Today we consider the problem of classification or discriminant analysis.

In particular, we will consider the simple $k$-nearest neighbours ($k$NN) approach to classification.
• Consider the olive oil data mentioned in the first class.

• The following dendrogram arises from a complete-linkage hierarchical clustering with Euclidean dissimilarity analysis on the standardized data.

• There appears to be some grouping of the data.
Olive Oil Groups

- If we considered there to be three groups, then we can see if these groups correspond to the known geographic areas of where the oils were collected.

- A cross tabulation shows some agreement (cluster 1 is identified with area 2, cluster 2 with area 3, and cluster 3 with area 1).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
<td>3</td>
<td>226</td>
</tr>
<tr>
<td>Area</td>
<td>2</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>37</td>
<td>114</td>
</tr>
</tbody>
</table>

- The Rand Index is 0.736 and the Adjusted Rand Index is 0.446.

- Can we do better?
Classification

- The olive oil data was collected from three distinct regions of Italy, with their fatty acid compositions recorded.

- For example, a plot of the oleic and linoleic values are shown below (coloured by region of origin).
Principal Components

- A plot of the first two principal components gives a summary of the data (and accounts for 68.6% of the variance).
Classification

• Can we use the data and the known area information to determine the geographic origin of new olive oil samples?

• This is the aim of classification/discriminant analysis.

• Many methods of classification exist. For example, within this course we will consider:
  – Linear discriminant analysis
  – Quadratic discriminant analysis
  – \( k \)-nearest neighbours classification.
  – Logistic Regression.

• Today we will consider the simplest of these: \( k \)-nearest neighbours classification.
$k$-Nearest Neighbours Classification

- As opposed to alternative classification techniques we will later consider, $k$-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 do make distributional assumptions at least allow quantification of the uncertainty in group membership.
\(k\)-Nearest Neighbours Classification

- Consider a subset of the olive oil data.
- The plot of the first two principal component loadings for a sample of five oils from each area is given below.

- Suppose we now had a new observation of unknown origin. How do we infer where it came from?
**$k$-Nearest Neighbours Classification**

- What could we say about five new observations with first two principal component loadings as shown below?
$k$-Nearest Neighbours Classification

- $k$-Nearest Neighbours simply looks at the $k$ closest points of known origin to the point of unknown origin.

- The point is then classified as belonging to the group which contains the most of these $k$ points.

- Remember there can be issues with the scaling of the axes.

- The results are not invariant to the scaling of the original variables, nor to the method in which distance is calculated, as these are likely to change the nearest neighbours for any particular point.
Salmon Data

- Salmon \((n = 100)\) were collected from Alaska (○) or Canada (△).
- Numeric variables ‘Freshwater’ and ‘Marine’ were also recorded.
Salmon Data

- The salmon fishery is a valuable resource for both the US and Canada. As it is a limited resource it must be managed efficiently. Also, as more than one country is involved, problems must be solved equitably, so Alaskan fishermen cannot catch too many Canadian salmon and vice versa.

- These fish are born in freshwater and after a year or two swim to the ocean. After a couple of years they return to their birth place to spawn and die. As they are about to return they are harvested while still in the ocean. To regulate catches samples of fish taken during harvest must be identified as coming from Alaskan or Canadian rivers.

- This information is measured in the growth rings on their scales, which are typically smaller for Alaskan-born than Canadian-born salmon.

- Freshwater is diameter of rings for the first-year freshwater growth, whilst Marine is diameter of rings for the first-year marine growth.
\textit{k}-Nearest Neighbours

- We can employ the \textit{k}-nearest neighbours method so as to use this data for classifying a future Salmon of unknown origin.

- For each new data point consider the class assignment of its \textit{k} closest neighbours.

- Then the \textit{k}-nearest neighbours method classifies the new observation as belonging to the class that was most prevalent in those \textit{k} labelled neighbours.
Example: Salmon Data

- Suppose six new salmon are observed and are included in the plot as below.
Example: Salmon Data

- The classification changes as a function of $k$ in the following way:

<table>
<thead>
<tr>
<th></th>
<th>k</th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>p6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- Alaska=1 and Canada=2.
Choosing $k$

- Classification will vary with the choice of $k$, e.g., Salmon 5.

- One way to choose $k$ is to split the labelled data into three parts:
  - **Training Set:** Points whose labels are used to classify unlabelled points.
  - **Test Set:** Points that are treated as unlabelled that are used to find the value of $k$ that is best for classification.
  - **Validation Set:** Remaining points that are also treated as unlabelled, used to estimate the classification error using the best $k$ identified by the test set.

- We will apply this rule to the Salmon data using a (50%, 25%, 25%) split.
Choosing $k$

- The proportion incorrectly classified at the test step is plotted below as a function of $k$.

- The value $k = 8$ achieves the best rate and for the validation set this achieves a 96% correct classification rate.
Training Data
The cross-hair indicates a misclassified point.
The cross-hair indicates a misclassified point.
Why validate?

- The classification rate for the test set can overestimate the classification rate in the validation set. This is because the value of $k$ is chosen specifically for the test set, and may not be perform as well when applied to another unlabelled data sample.

- The validation set is not used at any stage of the model fitting, and hence offers a more reasonable estimate of the correct classification rate.

- In the Salmon data we achieved the same classification rate for both the test data and the validation data. In general this will not be the case.
Training, Test & Validation Data

- The idea of splitting a data set into three parts to find the best model and to evaluate its classification rate is a general statistical technique.

- In many situations the following methodology can be applied:
  - **Training data**: Data used to fit several models.
  - **Test data**: Data used to compare competing models and choose best model.
  - **Validation data**: Data used to assess the performance of the chosen model.

- A split of 50% training data, 25% test data and 25% validate data is common.
Cross-Validation

- A related idea approach for choosing $k$ is to use cross-validation.

- In this problem (leave-one-out) cross-validation is used as follows:
  - For each value of $k$ remove a single data point and determine if that data point would be correctly classified knowing the labels of all other data points. Do this for each observation in the data set.
  - If there are 100 observations, this means making 100 classifications of 1 point using the labels of the other 99 points.
  - Select the value of $k$ that has the best classification rate.

- Another option is to split the data into, e.g, ten subsets. Then we determine how many data points in one of the subsets would be correctly classified using the labels of the other subsets. Do this for each of the ten subsets of the data. This is called 10-fold cross validation (and $k$-fold cross validation in general).
Cross-Validation

- Leave-one-out validation on the Salmon data leads to the following:

\[
\begin{array}{ccccccc}
0 & 20 & 40 & 60 & 80 & 100 \\
20 & 40 & 60 & 80 & 100 \\
\end{array}
\]

- The best results are for \( k = 9 \).