Maths Skills (MTCS) G5071

Lecture 8

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

Lecture 8

- Introduction to probability and statistics
  - Statistical independence
  - Normal and uniform distributions
  - Bayes' theorem
  - Variance and standard deviation

Probability and statistics

- A probability is a number between 0 and 1 and we can manipulate it according to certain well established and well understood rules.
- The probability of an event E is written P(E) or sometimes as Pr(E).
- For example, if H is the event that a tossed coin turns up heads, and the coin is fair, then we can write P(H)=0.5.

Probability and statistics

- If two events are mutually exclusive I.e. they cannot occur at the same time, then P(A+B) = P(A)+P(B).
- For example, if P(H) = 0.5 and P(T)=0.5 and H and T cannot both occur, because they stand for heads and tails, then P(H or T)=1 (excluding the possibility that the coin might land on its side).
- If a set of events is complete, I.e. one of the events in the set must occur, and all the events are mutually exclusive, then their probabilities sum to 1.
- If events X_1 to X_N are complete and mutually exclusive then:

$$\sum_i P(X_i) = 1$$

Probability theory and estimation

- If A and B are two events, then :
  - P(A or B)=P(A)+P(B) – P(A and B) where “A and B” means that both events happen.
- Imagine that genetic code strings of length 5 are generated by selecting random characters G,A,T and C with equal probability. What is the probability of combination AAAA somewhere in string?
  - Space of all possible strings (AAAAA, AAAAG, AAAAC etc all the way to CCCCC) – each has the same probability and there are 4*4*4*4*4 of them – so approximately 0.001.
  - 4 strings have AAAA at left and 4 have AAAA at right end but AAAAA is common to both so total containing AAAA is 7 – so probability required is about 0.007.
- Another way is the Monte Carlo method, generating strings at random and estimating the probability by counting those that have AAAA.

Conditional probability and independence

- Conditional probability is the probability of one event given that another event is known to have occurred.
- For example, for an autonomous robot, consider the probability that there is a particular object – say an apple – in front of its camera, given that the robot sees a particular image feature – say a circular shape.
- If A is the event that an apple is present and C is an event that a circular shape has been detected, we can write P(A|C) (probability of A given C).
- We can think about this as the “possible worlds” model of probabilities.

Conditional probability and independence

- On the other hand, consider another event B – knowing the colour of the tree.
- Knowing C does not tell the robot anything about the likelihood of B. So here P(B|C)=P(B). We say that B is statistically independent of C.
- We can express such independence numerically – if B and C are independent, the the probability of both events occurring is given by P(B and C) = P(B)*P(C).

Bayes' Theorem

- Suppose our robot "knows" that:
  - apples happen to be in front of the camera on one-tenth of the occasions that is looks, so P(A)=0.1, and
  - that it detects a circular shape in the image on one-fifth of the occasions that it looks, so P(C)=0.2, and
  - from previous experience, it detects a circular shape three-quarters of the time when an apple is in front of it so P(C|A)=0.75.
- The robot detects a circular shape, what is the chance that an apple is indeed in front of it: P(A|C) ?
  - P(A and C)=P(A)P(C|A)=0.1*0.75=0.075
  - P(A|C)=P(A and C) / P(C) = 0.075 / 0.2 = 0.375
- We say that the evidence of the visible circle has increased probability of hypothesis that an apple is in front of the camera from 0.1 to 0.375

Quick summary & Bayes' Theorem

$$P(A \vee B) = P(A) + P(B) \text{ for mutually exclusive events}$$

$$P(A \,\&\, B) = P(A).P(B) \text{ for statistically independent events}$$

$$P(A \mid B) = P(A) \text{ for statistically independent events}$$

$$P(A \mid B) \text{ reflects conditional dependence}$$

$$P(A \,\&\, B) = P(A \mid B).P(B) \text{ for conditionally dependent events}$$

$$P(A \,\&\, B) = P(B \,\&\, A) \text{, so}$$

$$P(A \mid B).P(B) = P(B \mid A).P(A)$$

$$P(A \mid B) = \frac{P(B \mid A).P(A)}{P(B)} \text{ which is Bayes' Theorem}$$

Random variables and probability distributions

- If a random variable X can take on any of a continuous set of values, for example any value in the range 0 to 1, then we have a continuous distribution.
- Assignment of probabilities needs some extra formalism – we can handle this by using probability that X is less than some particular value x, we write P(X<x).
- This probability is a a function of x – we can write it as F(x) and we call if the cumulative probability distribution.

Random variables and probability distributions

- Another approach is the consider the probability that X lies within a range of values $P(X \geq x \text{ and } X < x+\delta x) = f(x)\delta x$.
- The function f(x) is called a probability density function – the same relationship to probability that density of a substance does to mass.
- We multiply density by volume to get mass, and so we can multiply probability density by the size of part of the sample space to get probability.
- Relationship between cumulative distribution and probability density is:

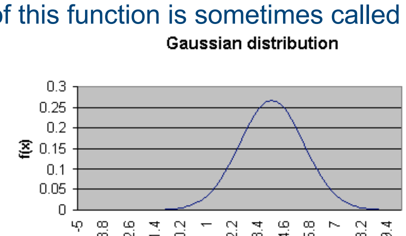
$$f(x) = \frac{d}{dx} F(x)$$

Gaussian or normal distribution

- Probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

- where μ and σ are called mean and standard deviation respectively of the distribution.
- The graph of this function is sometimes called the bell-shaped curve:



Statistics of distributions

- It's often necessary to summarise a distribution of random variables using a few numbers called its parameters.
- Quantities that help describe distributions and that are calculated from data are called statistics and the process of estimating parameters from statistics is called statistical inference.

Statistics of distributions

- One of the most important descriptors of a distribution is its mean.
- The mean or average can be obtained by adding together a set of observed numbers and dividing by the number of observations. In the case of a Gaussian distribution, this statistic is a good estimator of the parameter μ and for any discrete distribution with a finite number of values of the variable, mean is:

$$\sum_i x_i P(X = x_i)$$

Statistics of distributions

- Thus if we average the observation values, adding and dividing them by N, we expect to get the approximate results:

$$\frac{\sum_i x_i N P(x_i)}{N}$$

- The mean of this distribution is often written <X>

Variances and standard deviations

- The mean of a distribution says something about where its centre is. The next most useful thing to know is how spread the distribution is.
- In practice, squaring differences between values and the mean not only avoids negative numbers. but makes various calculations simpler.
- The average squared distance from the mean is called the variance, and its formula is given by:

$$\mathrm{var}(x) = \sum_i (x_i - < X >)^2 P(x_i)$$

Variances and standard deviations

- The square root of the variance is called standard deviation and if the sum shown in the previous slide is applied to the Gaussian distribution, using integration, then standard deviation is just the parameter σ.
- Finally the smoothness of the distribution is given by its entropy (used in neural networks and in information theory generally):

$$entropy(x) = -\sum_i (\log P(x_i))P(x_i)$$

Next time …

- Applied probability and statistics …