FIT2086 Lecture 1 Summary
Introduction/Probability and Random Variables

Dr. Daniel F. Schmidt (with material from Lachlan O'Neill)

August 13, 2019

1 Part I: Introduction

Data science. The science of analysing data and learning models to describe

Classifiers. A classifier guesses which of a set of groups a given object belongs to. For example, you might want to determine whether a person is likely to contract a disease such as cancer, given some other information about them like their age, height, sex and typical exercise levels. In this case you would be looking to categorise people into a small discrete number of states, such as "no cancer' vs. "cancer".

Regression. A regression is similar to a classifier. The main difference is that we are now guessing a continuous property rather than a discrete one. For example, YouTube might try to guess how long you are going to stay on the website for (e.g. one user might only watch videos for 20 minutes, and another might watch them for 4 hours).

Clustering. Classification and regression are both forms of "supervised" learning. Clustering is a form of "unsupervised" learning, because we don't necessarily want to guess "a value" - we just want to find useful information hidden in the data. In a clustering system, we have a set of "unlabelled" information (information that is not in groups) and we wish to create groups based on the data. As an example, we might have genetic information for some number of people, and wish to group them based on similarities or shared attributes (e.g. to find shared ancestry). This would be done through clustering.

Recommender Systems. Recommendation systems are also a type of unsupervised learning technique. Rather than trying to guess whether a person likes comedies or not (that would be a classifier) we instead utilise the information we have a user's past activities (watching comedies, horror movies, etc.) to predict what other things they might. For example, YouTube might want to suggest certain videos based on videos that you have watched in the past. How it discovers what those certain videos are is another question - they might use a clustering algorithm (clustering users into groups based on what they've watched and suggesting videos that others within that group have liked - we don't have any defined groups but instead look for "clusters" of similar users - the clustering algorithm - and then go on).

Forecasting. Forecasting is similar to classification and regression in some ways, but instead of predicting one value the aim is to predicting a "change" in value, or a sequence of values many steps into the future. For example, you might forecast a stock price (based on prior stock price data and other information) to predict whether it's going to increase or decrease tomorrow, or what it's long-term trend might be. You could implement this using a classifier (should you "buy", "hold" or "sell" a stock given some information about it?) but the point is to determine what will happen in the future as accurately as possible, given what has happened in the past. That's what makes it a forecaster!

Anomaly Detection. Lastly, anomaly detection is the analysis of repeated events to determine when something is "out of the ordinary" (i.e., anomalous). For example, one recently developed product is the "smart cane" which older people can use - it tracks their usual daily activity and, for example, if it notices they're not moving much today, might alert someone that they've fallen over or become unconscious. As another example, a credit card company might monitor a user's typical transactions (e.g. usually in the eastern Melbourne area, spends about $25 per day, mostly at around 8:30am and 12:30pm - morning coffee and lunch, respectively, but the bank doesn't know that - it just sees patterns!). Then, if one day the bank notices several $300 transactions up in Queensland at midnight, it might detect this activity as anomalous.

Population, sample, model. A population is a large collection of objects, or items, with measurable traits that we wish to model or learn about. We usually assume that the population is infinitely large, at least relative to the size of the sample we can take. A sample is a finite number of recordings of attributes taken ("sampled") from the population, usually much smaller in size than the population. The size is often constrained through the costs of data collection. This sample will be used as a surrogate for the population when we build our model. A model is a mathematical or algorithmic description of the population, usually "learned" or inferred from the sample. No model is correct, but some are more useful than others (i.e., they capture different aspects of the population more accurately). We often use models to make predictions or statements about likelihood of events occuring within our population.

Types of data. There are four general classes of data types found in real world datasets:

1. Categorical-nominal. This type of data has a discrete, finite number of values, with no inherent ordering between the categories; for example, sex and country of birth.

2. Categorical-ordinal. This type of data has a discrete, finite number of values, but with an inherent ordering; for example, education status (primary school, high school, tertiary, postgrad) or state of disease progression.

3. Numeric-discrete. Numeric data, but the values are enumerable (i.e., integers, or non-negative integers, etc.). Examples include number of live births or age measured in whole years.

4. Numeric-continuous. Numeric data, but the values are not enumerable (i.e., they are continuous, real numbers). Examples include weight, height, distance from CBD, etc.

2 Part I: Random Variables and Probability Distributions

Probability distribution over a random variable. We say X is a random variable (RV) if it takes on values from a set of possible values $\mathcal{X}$ with specified probabilities. We sometimes call $\mathcal{X}$ the event space, and seeing any x from $\mathcal{X}$ as observing the event $X = x$. We use the language

$$\mathbb{P}(X = x),\ x \in \mathcal{X}$$

to describe the probability that the RV X will take on the value x from $\mathcal{X}$. A probability distribution satisfies two important properties:

$$\begin{aligned} \mathbb{P}(X = x) &\in [0, 1], \\ \sum_x \mathbb{P}(X = x) &= 1, \end{aligned}$$

which says that the probability of any event lies between zero and one, and the total probability over all the possible values $\mathcal{X}$ of x is always equal to one. Another important property is the additivity of the probability of mutually exclusive events. What this means is that the probability of (X taking on values from some set A) OR (X taking on values from another set B) is equal to

$$\mathbb{P}(X \in A \text{ or } X \in B) = \mathbb{P}(X \in A) + \mathbb{P}(X \in B)$$

if A and B have no values in common.

Probability distributions over multiple variables. Let X and Y be random variables over some sets $\mathcal{X}$ and $\mathcal{Y}$. We define

$$P(X = x, Y = y), \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

as the joint probability of $X = x$ and $Y = y$; that is, the probability of both X taking on the specific value x and Y taking on the specific value y at the same time. The marginal probability of $X = x$ is given by

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y).$$

which is the probability of seeing $X = x$ irrespective of the value Y takes on, and the conditional probability is given by

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

which is the probability of observing $X = x$ if we know that $Y = y$.

Independent random variables. Independent random variables play a very important role in probability, because they greatly simplify calculations. Two RVs X and Y are considered independent if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$$

for all values of x and y. In words this says that if X and Y are independent, the joint probability of X taking on the value x and Y taking on the value y is equal to the product of the marginal probabilities of $X = x$ and $Y = y$. An important implication of independence is that

$$\mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x).$$

In words this says that knowing the value of Y tells us no new information about what value X may take on. A particularly important sub-class of independent RVs are the independent and identically distributed (i.i.d.) RVs; X_1 and X_2 are i.i.d. if they are independent and also if

$$\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x)$$

for all values of $x \in \mathcal{X}$. In words this says the two RVs have exactly the same marginal distribution over $\mathcal{X}$, and are independent.

Probability density functions. If the set of values $\mathcal{X}$ that a random variable X can take is continuous (i.e., real numbers), then we say that X follows a probability density function (pdf). A pdf satisfies

$$p(x) \geq 0 \text{ for all } x \in \mathcal{X}$$

and

$$\int_{\mathcal{X}} p(x)dx = 1.$$

The term "density function" is used because it describes how densely the probability is distributed across the set $\mathcal{X}$. To find the probability of an interval such as $X \in (a, b)$ we integrate the pdf from a to b:

$$\mathbb{P}(a < X < b) = \int_a^b p(x)dx.$$

One of the more confusing properties of continuous variables is that the probability of X taking on any specific, exact real number is zero. Why is this? Consider the probability of $X \in (x_0 - \delta/2, x_0 + \delta/2)$, where $\delta > 0$ is the width of the region around a value x_0. Then, we can approximate the integral as the product of the width of the region δ times the height of the pdf at the point $p(x_0)$:

$$\begin{aligned} \mathbb{P}(x_0 - \delta/2 < X < x_0 + \delta/2) &= \int_{x_0-\delta/2}^{x_0+\delta/2} p(x)dx \\ &\approx p(x_0)\delta \end{aligned}$$

From this it is clear that the smaller the interval δ around a point x_0, the smaller the probability; taking $\delta \to 0$ clearly shows that $\mathbb{P}(X = x_0) = 0$.

Cumulative distribution functions. The cumulative distribution function (cdf) plays an important role in probability theory. Let us begin by introducing some shorthand notation for discrete RVs; namely, that

$$\mathbb{P}(X = x) \equiv p(x)$$

where $a \equiv b$ is read as "a is equivalent to b". The cdf of a discrete random variable over the integers is then

$$\mathbb{P}(X \leq x) = \sum_{x' \leq x} p(x')$$

which can be interpreted as the probability of X taking on any value less than or equal to x. For a continuous RV the cdf is

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(x')dx'.$$

From the fact that the total probability of a distribution is one, we have the following useful properties:

$$\begin{aligned} \mathbb{P}(X > x) &= 1 - \mathbb{P}(X \leq x) \\ \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) \end{aligned}$$

With these two rules the cdf can be used to obtain probabilities over any interval or combination of intervals.

The inverse of the cdf is called the quantile function $Q(p)$. This function takes as an argument a value from zero to one, say p, and returns the value x such that the probability of X being less than $Q(p)$ is equal to p. Formally we write

$$Q(p) = \{x \in \mathcal{X} : \mathbb{P}(X \leq x) = p\}$$

which we read as "find the value of x in the set $\mathcal{X}$ such that $\mathbb{P}(X \leq x)$ is equal to p". The quantile function is frequently used in statistics; one of its uses is to define quantities such as the median, $Q(p = 1/2)$, or the first and third quartiles, $Q(p = 1/4)$ and $Q(p = 3/4)$.