STATISTICS CONCEPTS USED IN THE DOMAIN OF DATA SCIENCE AND MACHINE LEARNING

XAVIER TANG

Ver 0.1

Contents

Chapter 1

Distribution

This chapter deals with concepts mainly related to various probability distribution.

1.1 Probability Mass Function (PMF)

The Probability Mass Function (PMF) gives the set of probabilities of discrete outcome, e.g. discrete uniform PMF: roll one dice, each outcome is 1/6.

More formally, a probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value. In Eqn. (1.3) below, the PMF gives the probability of getting exactly k (discrete) successful Bernoulli trails.

Formal definition:
PMF is the probability distribution of a discrete random variable, and provides the possible values and associated probabilities. The probabilities associated with each possible values must be positive and sum up to 1. For all other values, the probabilities need to be 0.

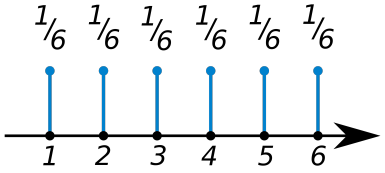

Figure 1.1: The probability mass function of a fair die. All the values of this function must be non-negative and sum up to 1

1.2 Probability Density Function (PDF)

The Probability Density Function (PDF) of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.

In a more precise sense, the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variable's PDF over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to 1. A PDF must be integrated over an interval to yield a probability, which is different from PMF (other than continuous VS discrete random variable).

More formally, the PDF is most commonly associated with absolutely continuous univariate distribution, a random variable X has PDF f_X and the probability of this variable taking values between a and b, i.e. $a \leq X \leq b$ will be:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx \tag{1.1}$$

If F_X is the cumulative distribution function (CDF) of X, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{1.2}$$

1.3 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of a real-valued random variable X (continuous or discrete), evaluated at x, is the probability that X will take a value less than or equal to

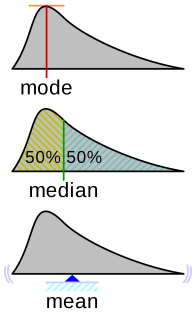

Figure 1.2: Geometric visualisation of the mode, median and mean of an arbitrary probability density function.

x. In the case of a scalar continuous distribution, it gives the area under the probability density function (PDF) from $-\infty$ to x.

1.4 Bernoulli Trials

This is a random experiment with exactly 2 possible outcomes. The probability of success is the same every time experiment is conducted. A similar analogy: Flipping a (possibly) biased coin, each coin has probability p of landing heads (success) and probability $1-p$ of landing tails (failure).

Closely related to a Bernoulli trial is a binomial experiment, which consists of a fixed number n of statistically independent Bernoulli trails, each with a probability of success p, and counts the number of success. The number k of success in n Bernoulli trials is Binomially distributed. The probability of exactly k success (out of n) is given by the probability mass function (PMF):

$$f(k,n,p) = P(k;n.p) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.3}$$

where $\binom{n}{k}$ is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.4}$$

Example:

Let's say a bank made 100 mortgage loans. It is possible that anywhere between 0 and 100 of the loans will be defaulted upon. You would like to know the probability of getting a given number of defaults, given that the probability of a default is $p = 0.05$. To investigate this, you will do a simulation. You will perform 100 Bernoulli trials. Here, a success is a default. (Remember that the word 'success' just means that the Bernoulli trial evaluates to be True, i.e., did the loan recipient default?) You will do this for another 100 Bernoulli trials. And again and again until we have tried it 1000 times. Then, you will plot a histogram describing the probability of the number of defaults. So we have performed 1000 times of 100 mortgage trials, the histogram should show a maximum at about 5 (because 0.05×100) and probability is given in Eqn. (1.3).

Example:

Consider the simple experiment where a fair coin is tossed four times. Find the probability that exactly two of the tosses result in heads.

$$\begin{aligned} P(2) &= \binom{4}{2} p^2 (1-p)^{4-2} \\ &= 6 \times (0.5)^2 \times (0.5)^2 \\ &= \frac{3}{8} \end{aligned}$$

When multiple Bernoulli trials are performed, each with its own probability of success, these are sometimes referred to as Poisson trials.

1.5 Discrete variable: Binomial Distribution, Bin(n, p)

The Binomial Distribution with parameters n and p, denoted Bin(n, p) is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own boolean-valued outcome True (with

probability p), or failure (with probability $q = 1 - p$). A single success or failure experiment is also called a Bernoulli trial.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, then the draws are not independent and so the resulting distribution would be a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used.

The PMF of Binomial Distribution is given in Eqn.(1.3). The CDF of Binomial Distribution can be expressed as:

$$F(k; n, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{1.5}$$

where we just add up all the probability for all the previous k values.

1.6 Discrete variable: Bernoulli distribution, Ber(p)

The Bernoulli distribution is a special case of the binomial distribution, where a single trial is conducted (i.e., a binomial distribution with $n = 1$). The PMF of Bernoulli distribution, over possible outcomes j is:

$$f(k; p) = \begin{cases} p, \text{if } j = 1 \\ q = 1 - p, \text{if } j = 0 \end{cases} \tag{1.6}$$

1.7 Discrete variable: Geometric Distribution, Geo(p)

The Geometric Distribution is either of two discrete probability distribution described below:

1) The probability distribution of the number X of Bernoulli Trials needed to get one success, supported on the set {1, 2, 3, ... }.

This geometric distribution, sometimes denoted $Geo(p)$, gives the probability that the first occurrence of success requires k independent trials, each with success probability (Bernoulli Trial). The Probability Mass Function (PMF) is (probability that the kth trial (out of k trials) is the first success):

$$P(X = k) = (1 - p)^{k-1}p, k = 1, 2, 3, ... \tag{1.7}$$

and the Cumulative Distribution Function (CDF) is:

$$P(X \leq k) = 1 - (1 - p)^k \tag{1.8}$$

This distribution is used for modelling the number of trials up to and including the first success.

2) The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set {0, 1, 2, 4, ... }.

This form of geometric distribution is used for modelling the number of failures until the first success:

$$P(X = k) = (1 - p)^k p, k = 0, 1, 2, 3, ... \tag{1.9}$$

and the Cumulative Distribution Function (CDF) is:

$$P(X \leq k) = 1 - (1 - p)^{k+1} \tag{1.10}$$

The geometric distribution is an appropriate model if the following assumption is true:

1) The phenomenon being modelled is a sequence of independent trials.

2) There are only two possible outcomes for each trial.

3) The probability of success p is the same for each trial.

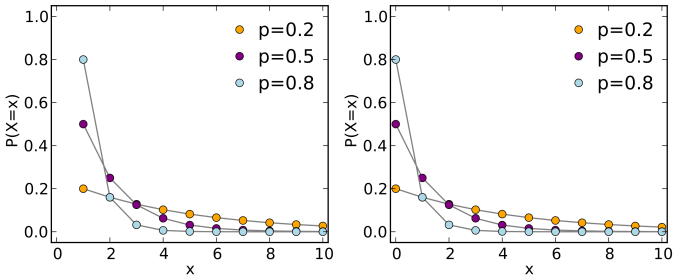

Figure 1.3: Probability Mass Function (PMF) of geometric distribution for case 1) (left) and case 2) (right).

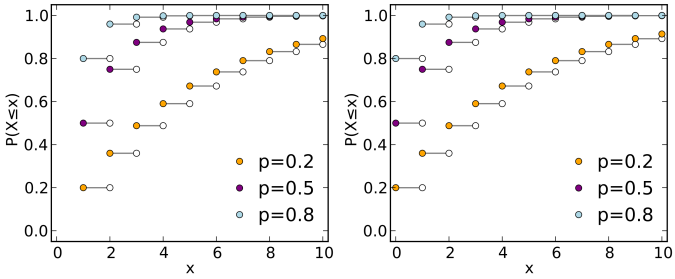

Figure 1.4: Cumulative Distribution Function (CDF) of geometric distribution for case 1) (left) and case 2) (right).

1.8 Discrete variable: Poisson Distribution, Pois(λ)

The Poisson Distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of times or space if these events occur with a known constant mean rate and independently of the time since the last event.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution.

The Poisson Distribution is popular for modelling the number of times an event occurs in an interval of time or space. A discrete random variable X is said to have a Poisson Distribution with parameter $\lambda > 0$, if, for $k = 0, 1, 2, ...$, the Probability Mass Function (PMF) of X is given by:

$$f(k; \lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} = P(\text{k events in interval}) \tag{1.11}$$

where λ is equal to the expected value of X and also its variance.

The Cumulative Distribution Functions (CDF) is:

$$F(k; \lambda) = e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!} \tag{1.12}$$

The Poisson Distribution is an appropriate model if the following assumptions are true:
1) k is the number of times an event occurs in an interval and k can take values 0, 1, 2, ...
2) The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
3) The average rate at which events occur is independent of any occurrences. For simplicity, this is usually assumed to be constant, but may in practice vary with time.
4) Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur.

The Poisson Distribution is also the limit of a binomial distribution, , for which the probability of success p for each trial equals to $\frac{\lambda}{\text{num. of trials}}$, as the number of trials goes to infinity.
If the number of Bernoulli trials goes to infinity (or every large), then Binomial distribution can be converted into Poisson distribution. (p will be small, it is a rare event.)

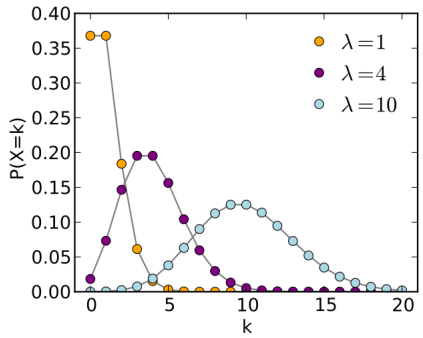

Figure 1.5: Probability Mass Function (PMF) of Poisson distribution.

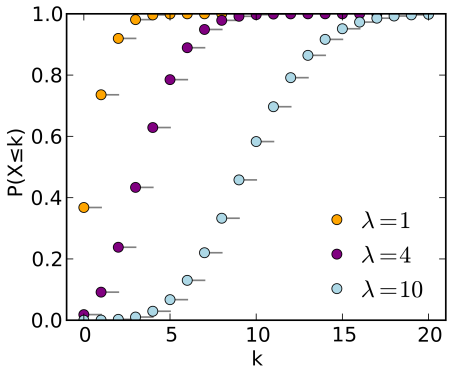

Figure 1.6: Cumulative Distribution Function (CDF) of Poisson distribution.

1.9 Continuous variable: Normal Distribution, $\mathcal{N}(\mu, \sigma^2)$

A Normal (Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its Probability Density Function (PDF) is: The Cumulative Distribution Functions (CDF) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{1.13}$$

where μ is the mean or expectation of the distribution (and also its median and mode), while σ is the standard deviation. The variance of the distribution is σ^2. Note that these μ and σ is associated with Normal Distribution and not the one computed directly from the data. Note also that Normal Distribution are affected by outliers because probability of any event happen more than 4 standard deviations from the mean is very low.

Normal distributions are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the Central Limit Theorem (CLT). CLT establishes that in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve), even if the original variables themselves are not normally distributed.

The Cumulative Distribution Function (CDF) for Normal Distribution is:

$$F(x) = \frac{1}{2}[1 + \text{erf}\,(\frac{x-\mu}{\sigma\sqrt{2}})] \tag{1.14}$$

where erf is the error function defined as:

$$\text{erf}\; z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{1.15}$$

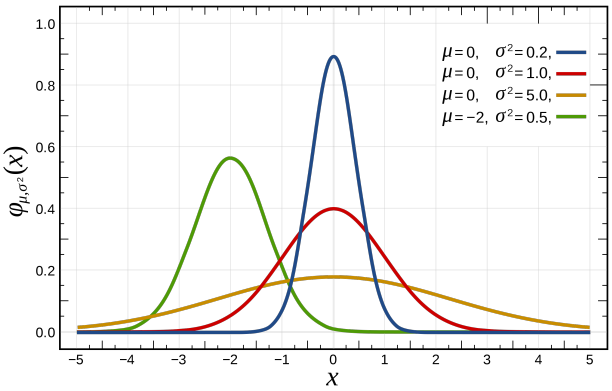

Figure 1.7: Probability Density Function (PDF) of Normal distribution.

About 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean μ. About 95% of the values lie within two standard deviations and about 99.7% are within three standard deviations. This is known as the 68-95-99.7 rule.

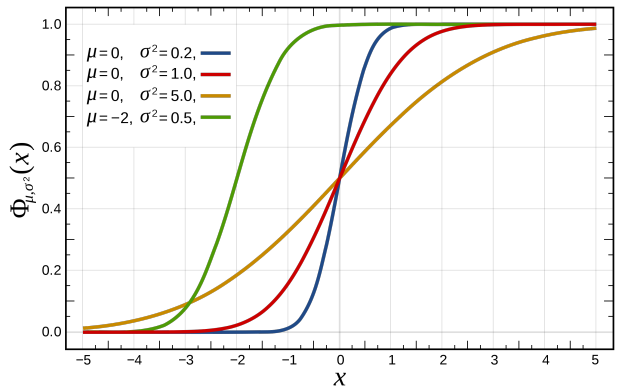

Figure 1.8: Cumulative Distribution Function (CDF) of Normal distribution.

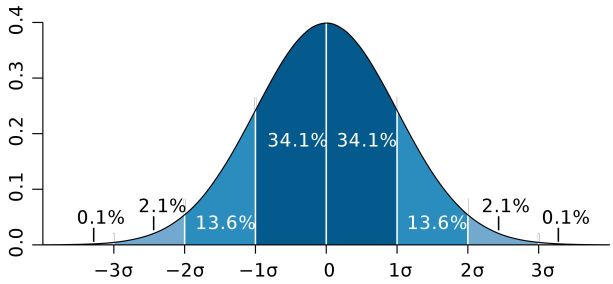

Figure 1.9: Normal distribution with standard deviation

1.10 Continuous variable: Exponential Distribution, Exp(λ)

Poisson process:

the timing of the next event is completely independent of when the previous event happened (memoryless).

The number of arrivals of a Poisson process in a given amount of time is Poisson Distributed.

The waiting time between arrivals of a Poisson process is Exponentially Distributed.

The Probability Density Function (PDF) of an exponential distribution is:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{1.16}$$

where λ is the rate parameter used in the Poisson Distribution.

The Cumulative Distribution Function (CDF) is given by:

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{1.17}$$

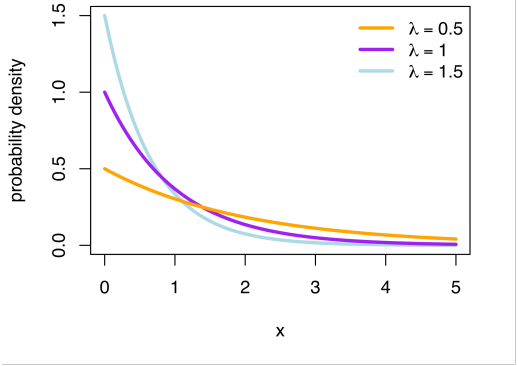

Figure 1.10: Probability Density Function (PDF) of Exponential Distribution.

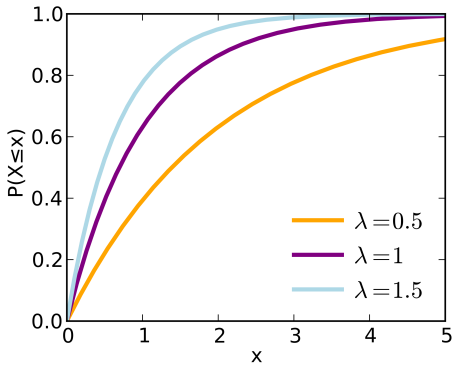

Figure 1.11: Cumulative Distribution Function (CDF)of Exponential Distribution

1.11 Continuous variable: Chi-Square Distribution, χ^2

The chi-square (χ^2) distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-square distribution is used in the common chi-square chi-square tests for goodness of fit of an observed distribution

to a theoretical one.

Let's say we have some random variables, each of them are independent standard, normally distributed random variables. Say we have random variable X_1:

$$X_1 \sim N(0,1) \tag{1.18}$$

This variable has mean = 0, variance/standard deviation = 1 and follows Normal Distribution.

Say now we have a new random variable $Q = X_1^2$, i.e. we sample from the distribution of X_1 and square whatever number we have. Q will have a chi-square distribution with k = 1 degree of freedom, i.e. $Q \sim \chi_1^2$. If $Q = X_1^2 + X_2^2$ then Q will have a degree of freedom = 2, $Q \sim \chi_2^2$.

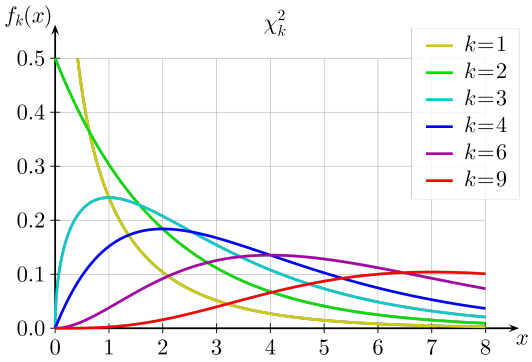

Figure 1.12: Probability Density Function (PDF) of Chi-Square distribution.

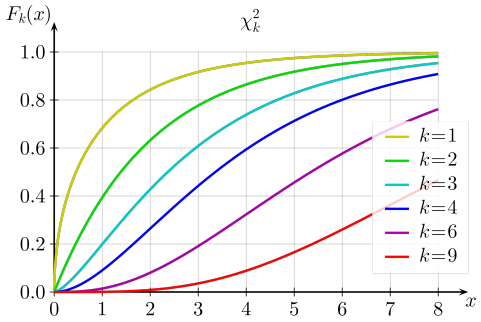

Figure 1.13: Cumulative Distribution Function (CDF) of Chi-Square distribution.

The Probability Density Function is shown in Fig.(1.12). For k = 1 the probability is high when x = 0. If we just sample once from X_1 we are very likely to get a value that close to 0.

Chapter 2

Metrics

2.1 Variance, σ^2 or Var(X)

Variance is the expectation of the squared deviation of a random variable from its mean. Note that the unit of variance is the square of the variable's unit.

1) discrete random variable

This is for discrete random variable (applicable to most dataset), if the generator of random variable X is discrete with Probability Mass Function (PMF) that maps value x_i to probability p_i, (certain x_i and p_i pairs can be same value due to same x value in the dataset) then the variance will be:

$$Var(X) = \sum_{i=1}^{n} p_i(x_i - \mu)^2 \tag{2.1}$$

where μ is the expected value:

$$\mu = \sum_{i=1}^{n} p_i x_i \tag{2.2}$$

If the each value in the n data points are equally likely, then the variance will be:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2.3}$$

2) continuous random variable

If the random variable X has a Probability Density Function (PDF) $f(x)$, then the variance will be:

$$Var(X) = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx \tag{2.4}$$

where μ is the expected value:

$$\mu = \int_{\mathbb{R}} x f(x) dx \tag{2.5}$$

2.2 Standard Deviation, σ

The standard deviation is a measure of the amount of variation or dispersion of a set of values. For both discrete and continuous random variable, the standard deviations are $\sqrt{\text{variance}}$ and note that this quantity has the same physical units as the random variable.

2.3 Standard Error

The standard error is a type of standard deviation for the distribution of the means.

There will be, of course, different means for different samples (from the same population), this is called "sampling distribution of the mean". This variance between the means of different samples can be estimated by the standard deviation of this sampling distribution and it is the standard error of the estimate of the mean. Standard error measures the precision of the estimate of the sample mean. The standard error is strictly dependent on the sample size and thus the standard error falls as the sample size increases. It makes total sense if you think about it, the bigger the sample, the closer the sample mean is to the population mean and thus the estimate of it is closer to the actual value.

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} \tag{2.6}$$

where σ is the standard deviation of the population (although sometimes population standard deviation is unknown, we can replace it with sample standard deviation as an estimate) and n is the size (number of observations) of the sample.

2.4 Confusion Matrix

The confusion matrix is a table specifically for the problem of statistical classification. A typical confusion matrix table is shown below:

		Actual class	
		P	**N**
Predicted class	**P**	**TP**	FP
	N	FN	**TN**

Figure 2.1: Confusion Matrix

where:

TP : True Positive or Hit

TN : True Negative or Correct Rejection

FP : False Positive or Type I error

FN : False Negative or Type II error

A more graphical way of seeing this is shown in Fig. (2.6).

2.4.1 Recall

Recall is also called Sensitivity or True Positive Rate (TPR), it measures how many relevant items are selected (among the total numbers of relevant elements):

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2.7}$$

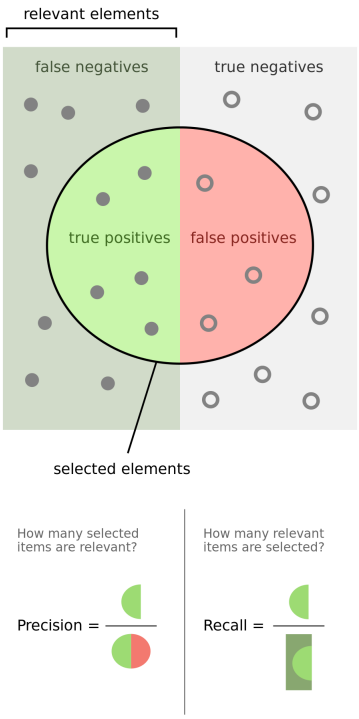

Figure 2.2: Graphical representation of confusion matrix

2.4.2 Precision

Precision is also called Positive Predictive Value (PPV), it measures how many selected items are relevant:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2.8}$$

2.4.3 F_1

The F_1 score is a measure of a test's accuracy. It is basically the harmonic mean of the precision and recall. The highest possible F_1 score is 1, indicating perfect precision and recall, and the lowest possible score is 0, if either precision or recall is 0.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.9}$$

2.5 ROC

The Receiver Operating Characteristic Curve, or ROC curve, is a graphic plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

To obtain the ROC curve, we need to plot the True Positive Rate (TPR) against the False Positive Rate (FPR):
True Positive Rate (TPR): also known as Sensitivity, Recall.
False Positive Rate (FPR): $\frac{\text{FP}}{\text{FP+TN(i.e. total actual negative)}}$, and is basically $1 - \text{Specificity}$.

In an ROC curve, the best possible prediction model would yield a point (for a certain discrimination threshold) in the upper left corner. This represent 100% Recall (No False Negative) and 100% Specificity (No False Positive). A random guess would give a point along a diagonal line. An intuitive example of random guessing is a decision by flipping coins. As the size of the sample increases, a random classifier's ROC point tends towards the diagonal line, specifically for the case of a fair coin, it will tend to the point (0.5, 0.5)

The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random); points below the line represent bad results (worse than random). The table below shows 4 prediction model for 100 positive and 100 negative instances:

A			B			C			C′		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112	TP=76	FP=12	88
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88	FN=24	TN=88	112
100	100	200	100	100	200	100	100	200	100	100	200
TPR = 0.63			TPR = 0.77			TPR = 0.24			TPR = 0.76		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.12		
PPV = 0.69			PPV = 0.50			PPV = 0.21			PPV = 0.86		
F1 = 0.66			F1 = 0.61			F1 = 0.23			F1 = 0.81		
ACC = 0.68			ACC = 0.50			ACC = 0.18			ACC = 0.82		

Figure 2.3: The prediction from 4 models, each at a certain discrimination threshold.

Plots of the four models in the ROC space is given below. The result of A clearly shows

the best predictive power among A, B, C. The result from B lies on the random guess line, and it can be seen in the table that the accuracy of B is 50%. However, when C is mirrored across the center point (0.5, 0.5), the resulting model C' is even better than A. This mirrored method simply reverse the predictions from C.

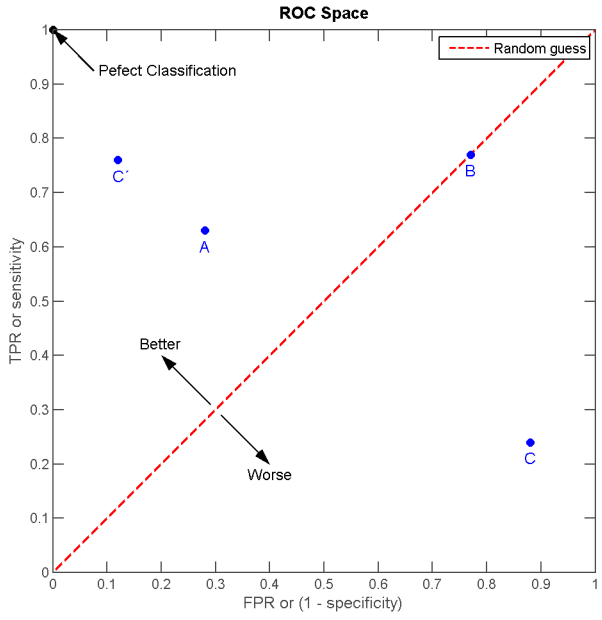

Figure 2.4: The prediction from 4 models, each at a certain discrimination threshold.

The closer the model can be at towards the upper left corner, the better. But the distance from the random guess line in either direction is the best indicator of how much predictive power a method has. If the result is below the line, all of the model's predictions must be reversed in order to utilize its power. Note that the output of a consistently bad predictor could simply be inverted to obtain a good predictor.

2.5.1 In the context of logistic regression

For binary classification, given certain features, logistic regression will output a probability p, with respect to the target variable.

if $p > 0.5$, we label the data as 1;

if $p < 0.5$, we label the data as 0;

By default the threshold is 0.5.

When the threshold = 0 , the model will predict 1 for all, which means that both TPR and FPR will be 1.
When the threshold = 1, the model will predict 0 for all, which means that both TPR and FPR will be 0.

If we vary the threshold between these 2 extremes, we will have a series of TPR and FPR pairs. If we try all possible threshold then we can map out the ROC curve for this particular model.

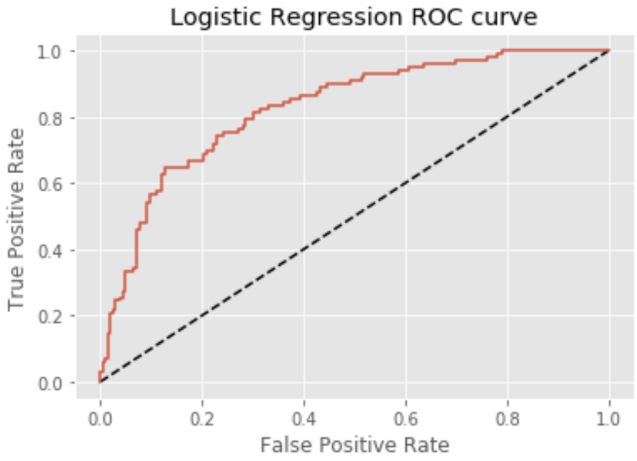

Figure 2.5: ROC curve for a logistic regression model.

When using normalised unit, the area under the curve (AUC) is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative ones. (assuming that positive rank higher than negative in the dataset).

2.6 Covariance

The covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, (i.e. the variables tend to show similar behaviour), then the covariance is positive. In the opposite case, when the greater values of one variable

mainly correspond to the lesser values of the other (i.e. the variables tend to show opposite behaviour), then the covariance is negative.

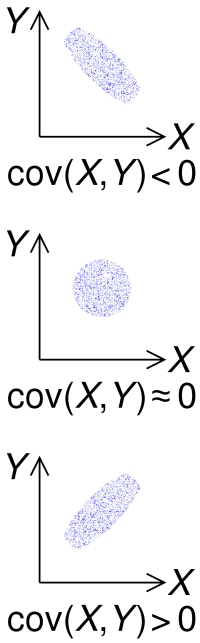

Figure 2.6: The sign of the covariance of two random variables X and Y.

The description above is for covariance of two random variables. We also have sample covariance. The formula to calculate covariance is:

$$\text{covariance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \tag{2.10}$$

where μ_x and μ_y are the mean of x and y in the dataset. When both x and y increase or decrease together, then they are positively correlated.

2.7 Pearson's Correlation Coefficient

The correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. The most familiar measure of dependence between two quantity is the Pearson product-moment correlation coefficient, or Pearson's Correlation Coefficient. This coefficient is dimensionless and takes value between (-1, 1). The correlation coefficient is +1 in the case of a perfect directly (increasing) linear relationship, -1 in the

case of a perfect inverse linear relationship.

If the variables are independent, Pearson's Correlation Coefficient is 0. But the converse is not true, because the correlation coefficient detects only linear dependencies between two variables.

The Pearson Correlation Coefficient $\rho_{X,Y}$ between two random variables X and Y with standard deviation σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{\text{covariance}(X,\, Y)}{\sigma_X \sigma_Y} \tag{2.11}$$

The Pearson Correlation is only defined only if both standard deviation are finite and positive.

2.8 Confidence Interval

2.9 Confidence Level

2.10 P value

See Chapter (3.4)

2.11 T value/t-statistic

The t-statistics See Chapter (3.2)

2.12 Z value

Details please see Chapter (3.1).

Z value, also known as z-scores and more commonly the Standard Score, is the number of standard deviations by which the value of a raw score (i.e. an observed value or data point)

is above or below the mean value of what is being observed or measured. Raw scores above the mean have positive z value, while those below the mean have negative z value.

If the population mean and population standard deviation are known, a raw score x is converted to z score by:

$$z = \frac{x - \mu}{\sigma} \tag{2.12}$$

where:
μ is the mean of the population and σ is the standard deviation of the population. Calculating z score using this formula requires the population mean and population standard deviation, NOT the sample mean or sample deviation. But knowing the true mean and standard deviation of a population is often unrealistic except in cases such as standardised testing, where the entire population is measured.

When the the population standard deviation are unknown, the z score may be calculated using sample standard deviations as estimates of the population. In these case, sometimes the z score is normalised by the sample size:

$$z = \frac{x - \bar{x}}{\bar{\sigma}} \tag{2.13}$$

where $\bar{x}$ and $\bar{\sigma}$ are the mean and standard deviation of the sample. Sometimes for sample-based z value, the z value is normalised by the $\sqrt{n}$ where n is the sample size used to get the z value:

$$z = \frac{x - \bar{x}}{\bar{\sigma}/\sqrt{n}} \tag{2.14}$$

The use of sample deviation instead of population standard deviation is alright if the sample size is large enough (Central Limit Theorem). If population mean is available use population mean instead.

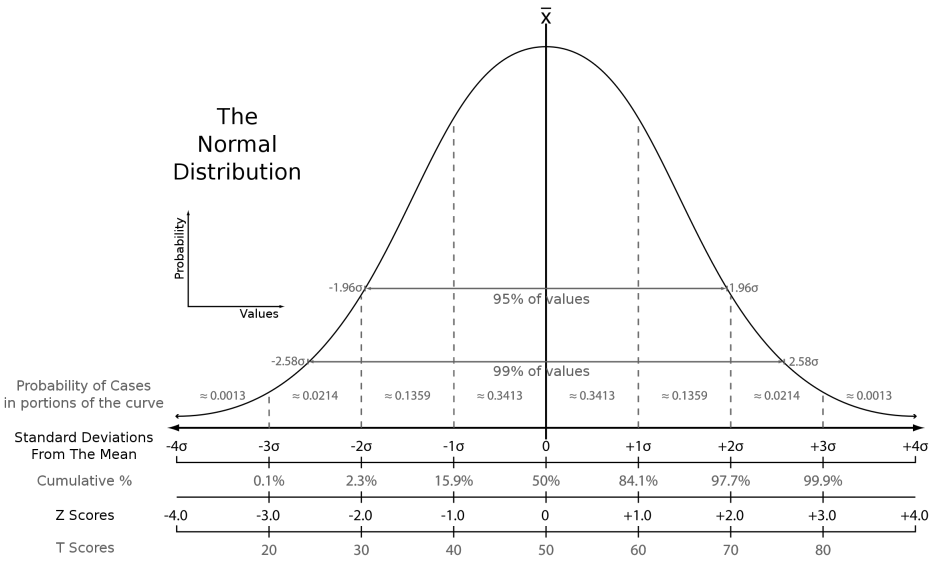

Figure 2.7: Compares the various grading methods in a normal distribution. Includes: Standard deviations, cumulative percentages, percentile equivalents, Z-scores, T-scores.

In educational assessment, T-score is a standard score Z shifted and scaled to have a mean of 50 and a standard deviation of 10.

Chapter 3

Testing

3.1 Z test

A Z-test is any statistical test for which the distribution of the test statistics (i.e. a quantity derived from the sample) under the Null Hypothesis can be approximated by a normal distribution. Z-test tests the mean of a distribution.

How to perform a Z test when T is a statistics that is approximately normally distributed under the Null Hypothesis:
1) estimate the expected value (mean) μ of T under the Null Hypothesis, and obtain the standard deviation σ of T.
2) determine the properties of T: one tailed or two tailed:
For Null Hypothesis $H_0 : \mu \geq \mu_0$ vs Alternative Hypothesis $H_1 : \mu < \mu_0$, it is right one-tailed test.
For Null Hypothesis $H_0 : \mu \leq \mu_0$ vs Alternative Hypothesis $H_1 : \mu > \mu_0$, it is left one-tailed test.
For Null Hypothesis $H_0 : \mu = \mu_0$ vs Alternative Hypothesis $H_1 : \mu \neq \mu_0$, it is two-tailed test. where μ is the true mean of the population under analysis, μ_0 is the hypothesised mean of the population under analysis.

3) calculate the Z score according to Eqn.(2.14), then the one-tailed or two-tailed p-value can be obtained. This p-value is then compared with the desired confidence level to determine if H_0 is rejected or not.

In summary, a simple case of Z test is that we have a population distribution (with a certain population mean and standard deviation). Now given a new sample, we would like to know if the mean of the new sample is significantly differ from the population mean, and we can use z test to find out assuming the distribution follow normal distribution.

Example 1

In a region (population) the mean and standard deviation of scores on a reading test is 100 and 12. Our interest (sample) is in the score of 55 students in a particular class who received a mean score of 96. We can ask whether this mean score is significantly lower than the region mean, that is, are the students in this school comparable to a simple random sample of 55 students from the region as a whole, or are their scores surprisingly low ?

Using Eqn.(2.14), we can calculate the Z score as:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{96 - 100}{12/\sqrt{55}} = -2.47 \tag{3.1}$$

In this example we treat the population mean and variance as known, which would be appropriate if all students in the region were tested. When population parameters are unknown, a t test should be conducted instead.

The classroom mean score is 96, which is -2.47 standard error (since we divided by $\sqrt{n}$, we have converted it into standard error) from the population mean of 100. Looking up the z-score in the table of standard normal distribution cumulative probability, we find that the probability of observing a standard normal value below -2.47 is around $0.5 - 0.4932 = 0.0068$. This is the one-tailed p-value for the Null Hypothesis that the 55 students are comparable to a simple random sample from the population of all test-takers. The Z test tells us that the 55 students of interest have an unusually low mean test score compared to the population mean.

Example 2

A complaint was registered stating that the boys in the school were underfed. Average weight of boys of age 10 is 32kg with 9kg standard deviation (population). a sample of 25 boys was selected from the school and the average is found to be 29.5kg. At 0.05 significance level ($\alpha = 0.05$), we need to check whether the complaint is true or not.

The Null Hypothesis says that there is no significant difference between the boys and the whole population. i.e.
$H_0 : \mu = 32$, no significant difference between the students.
The Alternate Hypothesis says that there is a significant difference between the boys and the whole population and the complaint is true, i.e.:
$H_1 : \mu < 32$, there is a significant difference between the students.

Using the Eqn. (2.14), we have Z value of $\frac{29.5-32}{9/\sqrt{25}} = -1.39$. This Z value can be used to find the p-value of the same selected samples. The corresponding p-value is 0.0823. This p-value is greater than $\alpha = 0.05$, this means that in this left-tailed test, we can accept (cannot reject) the Null Hypothesis. This means that there is a probability of 0.0823 that a sample from the population can have more than 1.39 standard error from the population mean, as demonstrated by our sample (because we want to know if a sample drawn from this population is the same as the sample we are studying.), and it is larger than our α. In this case, there is no difference between the students (at this confidence level) and the complaint is not true.

3.2 T test

T test are very similar to Z test described above. Both test usually requires the sample means to exhibit normality for exactness. In addition, whether using T test or Z test depends on:

1) The size of our sample. The magic number is usually around 30-50. Below that is considered a small sample. When the sample size is large enough, the central limit theorem kicks in and we do not need to worry too much about the population is normally distributed.

2) If we know the population standard deviation (σ). In real life we usually do not know this value. But in certain cases we can have this information.

If sample size is large and σ is known, we can use Z test. If sample size is large and σ is unknown, use T test. If the population is small, usually use T test instead unless population σ is known. The formula is similar to Z-statistics:

$$t - statistics = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{3.2}$$

where μ is the population mean and S is the sample standard deviations.

3.3 Chi-squared test

A chi-squared test (χ^2 test) is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the Null Hypothesis, specifically Pearson's chi-squared test. The Pearson's chi-square test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories.

Suppose that n observations in a random sample from a population are classified into k mutually exclusive classes with respective observed numbers x_i (for i = 1, 2, ... , k) and a Null Hypothesis gives the probability p_i that an observations falls into the ith class. So we have expected numbers $m_i = np_i$.

The chi-square statistics:

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} \tag{3.3}$$

The expected number m_i is large enough known numbers in all cells assuming every x_i maybe taken as normally distributed. And then in the limit that n becomes large, X^2 follows the χ^2 distribution with $k-1$ degrees of freedom. We can then look up the χ^2 table to obtain (one- or two-tail) p-value to determine if Null hypothesis should be rejected.

Example

For example, we want to investigate what is the distribution of the number of customers we get each day for a particular restaurant ? The restaurant owner claimed that the distribution is the expected number in the table below. We would like to see how good this distribution he claimed actually fits the observed data.

Day	M	T	W	T	F	S
Observed	30	14	34	45	57	20
Expected	20	20	30	40	60	30
Expected %	10	10	15	20	30	15

We can make the Null Hypothesis that the owner's distribution is correct, i.e. H_0: The expected number of customer is right. Then H_1: The expected number is wrong. And I want to do this with a significance level of 5% or $\alpha = 0.05$. We are going to calculate a test statistics here based on the table, this statistics is the chi-square statistics. Another way to understand it is that the statistics we are calculating has approximately a chi-square distribution. And given that it does have a chi-square distribution with a certain number of degrees of freedom, we would like to see the probability of getting the observed results (or more extreme) is less than 5%, in that case we can reject the Null Hypothesis.

From the observed data, we obtained that the total number of customers is 200, which can be used to calculate the Expected number of customers in the table above. Then using the chi-square formula, we can obtain the chi-square statistics (approximate to χ^2):

$$\begin{aligned} X^2 &= \frac{(30-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(34-30)^2}{30} \\ &+ \frac{(45-40)^2}{40} + \frac{(57-60)^2}{60} + \frac{(20-30)^2}{30} \\ &= 11.44 \end{aligned} \tag{3.4}$$

If we assume this test statistics follows a chi-square distribution, what is the probability of getting a result this extreme ? For our particular case, the degree of freedom is $(6-1) = 5$. If we check out a chi-square table of the graph of chi-square PDF (Fig.(1.12)), the value for 11.44 is a bit less than 5%. Hence we can reject H_0.

3.4 A/B testing

A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomised experiment with with two variants, A and B. It includes application of statistical hypothesis testing (Null vs Alternative hypothesis) or two-sample hypothesis testing.

3.4.1 Null hypothesis H_0

The Null hypothesis usually states that there is no difference between groups in study. For example: there is no relationship between the risk factor (or treatment) being studied AND the occurrence of the health outcome in a medical study. (In this case, Group A is a placebo group and Group B receiving a new drugs).

3.4.2 Alternative hypothesis H_1

The Alternative hypothesis states that there is a difference between groups in study. There IS a relationship between the treatment or risk factor AND the outcomes of the experiment.

By default, we assume that the Null Hypothesis is true, until we have enough evidence to support rejecting this hypothesis. A bummer if Null Hypothesis is True.

But we can NEVER prove the Alternative Hypothesis is true, the best we can do is to REJECT a hypothesis (saying it is false) or FAIL to reject a hypothesis (could be true, but never sure). So usually we want to reject the Null Hypothesis, because that is the as close as we can get to prove the Alternative Hypothesis.

3.4.3 Type I/II error

Rejecting the Null Hypothesis when H_0 is True is called Type I error (also known as False Positive), i.e. the researcher say there is a difference between the groups when there really isn't. This error is usually the focus because researcher wants to show H_0 is False. The probability of making a Type I error is called α,

Type II error (also known as False Negative) is when we failed to reject H_0 when we should reject it. The probability of making a Type II error is called β.

3.4.4 Statistical Power

Power = probability of finding a difference between groups if one truly exist. (i.e. H_0 is False)
= probability of not making a Type II Error
$= 1 - \beta$

A good power is around 0.8. Power matters during experiment design. We should do power calculations based on projections.

Power increases when:
1) increase in sample size. i.e. you have more data to make a conclusion.
2) (big) actual difference between groups, i.e. effect size

3) (good) precision of results, i.e. multiple samples show consistent results instead of all over the place.

3.4.5 Test of Significance

A common indicator when testing significance is the p-value.

p-value is the probability of obtaining a sample more extreme than the ones observed in our data, assuming H_0 is True. If this value is low, then it means either our power is low or there is a low probability of observing this value if the Null Hypothesis is True. This represents a measure of evidence against retaining H_0. We don't have to prove H_0 to use this, we just assume it is true before using the p-value.

For example, we can calculate z-test value for a right tailed test assuming normal distribution. Then the p-value is just the area under the curve to the right of the z-test score.

What determines if p-value is low or high ?
We use α, which is also called level of significance. It is a selected cut-off point to determine if the p-value is acceptably high or low.

We define our probability of not making type I error as the confidence level, a common value for this is 0.95.

Standard p-value:
<0.01: very strong evidence against the Null hypothesis.
0.01 - 0.05: strong evidence against the Null hypothesis.
0.05 - 0.10: very weak evidence against the Null hypothesis.
more than 0.1: small to no evidence against the Null hypothesis.

3.5 Hypothesis testing (one-way and two-way)

3.6 ANOVA

3.7 ANCOVA

3.8 One-sample/Two-sample bootstrap hypothesis test

3.9 Time series: p, d, q parameters, unit root and box test

Chapter 4

Thoerem

4.1 Central Limit Theorem

4.2 Law of the large number

4.3 Naive Bayes Algorithm

4.4 Bayesian Statistics/Bayes Theorem

4.5 Sampling Theory

Chapter 5

General

5.1 Confidence Interval

5.2 Conditional Probability

Bayes's theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5.1}$$

where A and B are events and $P(B) \neq 0$.

P(A|B) is a conditional probability: the likelihood of event A occurring given that B is true. $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively and they must be different event.

For conditional probability, $P(A|B)$ may or may not be equal to $P(A)$ (the unconditional probability of A). If $P(A|B) = P(A)$, then the events A and B are said to be independent. In this case, knowledge about either event does not alter the likelihood of each other. The definition (defined this way, not theoretical results) of the conditional probability is often quoted as :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5.2}$$

This maybe visualised as restricting the sample space to situations in which B occurs. The logic for this equation is that if the possible outcomes for A and B are restricted to those in which B occurs, this set serves as the new sample space.

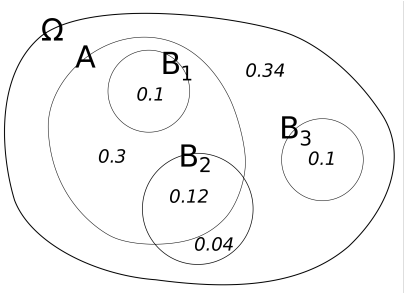

Figure 5.1: Conditional probability with an Euler diagram

The unconditional probability $P(A) = 0.3 + 0.1 + 0.12 = 0.52$. However, the conditional probability $P(A|B_1) = 1$, $P(A|B_2) = 0.12/(0.12 + 0.04) = 0.75$, and $P(A|B_3) = 0$

5.3 Normalisation

5.4 Standardisation

5.5 Least-squared error

5.6 R-squared error

5.7 Mean-squared error

5.8 Inferential Statistics

5.9 Bias-variance trade off

The bias-variance tradeoff is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters. The bias-variance dilemma is the conflict in trying to simultaneously minimise these two sources of error that prevent supervised learning algorithm from generalising beyond their training set.

The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs, aka, underfitting.

The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs, aka, overfitting.

This trade-off is universal: it has been shown that a model that is asymptotically unbiased must have unbounded variance.

Dimensionality reduction and feature selection can decrease variance by simplifying mod-

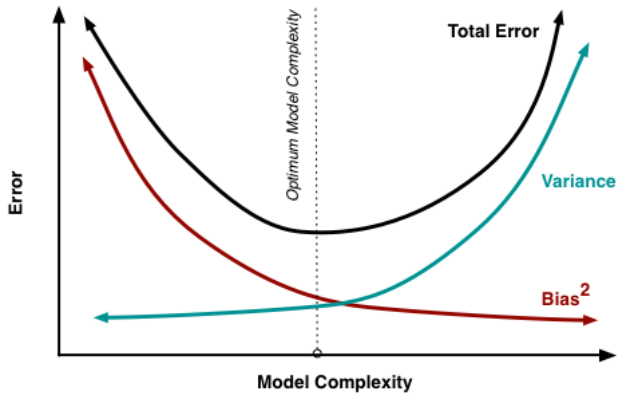

Figure 5.2: Bias-variance trade-off.

els. Similarly, a larger training set tends to decrease variance. Adding features tends to decrease bias, at the expense of introducing additional variance. Learning algorithm typically have some tunable parameters that control bias and variations, some of the examples are:

1) Linear models can be regularised to decrease their variance at the cost of increasing their bias.

2) In neural network, variance increases and the bias decreases as number of hidden units hidden units increase.

3) In KNN, a high value of k leads to high bias and low variance.

4) In decision trees, the depth of the tree determines the variance. Decision trees are commonly pruned to control variance. (e.g. Reduced Error Pruning: starting at the leaves, each nodes is replaced with its most popular class, if the accuracy is not affected then the change is kept.) Pruning a node consists of removing all subtrees, making it a leaf, and assigning it the most common classification of the associated training examples.