Maths Skills (MTCS) G5071

Lecture 10

Kingsley Sage
Room 5C16, Pevensey III
khs20@sussex.ac.uk

Lecture 10

- Introduction to data analysis
  - Correlation
  - Histograms
  - Hypothesis testing

Introduction to data analysis

- In traditional AI, it has been common for researchers to make their points by building systems that illustrate particular techniques or demonstrate particular competencies – like engineering.
- Increasingly, there is a kind of investigation that demands a different approach, more like that of a behavioural scientist.
- This occurs particularly when systems cease to be transparent. Then it's not enough to just build a system, its properties must be explored.

Introduction to data analysis

- In artificial life and evolutionary simulation, systems involve random numbers to mimic environmental variability, and robotic systems are subject to the genuine thing.
- Characterising systems that involve variability involves the use of statistical methods.
- Various kinds of descriptive statistics are useful in the exploration of a system and are the main method of trying to obtain some degree of understanding of it.

Graphs and numerical statistics

- When looking at the results from an experiment, the first set of tools to turn to are graphical ones.
- Tools such as MATLAB provide a wide variety of ways to display data graphically – the area is large and complex but methods such as 2D and 3D graphs and bar charts are probably familiar already.
- When data has multiple dimensions, it can be difficult to find the appropriate combinations to display.
- It is essential to spend time finding the right way to display data in order to reveal relationships that may be present.

Graphs and numerical statistics

- The mean and standard deviation of a set of data have already been discussed.
- Calculating these statistics for results obtained when as experiment is repeated is often the first step in gaining a clear view of what is going on.
- The mean gives the the measure of the location of the centre of some numerical data and the standard deviation gives a measure of its spread.

Correlation

- An additional descriptive statistic, the correlation coefficient between two sets of data, can be used when the individual data values can be paired off between the two sets.
- This gives a measure of whether the two random variables being sampled vary together or are independent.
- For example, in weather forecasting we might be interested in whether there is a correlation between wind speed and rainfall …

Correlation

- If two random variables *X* and *Y* are being sampled, the correlation coefficient is defined as:

$$r = \frac{\left\langle (X - \langle X \rangle)(Y - \langle Y \rangle) \right\rangle}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}$$

- That is, it is the average of the products of deviations of the variables from their means, normalised using their variances.
- Coefficient lies between –1 and +1, and either –1 or +1 means that there is a perfect linear relationship between the two variables, 0 corresponds to no linear relationship between the two (de-correlated).

Histogram

- A graphical way of looking at a distribution is to use a histogram of the values found.
- The simplest way to produce a histogram is to create a set of bins covering the range of values of the variable – each bin initially contains the value zero.
- After each trial, the value of the variable being measured is used to pick out a bin, and the value held in the bin is incremented.
- For example, in a simple case, a measure might range from 0 to 99 – we create 10 bins covering the ranges 0-9, 10-19, 20-29 and so on.
- If a trial yields the value 63, we increment the 60-69 bin, and after a large enough number of trials, values in the bins can be used to provide an approximation of the underlying probability distribution of the variable.

Histogram

- We can use this histogram approach to model both discrete and continuous distributions.
- There is a trade off between the number of bins and the accuracy of the probability estimate each one represents – a lot of bins gives a narrow range of values for each, giving a higher accuracy on the position of any feature of the distribution, but lower accuracy on the probability estimates because fewer votes will be cast for each bin.

Hypothesis testing

- Suppose you run a simulation and measure the outcome – say average level of fitness in a population after a certain number of generations or the number of times a robot succeeds in reaching its goal.
- You then make some adjustment, perhaps by varying a parameter of the simulation such as the mutation rate of a genetic algorithm or rate of learning of a neural network, and repeat the simulation.
- If the outcomes changes, how can you say whether this was a result of the adjustment you made, or simply a random fluctuation which might have been expected to occur regardless?

Hypothesis testing

- This kind of question if at the heart of the dominant statistical methodology of the behavioural, social and medical sciences.
- The question might be whether a new drug has an effect on the outcome of a particular disease.
- The method generally used is called hypothesis testing – the approach is to ask whether it is reasonable to attribute any differences observed to random fluctuations, assuming that the manipulation has no effect.
- If the changes are too big for this to be reasonable, then there is evidence for a real effect due to manipulation in the experiment.

Formal framework

- The null hypothesis H_0 is the hypothesis that the difference in conditions between runs of an experiment have no effect.
- The alternative hypothesis H_1 is that H_0 is false I.e. there is an effect of manipulation.
- If we decide the experiment shows an effect when in fact there is none, we have a *Type1* error, and conversely if we decide that there is no evidence for the effect when in fact one exists, we have a *Type 2* error.

Formal framework

- Usually, differences between experimental results are summarised in a single statistic – say the change in a success rate of a robot, or the effectiveness of a drug.
- We then calculate the probability, assuming the null hypothesis, of getting the observed value of the statistic, or a more extreme value.
- This probability is always given the symbol *P* and is known as the significance level.
- If *P* is low, then the result we have is unlikely under the null hypothesis and it is likely we have a real effect.

Simple example

- Suppose we experiment with a simulation of a system by setting a pseudo-random number generator to a particular seed before we start, and using the value *a* for some parameter we are interested in.
- We then change the parameter to *b* and reset the random number generator to the same seed as before and rerun the experiment.
- We look to see whether the performance is better or worse than it was before and repeat the pair of tests some number of times – say 10.
- We record for each pair of tests whether the performance increased or decreased when the parameter was changed from *a* to *b* with different random numbers in each pair of tests.

Simple example

- Suppose performance gets better on 8 trials out of 10, and worse on 2 trials – how likely is this under the null hypothesis that changing the parameter from *a* to *b* produces no improvement in the performance?
- It is like flipping a coin, there are 2^{10}=1024 different equally likely ways the experiment can turn out.
- In 1 of these, performance will improve on all 10 trials (from TTTTTTTTTT to HHHHHHHHHH), in 10 of them the performance will improve on 9 trials (e.g. from HTTTTTTTTT to HHHHHHHHHH), and in 45 trials the performance will improved in 8 trials. There are 1+ 10 + 45 = 56 cases that give the observed result of more extreme.

Simple example

- P=56/1024 or about 0.055 I.e. on about 55 in 1000 repetitions of the whole sequence of trials, you would expect to get 8 or more improvements, just by random fluctuations in the total.
- We can also ask whether a result of 2 of fewer improvements out of 10 would not be just as “extreme” as a result of 8 improvements.
- This depends on whether we want to test that the change had effect of some sort, or whether we want to test that it specifically produced improvement.
- The concept of the two tailed test and the one tailed test.

Simple example

- Does the result P=0.055 obtained in our imaginary experiment mean that changing the parameter produces an improvement or not?
- The received wisdom of how to answer this is as follows – before doing the experiment decide on the critical value of P and denote it α.
- If P turns out to be less than α, you reject the null hypothesis I.e. accept the existence of the effect, otherwise you accept the null hypothesis.

Next time …

- Data visualisation …