FIT2086 Lecture 1 Summary
Introduction and Descriptive Statistics

Dr. Daniel F. Schmidt and Lachlan O'Neill

August 9, 2017

1 Part I: Introduction

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 dont 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 theyve watched and suggesting videos that others within that group have liked - we dont 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 its going to increase or decrease tomorrow, or what its 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. Thats 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 theyre not moving much today, might alert someone that theyve fallen over or become unconscious. As another example, a credit card company might monitor a users 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 doesnt 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 II: Descriptive Statistics

Statistics. What is a statistic? Technically speaking, a statistic is any function of a data sample. Some statistics are more useful than others, as they describe different aspects of the data.

Measures of centrality. These statistics attempt to give an idea of what the typical, or "average" value in a sample is – using different definitions of what it means to be average. There are three common measures of centrality used in statistics. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be our sample of n datapoints.

The most common measure of centrality is the sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The sample mean is particularly appropriate for numeric data, especially if the data does not contain any outliers (values much larger than the bulk of the sample). The sample mode is another commonly used measure of "average value", and is defined as the most commonly occurring value in our sample. For continuous data this is not a very good measure of typical value, as most almost every value in our sample will likely be unique. This is an appropriate measure for categorical data. The sample median is the third common measure of centrality, and is defined as the value $\mathrm{med}(\mathbf{y})$ such that half of the sample $\mathbf{y}$ has values less than $\mathrm{med}(\mathbf{y})$, and half of the sample has values greater than $\mathrm{med}(\mathbf{y})$. We can find this by sorting our data and taking the middle value. The sample median is appropriate for categorical-ordinal data, and for numeric data, particularly if the sample has several outliers.

Percentiles. The p-th sample percentile is the value $Q(\mathbf{y}, p)$ such that $p\%$ of the values of the sample are lower than $Q(\mathbf{y}, p)$. The sample median is given by $Q(\mathbf{y}, p)$. Other important percentiles are the 25-th and 75-th percentiles (also known as the 1st and 3rd quartiles).

Measures of spread. Measures of centrality tell us about the average value of the sample. The measures of spread instead tell us how much the values in the sample differ, on average, from the average, or typical value. The larger the spread, the more variable the values in the sample are. The most basic measure is the range, which is defined by

$$\mathrm{rng}(\mathbf{y}) = \sup\{\mathbf{y}\} - \inf\{\mathbf{y}\}$$

where $\sup\{\mathbf{y}\}$ denotes the largest value in $\mathbf{y}$, and $\inf\{\mathbf{y}\}$ denotes the smallest value in the sample $\mathbf{y}$. This is literally just a measure of the range of values that appear in the sample. The most common measure of spread is the sample standard deviation:

$$s(\mathbf{y}) = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2}.$$

The sample standard deviation is defined as square-root of the mean squared deviation of the datapoints in the sample, from the sample mean. The sample standard deviation has the same units of measurement as the data itself, and tells you how far, on average you expect a random datapoint from your sample to be from the sample mean. Sometimes the sample variance, $V(\mathbf{y}) = s^2(\mathbf{y})$ is preferred to the sample standard deviation for reasons of mathematical convenience.

Correlation coefficients. Imagine now our sample consists of two numeric measurements, $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ made on the same objects. The correlation coefficient is a measure of the association between the two:

$$R(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n s(\mathbf{x}) s(\mathbf{y})}.$$

The division by the standard deviations removes the unit of measurement from the coefficient correlation ("standardises" it) so that the value always lies between -1 (perfect negative correlation) and 1 (perfect positive correlation), with a correlation of zero denoting no correlation. Correlation can be interpreted as follows:

- A positive correlation implies that if a x_i is greater than the sample mean $\bar{x}$, then we expect the corresponding value y_i will be more likely to be greater than the sample mean $\bar{y}$.
- Conversely, negative correlation implies that if a x_i is greater than the sample mean $\bar{x}$, then we expect the corresponding value y_i will be more likely to be smaller than the sample mean $\bar{y}$.

As correlation measures the *linear* association between the two variables, it is very possible that two variables will be uncorrelated (correlation of zero, or small correlation) but strongly associated through a non-linear relationship.