The business of learning the distribution of data points.

The catch-all of learning from data.

In theory, every LFD problem is an issue of density estimation.

In practice good general density estimation is hard.

A generative approach. Answers the question "How was the data generated?"
Recap on Probabilities

- Probabilities of all events sum to one.
- Probability density: probability per unit length. Probability integrates to one.
- Sample from a distribution: pick one value with a chance proportional to the probability (density). In the long run the number of each value will be proportional to the probability.
Example from sheet. Length of Haggis. Evidence of a bimodal distribution.

Continuous variables: probability density. Integrates to 1.
Example: Discrete Distributions

- Have data for the number of characters in names of people submitting tutorial requests:
  9 10 10 11 11 11 11 11 12 12 12 12 12 12 12 12 13 13 13 13 13 13 13 13 13 14 14 14 14 14 14 14 14 14 15 15 15 15 15 16 16 16 16 16 16 16 17 17 17 17 17 17 18 18 19 19 19 19 20 20 20 20 20 20 21 21 21 21 21 21 22 22 22 22 22 22 22 22 24 24 25 25 25 25 25 25 25 25 25 25 25 25 27 27 30

- Discrete data.
- Can build a frequency table of the data.
Frequency

Amos Storkey, School of Informatics

Learning from Data: Density Estimation - Likelihood
Normalised frequency

Amos Storkey, School of Informatics
Learning from Data: Density Estimation - Likelihood
Histograms are like frequency tables for continuous variables.

Counts how many points between $a$ and $b$. Plot \( \text{area} = \text{count} \).
Framework

- Have some underlying probability distribution or density.
- This distribution is used to generate data.
- Each data point is generated independently from the same distribution.
- This is the *generative* model. It is the approach we could use to generate artificial data.
Example

Haggis again!

sample measurements

$p(x)$

$x$ (Haggis length in cm)
BUT what if we don’t know the underlying distribution. Want to *learn* a good distribution that fits the data we do have. How is *goodness* measured? Given some distribution, we can ask how likely it is to have generated the data. In other words what is the probability (density) of this particular data set given the distribution. A particular distribution explains the data better if the data is more probable under that distribution.
Likelihood

- $P(D|M)$. The probability of the data $D$ given a distribution (or model) $M$. This is called the likelihood of the model.

- This is

\[
P(D|M) = \prod_{i=1}^{N} P(x_i|M)
\]

i.e. the product of the probabilities of generating each data point individually.

- This is a result of the independence assumption.

- Try different $M$ (different distributions). Pick the $M$ with the highest likelihood $\rightarrow$ Maximum Likelihood Approach.
Boolean distribution

- Data 1 0 0 1 0 1 0 0 0 0 1 0 1 1 1 0 1.
- Three hypotheses:
  - $M = 1$: Generated from a fair coin. 1=H, 0=T
  - $M = 2$: Generated from a die throw 1=1, 0 = 2,3,4,5,6
  - $M = 3$: Generated from a double headed coin 1=H, 0=T
- Likelihood of data. Let $c=$number of ones:

$$
\prod P(x_i|M) = P(1|M)^c P(0|M)^{20-c}
$$

- $M = 1$: Likelihood is $0.5^{20} = 9.5 \times 10^{-7}$
- $M = 2$: Likelihood is $(1/6)^9 (5/6)^{11} = 1.3 \times 10^{-8}$
- $M = 3$: Likelihood is $0^9 1^{11} = 0$
Boolean distribution

- Data 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1.
- Continuous range of hypotheses: $M = k$ - Generated from a Boolean distribution with $P(1|M = k) = k$.
- Likelihood of data. Let $c$=number of ones:

$$\prod P(x_i|M = k) = k^c(1 - k)^{20-c}$$

- Maximum Likelihood hypothesis? Differentiate w.r.t. $k$ to find maximum

$$d \log P(D|M)/dk = c/k - (20 - c)/(1 - k)$$

- So $c(1 - k) - (20 - c)k = 0$. This gives $k = c/20$. Maximum likelihood is unsurprising.

- Warning: do we always believe all possible values of $k$ are equally likely?
Density and distribution estimation. Find the density from which the data was generated.

Given a density, can generate artificial independently and identically distributed (IID) data.


Given the data, and a model (a set of hypotheses - either discrete or continuous) we can find a maximum likelihood model for the data.

Next lecture: the Gaussian distribution, multivariate densities.