knowledge about a cat



We can formalise our model by specifying that:

- $R_l \simeq R_r$ (the eyes are approximately the same size)
- $S_i \simeq 2(R_l + R_r)$ (the eyes are spaced correctly)

Semantic description of a Cat

To find our cat we will extract every circle from the image and record its position and radius.

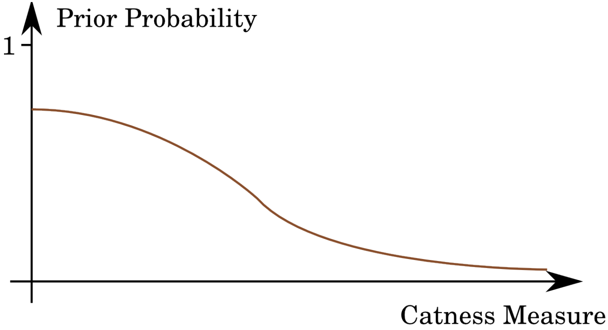
For each pair of circles we will calculate a catness measure which is:

$$Catness = |(R_l - R_r)/R_r| + |(S_i - 2 \times (R_l + R_r))/R_r|$$

The catness measure is zero for a perfect match to our model. Catness is a continuous variable.

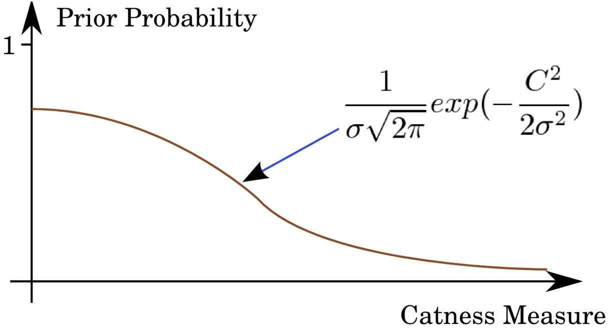
Turning measures into probabilities

We could choose some common sense way of changing our catness measure into a probability.



Turning measures into probabilities

We can do this conveniently by choosing a distribution.



Subjective Probabilities

If we choose a common sense approach to converting a measure to a probability, we are making a subjective estimate.

There may be no formal reason for choosing the distribution other than personal judgement

Objective Approach

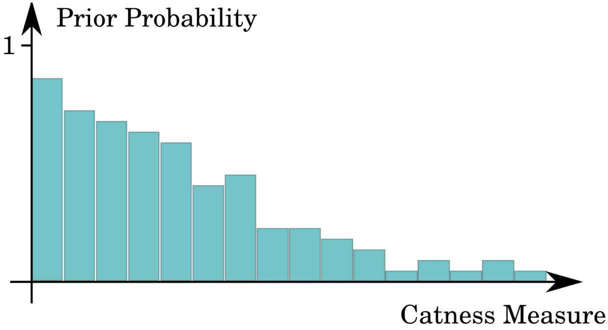
An alternative is to use data to create our probabilities. For example we could collect a large number of pictures, some including cats.

Every time we extract a pair of circles and apply a catness measure, we also get an expert to tell us whether the extracted structure does represent a cat.

For a set of histogram bins we calculate the ratio of correctly identified cats to the total.

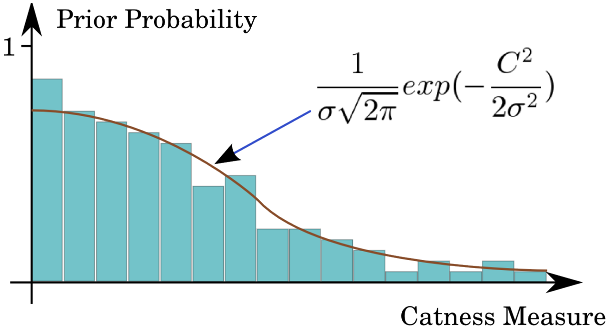
Measured Distribution

This allows us to construct a discrete distribution of the the probabilities



Measured Distribution

From which we can calculate a maximum likelihood estimate of a distribution



Subjective vs. Objective Probabilities

There is a long standing debate as to whether the subjective or the objective approach is the most appropriate.

Objective may seem more plausible at first, but does require lots of data and is prone to experimental error.

NB Some people say the Bayesian approach must be subjective. I do not subscribe to this.

Likelihood

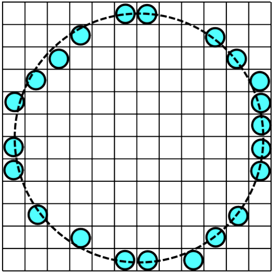
Our prior probabilities represent long standing beliefs. They can be taken to be our established knowledge of the subject in question.

When we process data, and make measurements we create likelihood information.

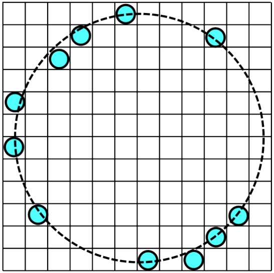
Likelihood can incorporate uncertainty in the measurement process.

Likelihood and Catness

Our computer vision process could not just extract a circle, but also tell us how good a circle it is, for example by counting the pixels that contribute to it.



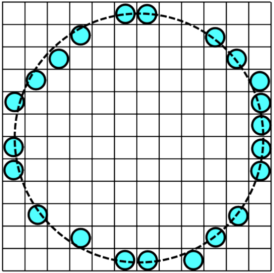
Likely Circle



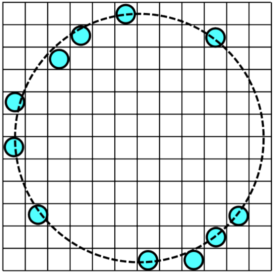
Unlikely Circle

Likelihood and Catness

- There are 31 possible circle pixels
- The first circle has 22, so Likelihood = 22/31 = 0.71
- The second circle has 11, so likelihood =11/31 = 0.35



Likely Circle



Unlikely Circle

Summary on Bayesian Inference and Cats

Bayes' Theorem for this example states that:

$$P(C|I) = \alpha \times P(C) \times P(I|C)$$

PRIOR:
$P(C)$ is the probability that two circles represent a cat, found by measuring catness and using prior knowledge to convert catness to probability.

LIKELIHOOD:
$P(I|C)$ is the probability of the image information, given that two circles represent a cat.

Prior and Subjective

- Should we use subjective or objective methods?
  (Many schools of thought exist)

- Prior information should be subjective. It represents our belief about the domain we are considering.
  (Even if data has made a substantial contribution to our belief)

Likelihood and Objective

- Likelihood information should be objective. It is a result of the data gathering from which we are going to make an inference.
- It makes some assessment of the accuracy of our data gathering
- In practice either or both forms can be subjective or objective.