Algorithms for Graphical Models (AGM)

Data and probabilities

$Date: 2008/10/15 15:37:56 $

AGM-04

In this lecture

- Variables and values
- Contingency tables
- Joint probability distributions
- Maximum likelihood estimation
- Saturated models

Fictitious health data

Bronchitis	Cancer	Smoking	
absent	absent	nonsmoker	35
absent	absent	smoker	18
absent	present	nonsmoker	0
absent	present	smoker	2
present	absent	nonsmoker	15
present	absent	smoker	27
present	present	nonsmoker	0
present	present	smoker	03

A primitive database

`cancer.dat` has 3 sections:

1. `Cancer` is a *variable*, with two *values*: `present` and `absent`. Similarly `Bronchitis` and `Smoking` are variables.

2. A field header

3. The data has a count for each of the 8 possible cases. A less compact possibility is to repeat e.g. the line `absent,absent,nonsmoker` 35 times!

A contingency table

Doing:

```
>>> from gPy.Examples import cancer_table
>>> cancer_table()
```

- . . . produces a *contingency table* with 8 *cells*.

- This is really a flattened version of a three-dimensional object: one dimension for each variable.

From data to probability

- A contingency table tells us what has been observed in the past: it contains data.

- One simple way to create a *probability distribution* from data (effectively a prediction of what's likely in the future) is to find the sum of all counts (100 in this case) and divide the count for each cell by this total.

- This produces a *joint probability distribution*.

Contingency table

Bronchitis	Cancer	Smoker	
absent	absent	nonsmoker	3
absent	absent	smoker	0
absent	present	nonsmoker	27
absent	present	smoker	15
present	absent	nonsmoker	2
present	absent	smoker	0
present	present	nonsmoker	18
present	present	smoker	35

Joint probability distribution

Bronchitis	Cancer	Smoker	
absent	absent	nonsmoker	0.030000
absent	absent	smoker	0.000000
absent	present	nonsmoker	0.270000
absent	present	smoker	0.150000
present	absent	nonsmoker	0.020000
present	absent	smoker	0.000000
present	present	nonsmoker	0.180000
present	present	smoker	0.350000

Joint probability distribution

- It is a *distribution* because a ‘probability mass’ of 1 has been *distributed* over the 8 cells.

- It is a *probability* distribution because each individual number is a probability.

- It is a *joint* probability distribution because each probability corresponds to a joint instantantiation of the 3 variables.

Maximum likelihood estimation (1)

- We have just seen an example of *maximum likelihood estimation (MLE)*.

- It is *estimation* since the distribution it produces is an estimate of some unknown true distribution.

Maximum likelihood estimation (2)

- Putting aside the joint structure of our distribution, our MLE distribution defines a probability distribution with 8 possible outcomes, like throwing a (biassed) 8-sided dice.

- Given a fixed data size, say 100, it defines a *multinomial distribution* over all possible datasets of that size. For example, it gives the probability for our original data as $\approx 7.510472 \times 10^{-5}$.

- Adopting a multinomial distribution is tantamount to assuming that each datapoint is independently 'drawn' from our probability distribution.

The calculation

Just for the record

$$
\begin{aligned}
& P(3, 0, 27, 15, 2, 0, 18, 35) \\
= \; & \frac{100!}{3!, 0!, 27!, 15!, 2!, 0!, 18!, 35!} \times \\
& 0.03^3 0^0 0.27^{27} \times \\
& 0.15^{15} 0.02^2 0^0 \times \\
& 0.18^{18} 0.35^{35} \\
\approx \; & 7.510472 \times 10^{-5}
\end{aligned}
$$

Maximum likelihood estimation (3)

- The probability of observed data (according to some distribution) is known as the *likelihood* of that data. Here *likelihood* is being used in a specific technical sense.

- Our MLE distribution is the distribution that maximises the likelihood of the data (just trust me). Hence the name.

- It is a reasonable way of estimating distributions, particularly when there is lots of data.

A saturated model

- A *probabilistic model* imposes structural constraints on what the ‘true’ probability distribution is.

- A graphical model is just one type of probabilistic model.

- Formally, a model is just a set of probability distributions.

- A *saturated model* is a special case where there are no constraints.

- So formally it is the set of all possible probability distributions for a given collection of variables.

MLE for a saturated model

- Let there be n datapoints in total, and $n(i)$ which fall into cell i.

- The MLE distribution is defined by a probability for each cell.

- Let $p(i)$ be the unknown true probability for cell i.

- MLE gives us $\widehat{p}(i) = n(i)/n$ as the estimate for $p(i)$ for all values of i.