# Clustering

## What is meant by a hierarchical clustering algorithm?

These methods construct a tree-like structure (dendrogram) to show groups of observations. It can be agglomerative (each point starts as its own cluster i.e. bottom up) or divisive (all points are part of one cluster and we split it i.e. top down).

You have n objects and wish to form k clusters:

1. Construct a distance matrix where the distance is Euclidean, Manhattan, etc.
2. Start with m=n clusters
3. Find the closest pair of clusters and merge them. Now there are m=n-1 clusters.
4. Repeat until m=1 i.e. there is a single cluster
5. Examine the sequence of partitions on the dendrogram and choose the best k.

## Within hierarchical clustering, choices are available concerning the dissimilarity measure and the linkage function. Explain what effect both of these have.

Hierarchical clustering aims to find groups of observations such that observations within a group are very similar and different groups are very dissimilar.

The following have been proposed as dissimilarity measures for measurement data, :

Euclidean

Manhattan

Maximum

Minkowski

There are various methods for measuring dissimilarity between two groups A and B where A = {xa1, xa2,…, xak} and B = {xb1, xb2,…, xbk} :

Single linkage

Complete linkage

Average linkage

## What is meant by an iterative clustering algorithm?

Iterative clustering has a predetermined number of clusters. Assign data points to a specific cluster and then iterate points between clusters until convergence is found i.e. we repeat previous clustering of the data and choose the cluster with the minimum internal dissimilarities e.g. k-means clustering.

## Describe the k-means clustering algorithm.

The aim of k means clustering is to divide the data into k distinct groups so that observations within a group are similar whilst observations between groups are different. K-means clustering is an iterative clustering algorithm (rather than hierarchical).

Step 1 – Choose the number of clusters k

Step 2 – Assign each data point to the cluster who’s centre is closest

Step 3 – Calculate each cluster’s centroid using:

Step 4 – Calculate the sum of squared distances of each object to its cluster centroid:

Where

* n = number of observations
* m = number of variables in an observation

Step 5 – Reassign each observation to the cluster whose centroid is closest

Step 6 – Repeat 3-5 until convergence

## What type of cluster structure is k-means clustering good at finding?

It is good at finding circular data. If a data point goes from cluster 1 to cluster 2 to cluster 1, convergence will never be satisfied and we have found circular data.

k-means is sensitive to the selection of starting points so running the algorithm several times for different starting values can help check whether the results are robust.

## Give a brief account on the use of standardization and data reduction in conjunction with cluster analysis methods.

Standardization standardizes the variance and so it changes the inter-point distances. Data reduction provides a lower dimension analysis of the data.

When constructing a dissimilarity matrix for use in clustering, we need to be aware of how our data is scaled. It is important that different variables are comparably scaled, otherwise the variable with the greatest variance will be most prominent in the clustering solution. Variables are standardized by dividing through by their standard deviation to give them a variance of 1.

Dimension reduction helps to reduce random noise and nuisance variables. It makes it easier to visualize the data. PCA can be used to reduce the dimensionality of a data set. We use the amount of PCs that account for the majority of the variance. Other methods of data reduction that may be used are factor analysis and multi-dimensional scaling.

## Explain whether scaling the data so that each variable has an equal standard deviation will have an effect on the results of a hierarchical clustering algorithm.

Yes – Scaling can give different variables different amounts of emphasis/contribution to the dissimilarities. It could potentially change the order of the closest data points, hence, the order in which the data points are merged can change, changing the composition of the clusters.

## Explain whether subtracting the mean vector of the data from each observation will have an effect on the results of a hierarchical clustering algorithm.

No – Here we are only shifting the data by a constant so we’re not affecting dissimilarities. Therefore, the clustering is not affected.

## Describe a real problem in which a clustering algorithm would offer an appropriate solution.

Clustering algorithms could be used in fraud detection. If there are groups of individuals who have similar bank transactions and one person is found to be fraudulent, then possibly the other people in that cluster have been fraudulent too.

## Using maximum dissimilarity, generate a dissimilarity matrix for the following:



Repeat for all 6 options and you end up with...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | UK | USA | Ire | China |
| UK | 0 | 3.4 | 2.4 | 23.6 |
| USA | 3.4 | 0 | 1.6 | 22.9 |
| Ire | 2.4 | 1.6 | 0 | 24.5 |
| China | 23.6 | 22.9 | 24.5 | 0 |

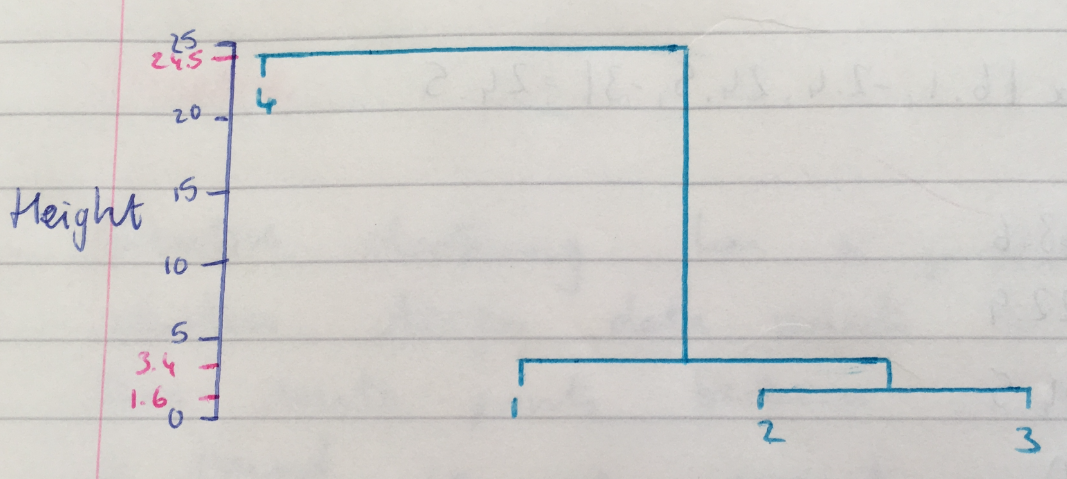
## Using complete linkage, produce a sketch of the resulting dendrogram.

Find the closest pair of objects (2,3) > 1.6. Place them in the same cluster. To get the distance between (2,3) and 1, look at the maximum distance.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | (2,3) | 4 |
| 1 | 0 | 3.4 | 23.6 |
| (2,3) | 3.4 | 0 | 24.5 |
| 4 | 23.6 | 24.5 | 0 |

Find the closest pair of objects [(2,3),1] > 3.4. Place them in the same cluster.

|  |  |  |
| --- | --- | --- |
|  | (1,2,3) | 4 |
| (1,2,3) | 0 | 24.5 |
| 4 | 24.5 | 0 |

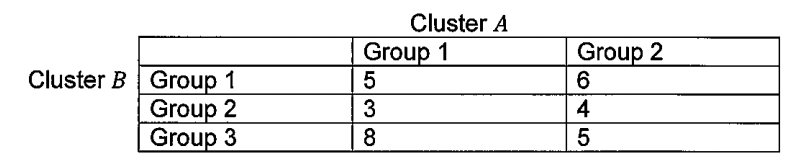


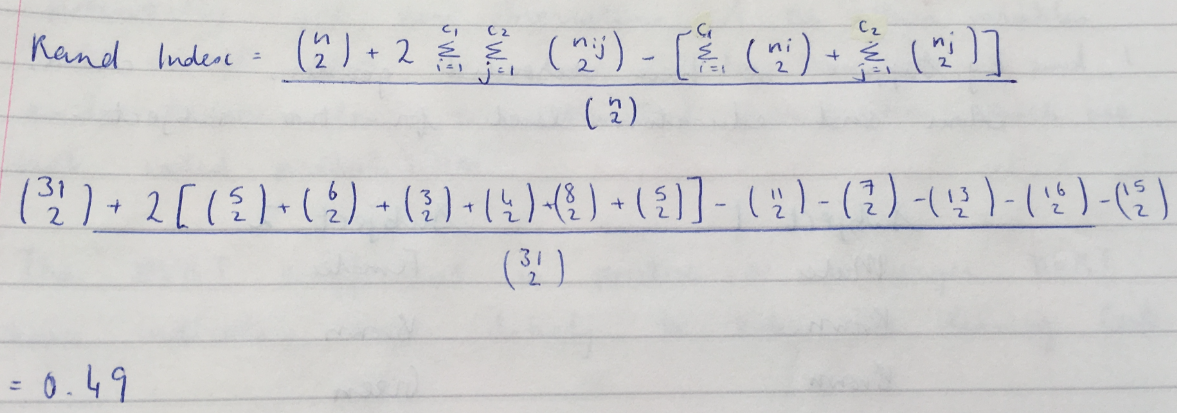
Two clusters are clearly present. Anything between 3.4 and 24.5 would be an appropriate cut value. An often suggested rule is the cut the tree at .

## Explain the role of Rand Index in cluster analysis.

Rand Index is used for measuring the agreement between two clusters. It is a number between 0 and 1 with 0 representing little agreement and 1 representing strong agreement.

## Calculate the Rand Index for the following table.





## Explain a method for calculating the dissimilarity from data points that contain binary, numeric and categorical data.

Take each collection of variables of the same type and calculate a dissimilarity assuming they were the only recordings (e.g. use Euclidean for numeric data, Simple matching for binary and proportion of times in agreement for categorical). Combine these in a weighted average with the weightings being given by the proportion of all variables that are of the type specified.

**Binary**:

Use simple matching (as described above). This measures the proportion of variables in agreement.

**Numeric**:

Euclidean

Manhattan

Maximum

Minkowski

**Categorical**:

Count the number of terms that differ.

## Contrast clustering and classification algorithms.

Classification algorithms are irrelevant when there is labelling already present. However, when a new data point is unlabelled, you use these algorithms to find out where it belongs e.g. k-nearest neighbours.

Clustering on the other hand occurs when there is no labelling present and you want to group similar points e.g. k-means clustering.

# Dimension Reduction

## Provide a brief description on the benefits of dimension reduction as a statistical technique.

Dimension reduction is useful for visualisation purposes. It also reduces computation intensity. It can show which variables are important and get rid of “nuisance variables” (noise). You might also do a dimension reduction before doing a cluster analysis to get more accurate results.

## Outline the problems associated with the lower dimensional representation of data.

Risk of losing important information and of distorting the relationship between variables.

## Explain the motivation behind, and usefulness of, PCA.

For data with many variables/dimensions, it is often difficult to comprehend or visualize inherent associations. PCA is a method for re-expressing the data so as to reveal its internal structure and explain its variation through the use of a few linear combinations of the original variables.

We are motivated to use PCA because:

1. It is a dimension reduction technique which is useful as when the data has many variables as it is often difficult to comprehend or visualize inherent associations.
2. It aids visualisation which helps us interpret the data.
3. It can assist other clustering or classification algorithms by removing noise.

The intention is that the first few PCs will account for a substantial amount of the variation in the original data, and as such, can be used as a convenient lower dimensional summary of it.

## Describe the pros and cons of doing PCA on standardized data.

Standardizing the data ensures that every variable is dimensionless and has a variance of 1 (by dividing through by variance). This allows analysis of the relationships between the data points.

If all the variables are measured in the same unit (e.g. mm) and you apply PCA on the standardized data, you’re not applying it on the data itself and so you may be missing out on useful information.

Note: If it says correlation matrix > standardized data and if it says covariance matrix > not standardized.

## How are the standard deviation values in PCA generated?

The standard deviation is the square root of the variance. The variance is the eigenvalue of the corresponding eigenvector. The eigenvector is given by principal components. The first principal component (eigenvector) is a particular line out of all possible lines where the data is most spread upon. Subsequent standard deviations represent lines which are orthogonal to the previous line.

## Describe Factor Analysis and compare it to PCA.

PCA is data driven, we take the covariance matrix and perform an eigen decomposition. We find linear combinations of the original variables which have maximum variance.

Factor analysis is model driven. It seeks unobserved linear combinations of the variables representing underlying fundamental qualities. We assume, on average, that the values of the underlying factors are 0, that they are independent and have a covariance of 1. We find interpretation for them by looking at factor loadings which are the correlations between the unobservable latent factors and the observable recordings.

We have additional noise which is explained for by specific factors which are assumed to be 0 on average, be independent but have different values for their variance depending on how well they explain particular recorded measurements.

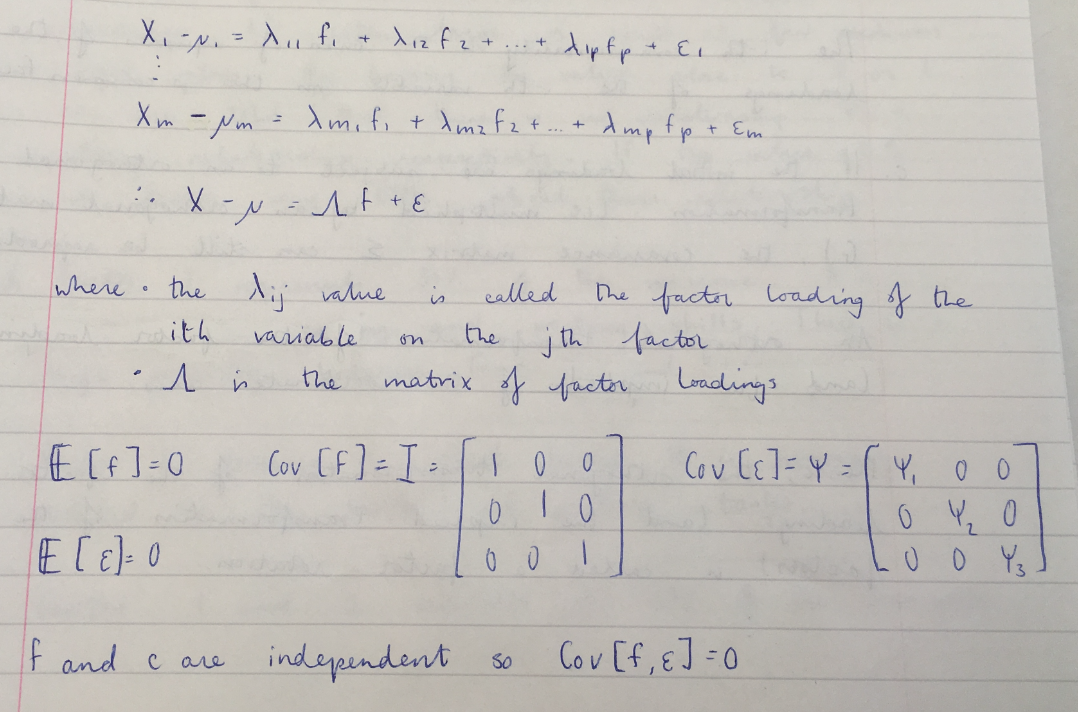
PCA makes no assumptions about the form of the covariance matrix whereas factor analysis assumes that the data comes from a well-defined model where specific assumptions hold e.g. E[f] = 0, E[ε] = 0.

The two analyses are often performed together. For example, you can conduct a PCA to determine the number of factors to extract in a factor analysis study.

## Explain the Factor model for multivariate data i.e. show how such multivariate data relates to the common factors and the specific factors, and indicate any assumptions about the expectation and covariance of these factor terms.

The observable random vector has mean μ and covariance matrix Σ. The factor model states that X is linearly dependent upon a few unobservable random variables called common factors and m additional sources of variance called specific factors.

Hence:



## In relation to the variance of a multivariate random variable, explain what is meant by the communality and uniqueness within FA and show how these terms are related.

The variance of λ can be split into two parts. The first portion of the variance for the i-th component arises from the m common factors and is referred to as the i-th communality. The remainder of the variance for the i-th component is due to the specific factor, and is referred to as the uniqueness.

Denoting the i-th communality by hi2, then:

The i-th communality is the sum of squares of the loadings of the i-th variable on the p common factors.

## Explain what is meant by a factor rotation, why it is relevant within factor analysis and state the objective of the varimax rotation.

If the initial loadings are subject to an orthogonal transformation (i.e. multiplied by an orthogonal matrix G), the covariance matrix Σ can still be reproduced. An orthogonal transformation corresponds to a rigid rotation or reflection of the coordinate axes.

Hence, the orthogonal transformation of the factor loadings (and the implied transformation of the factors) is called a factor rotation. From a mathematical viewpoint, it is irrelevant whether Λ or Λ\* = ΛG is reported since…

However, when the objective is statistical interpretation, it may be that one rotation is more useful than alternatives. Factor rotation allows us to explain the covariance matrix in fewer dimensions. The varimax procedure selects the orthogonal transformation G maximizing the sum of the column variances across all factors j = 1, … , p.

We aim for a rotation that makes the squared loadings either large or small i.e. few medium sized values. By having the values close to 0 or 1, we are told that there is no relationship or a strong relationship respectively. If the value = 0.5, we learn very little about this relationship.

## Describe Multi-Dimensional Scaling.

MDS attempts to find a map in a lower dimension so that inter-point dissimilarities match those in the original dissimilarity matrix as much as possible. We want to do this in a low dimension (2 or 3).

There are various ways that we implement MDS.

In the metric case, the lower dimensional distances arise as a result of applying a continuous monotonic function to the distances in the original dimensions. The results are the same between classical metric scaling with Euclidean distance and PCA.

In the non-metric version, only the rank order of the distances needs to be preserved.

## Explain the meaning and role of the stress value for a MDS analysis and explain how this value is found.

The stress is a measure of distortion between the original dissimilarities and what has been created.

Where

* is the distance between i and j in the lower dimensional plot
* is the distance between i and j in the dissimilarity matrix (original data)

## Compare Classical MDS and PCA.

Both are dimension reduction techniques and use eigen decomposition.

|  |  |
| --- | --- |
| Classical MDS | PCA |
| Computes eigen decomposition on a function of the dissimilarity matrix. | Computes eigen decomposition on the covariance/correlation matrix. |
| Seeks to match inter-point dissimilarities as closely as possible. | Seeks uncorrelated linear combinations which account for maximum variance. |
| Uses Euclidean distance. | Uses several distance measures. |

## Given a matrix D detailing the dissimilarities between any two multivariate observations xi and xj, explain how Classical MDS can be used to obtain a lower dimensional co-ordinate array X.

Note: Classical uses Euclidean distance and the objective is to minimize the stress value.

Step 1 – Obtain the dissimilarities {}

Step 2 – Form B, each element of which is given by , with i representing the origin of all observations.

Step 3 – Create matrix Λ from the eigenvalues λ1, … , λn-1 and the matrix V from the associated eigenvectors v1, … , vn-1 of B.

Step 4 – Choose an appropriate number of dimensions d using a suitable measure.

Step 5 – The coordinates of the n required points that are used to represent the n observations in d-dimensional space are given by for i = 1,…,n and j = 1,…,d.

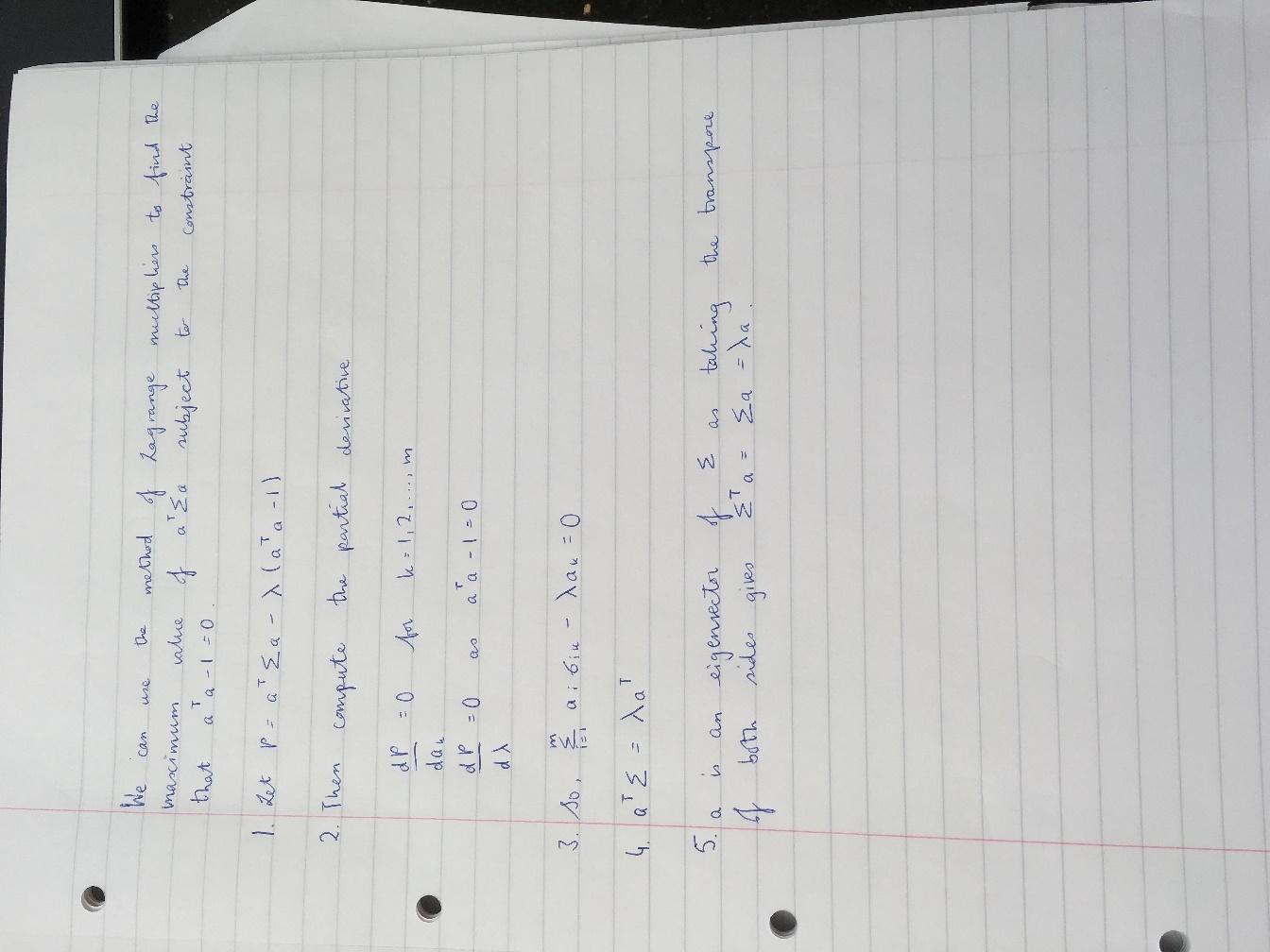
## Explain the role and application of a Procrustes Analysis within the context of MDS.

Procrustes Analysis matches one MDS configuration with another by dilation, rotation, reflection and translation. Taking the 2 lower dimensional coordinate matrices that arise from the two MDS analysis, a sum of squared distances between corresponding points in the different representations can be calculated.

This analysis requires calculation of a dilation matrix, a translation factor and an orthogonal matrix (which causes a rotation and/or reflection).

When these calculations are applied on one representation, the sum of squared distances between corresponding points in the two representations is minimized. The minimized sum of squares is called the Procrustes Sum of Squares and is interpreted as a measure of agreement between the two representations.

## Using the method of Lagrange multipliers, show that for a covariance matrix Σ, the expression , subject to the constraint , is maximized when a is an eigenvector of Σ.



# Classification

## Provide a motivation for the use of logistic regression, rather than linear regression.

Logistic regression is used when we have binary variables. Logistic regression ensures that regardless of the covariance we get, any estimate we get will fall within the interval 0 and 1. In linear regression, we cannot guarantee that. We could potentially get new observations of the other variables that would lead to values outside the region of 0 and 1 and hence, would lead to estimated probabilities which are not valid probabilities.

## Give the formula for P (Survived = 1).

Where

* α is an estimate of the intercept

## For data with only binary entries, explain the Simple Matching (Hamming) and Jaccard dissimilarity measures. Give a motivation for using the Jaccard measure instead of the Hamming measure.

|  |  |  |
| --- | --- | --- |
|  | **1** | **0** |
| **1** | a | b |
| **0** | c | d |

Simple Matching:

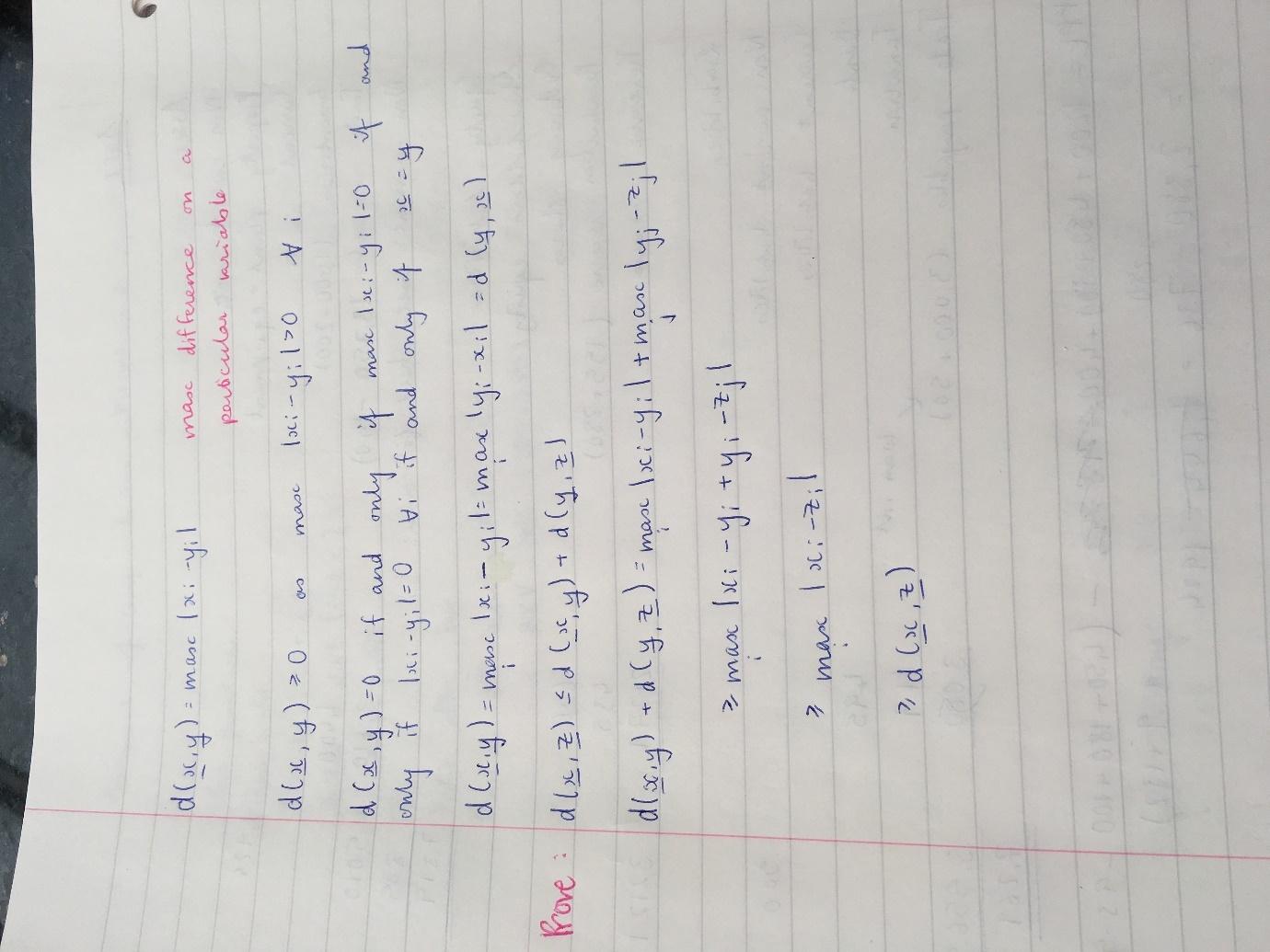
Where

* a is the number of times both observations share a value 1 in the same position
* b is the number of times one data point has a value 0 while the other has a value 1 in the same position
* c is the number of times one data point has a value 0 while the other has a value 1 in the same position
* d is the number of times both observations share a value 0 in the same position

Jaccard:

We would use Jaccard as it eliminates the problem where double absence doesn’t indicate similarity. For example, there are no kangaroos in Ireland or in the Antarctic but that doesn’t mean they are similar countries.

## Give the definition of the maximum dissimilarity measure and show how this measure satisfies the common dissimilarity properties of non-negativity, symmetry and the triangle inequality.



## Describe the role and use of interactions in logistic regression.

Interactions allow for the effect of one covariate to be altered depending on the value of another covariate. For example, household income may be important in predicting whether a child takes the Leaving Cert if they are a boy but not if they are a girl. Relevant and significant interactions can greatly increase predictor performance.

## Describe the modelling assumptions underpinning Linear Discriminant Analysis and Quadratic Discriminant Analysis.

LDA and QDA are used if it is assumed that there exists a set of k groups within the data and that there is a subset of the data that is labelled i.e. whose group membership is known. We want to be able to use knowledge of labelled data in order to classify the group membership of unlabelled data.

Discriminant analysis assumes that observations from group k follow a multivariate normal distribution with mean μk and covariance Σk. That is:

Discriminant analysis also assumes values for πk = P (x ϵ k) which is the proportion of population objects belonging to class k. Note that . Typically, is used. are sometimes referred to as prior probabilities.

Using all of this, we can compute and assign the data points to groups so as to maximize this probability.

## Outline the difference between logistic regression and linear discriminant analysis.

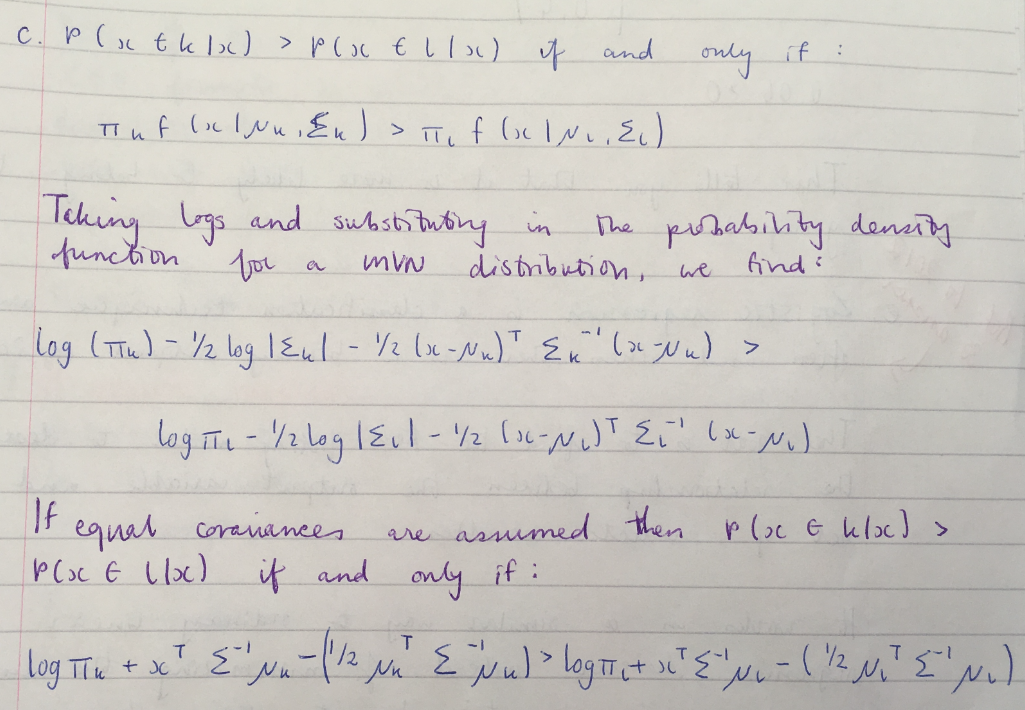
In LDA, the decision boundary between class k and class l is given by:

In logistic regression, this is:

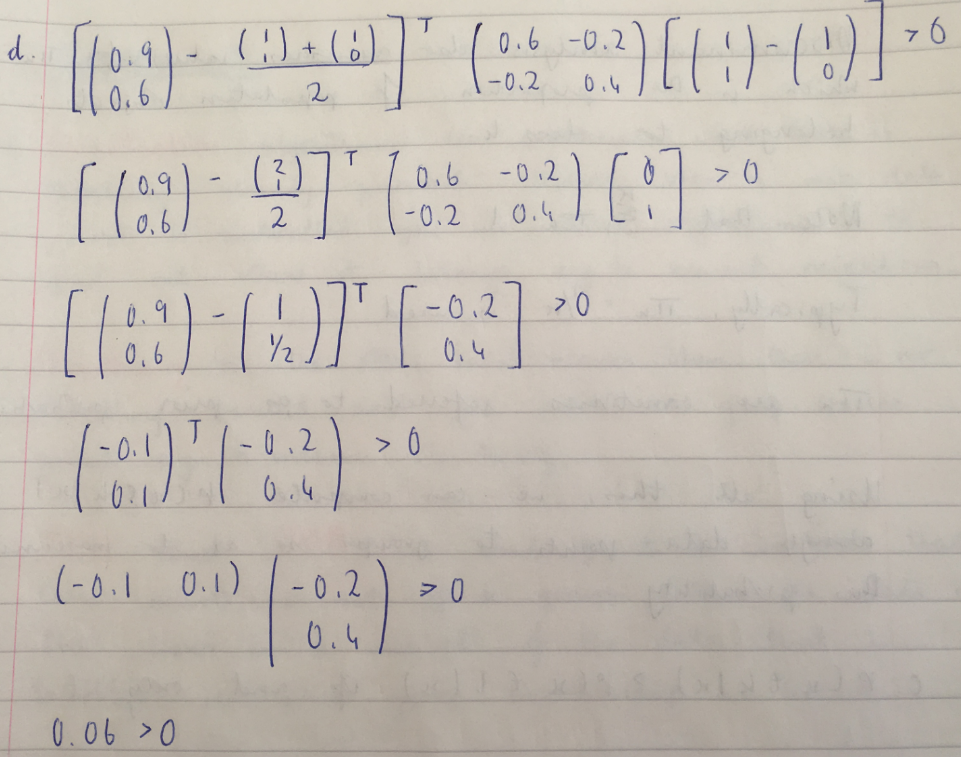
Both of these are linear expressions in x, however, the former is based on a multivariate normal distributional assumption concerning x and uses Bayes rule with prior probabilities , while the latter does not make a multivariate modelling assumption and its parameters are found through maximum likelihood estimation.

Similar to LDA and QDA, logistic regression is a parametric classification technique making distributional assumptions over the data so as to allow a probabilistic quantification of class assignment. Logistic regression is much more widely used in the way that we can use binary data and uneven data.

## Show that, in the case of LDA with equal prior probability of group assignment, the probability that a point x belongs to group 1 is larger than the probability of belonging to group 2 if:



## Group 1 has a mean vector , whilst group 2 has a mean vector , with pooled precision matrix . To which group would LDA assign the point ?



This tells you that it is more likely to belong to group 1.

## Compare k-nearest neighbours and LDA.

K-nearest neighbours classification is a non-parametric/distribution free method of assigning group membership. In other words, k-nearest neighbours classification makes no assumption on the spread of data within each class.

The consequence of this is that class assignment is fixed, with no measurement of uncertainty concerning any particular assignment. Classification techniques that make distributional assumptions at least allow quantification of the uncertainty in group membership.

Unlike k-nearest neighbours, both LDA and QDA are model based classifiers where is assumed to follow a MVN distribution:

1. The model based assumption allows for the generation of the probability for class membership
2. The MVN assumption means that groups are assumed to follow an elliptical shape.

## Explain what is meant by cross validation in the context of discriminant analysis.

The cross validation routine works by omitting each observation one at a time, recalculating the classification function using the remaining data and then classifying the omitted observation. It measures the performance of the model.