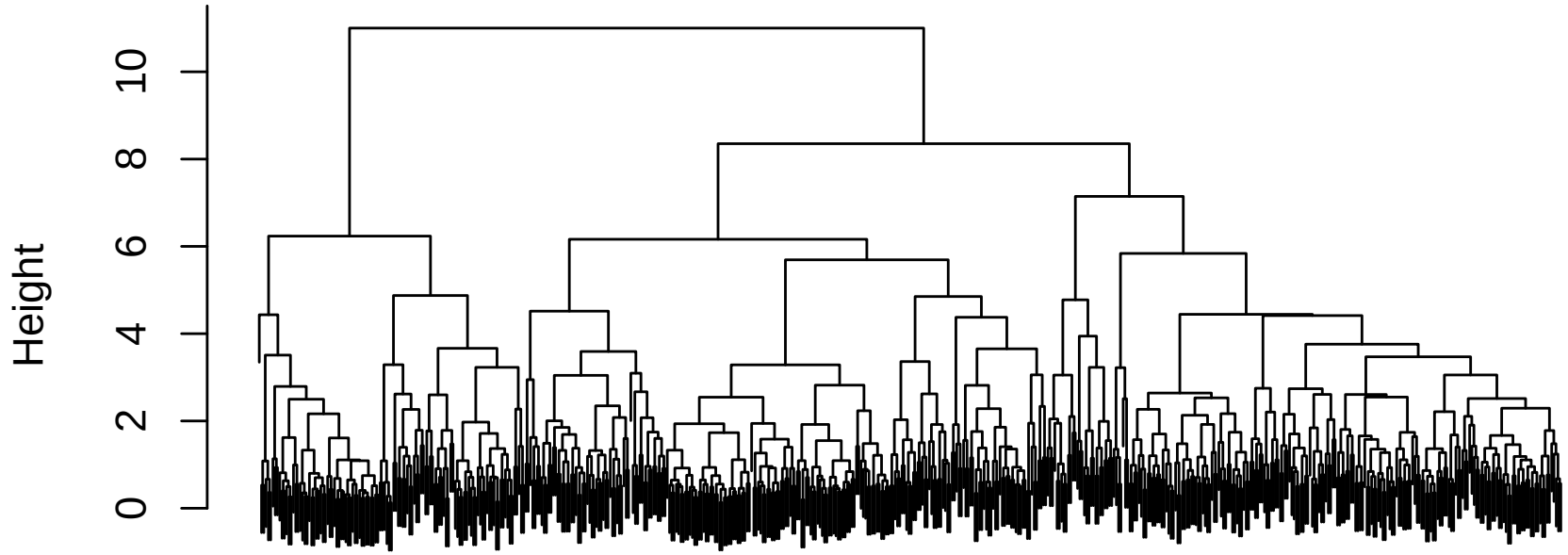
Multivariate Analysis (Slides 7)

- Today we consider the problem of classification or discriminant analysis.
- In particular, we will consider the simple k-nearest neighbours (kNN) approach to classification.

Olive Oil Example

- 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

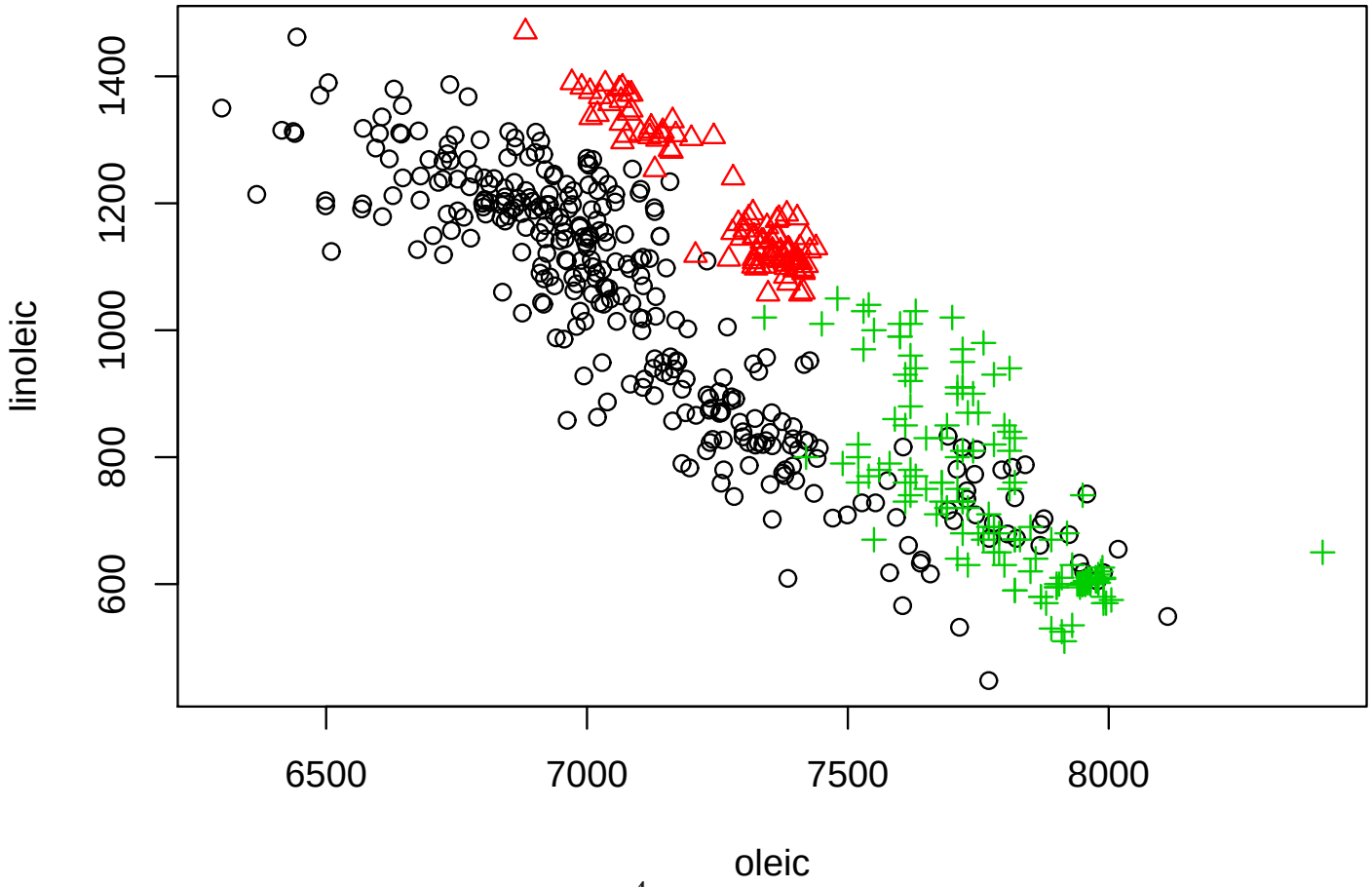
- If we considered there to be three groups, then we can see if these groups correspond to the known geographic areas of where the oils were collected.
- A cross tabulation shows some agreement (cluster 1 is identified with area 2, cluster 2 with area 3, and cluster 3 with area 1).

		Cluster		
		1	2	3
	1	94	3	226
Area	2	98	0	0
	3	37	114	0

- The Rand Index is 0.736 and the Adjusted Rand Index is 0.446.
- Can we do better?

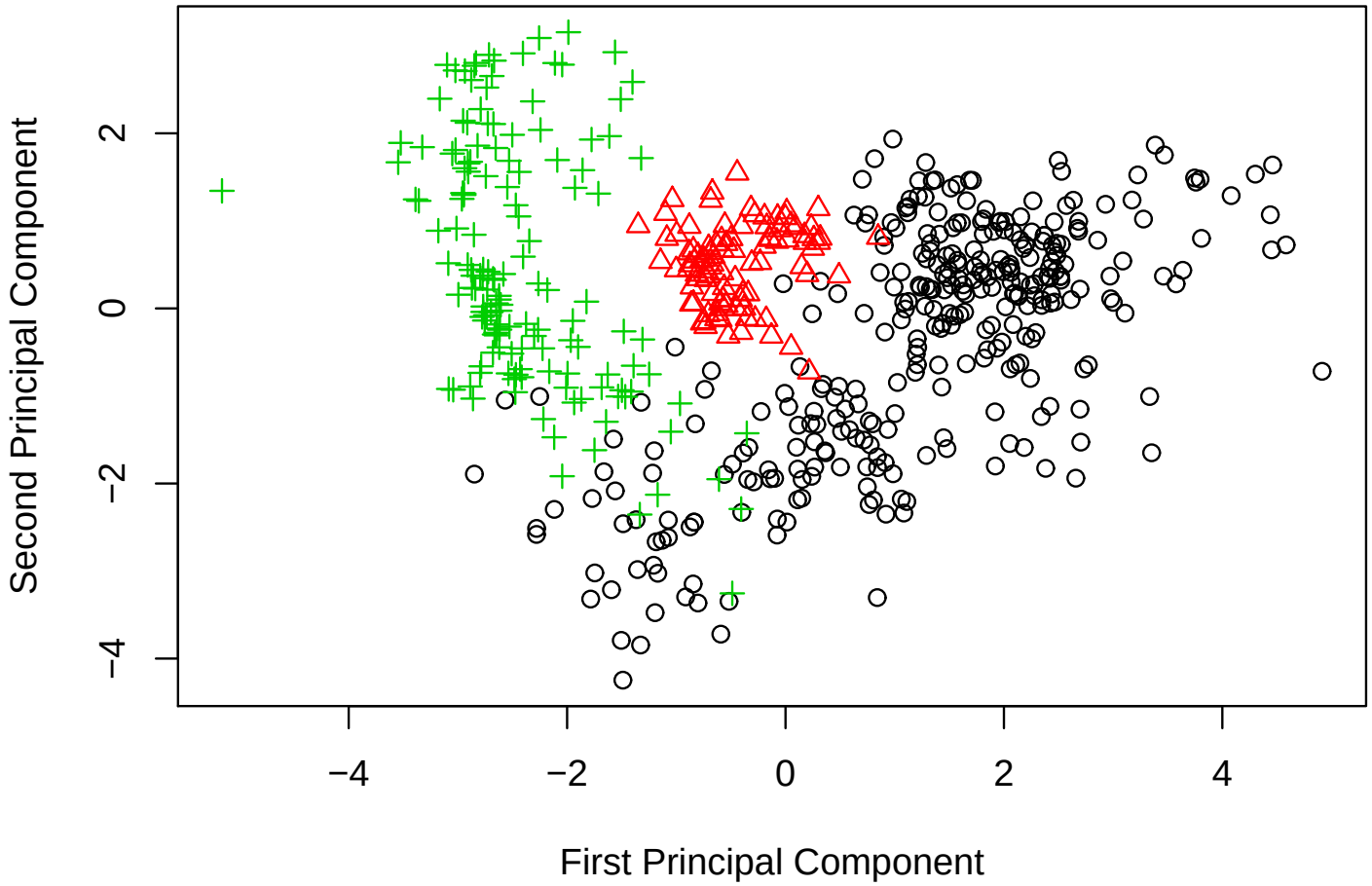
Classification

- 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

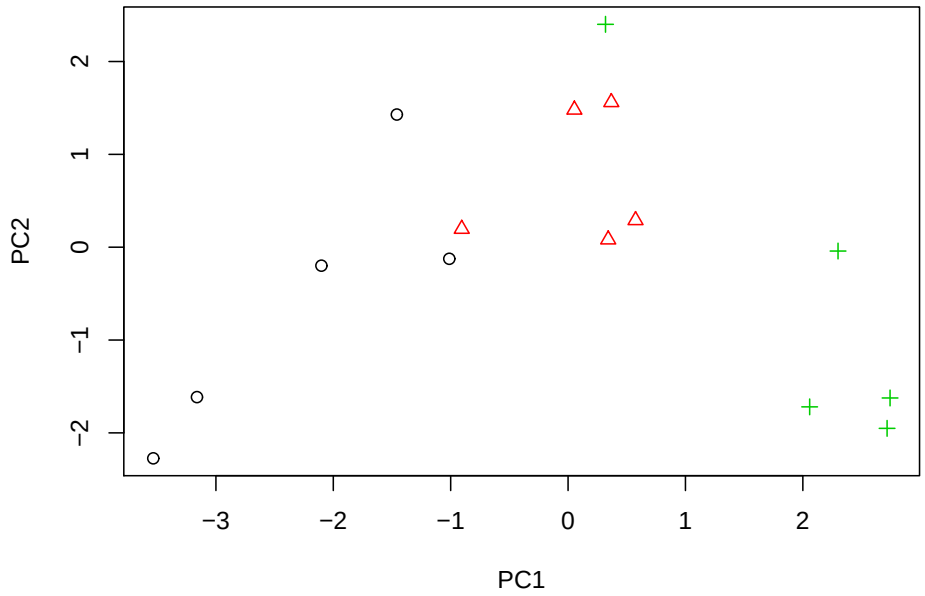
- Question: Can we use the data and the known area information to develop a way of determining the geographic origin of new olive oil samples?
- This is the aim of classification/discriminant analysis.
- Many methods of classification exist. For example, within this course we will consider:
  - Linear 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

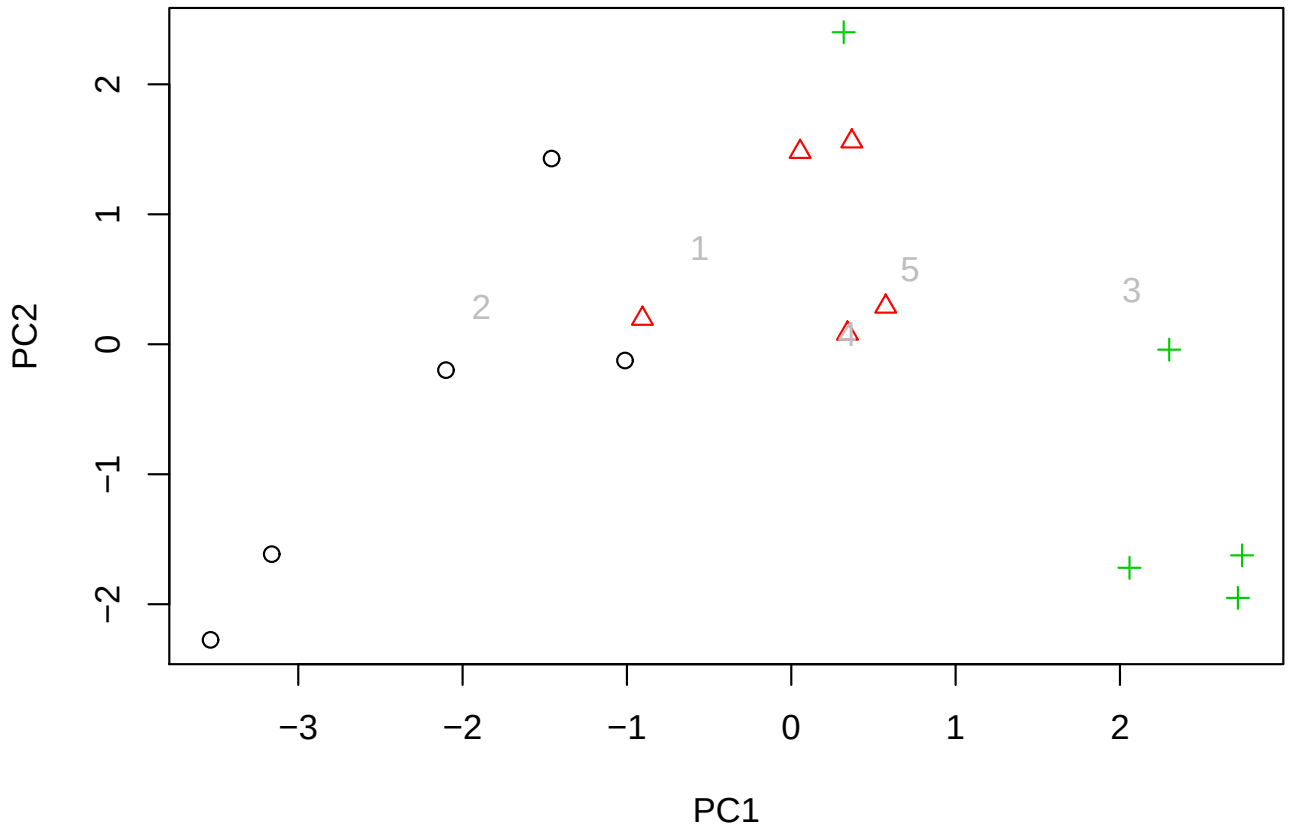
- Consider a subset of the olive oil data.
- The plot of the first two principal component loadings for a sample of five oils from each area is given below.



- Suppose we now had a new observation of unknown origin, how can 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?

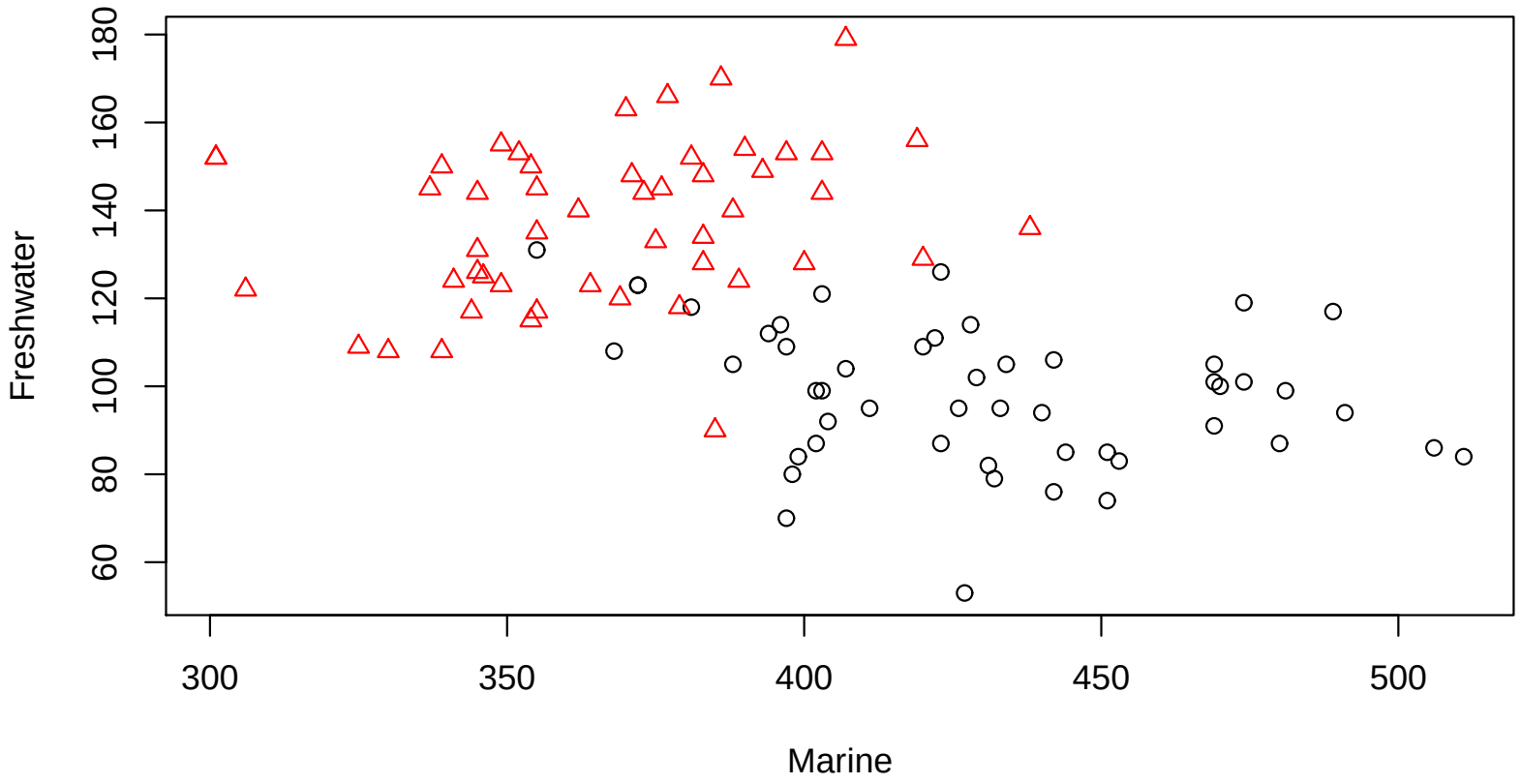


k-Nearest Neighbours Classification

- 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.
- If looking by eye remember there may be issues with the scaling of the axes.
- Of course the results are not invariant to the scaling of the original variables (consider standardization?), nor to the method in which distance is calculated, as these are likely to change the nearest neighbours for any particular point.

Salmon Data

- Consider the following Salmon data (100 data points). For each observation a record relating to whether the Salmon was collected from Alaska (○) or Canada (△) is recorded, along with two numeric variables titled 'Freshwater' and 'Marine'.



Salmon Data

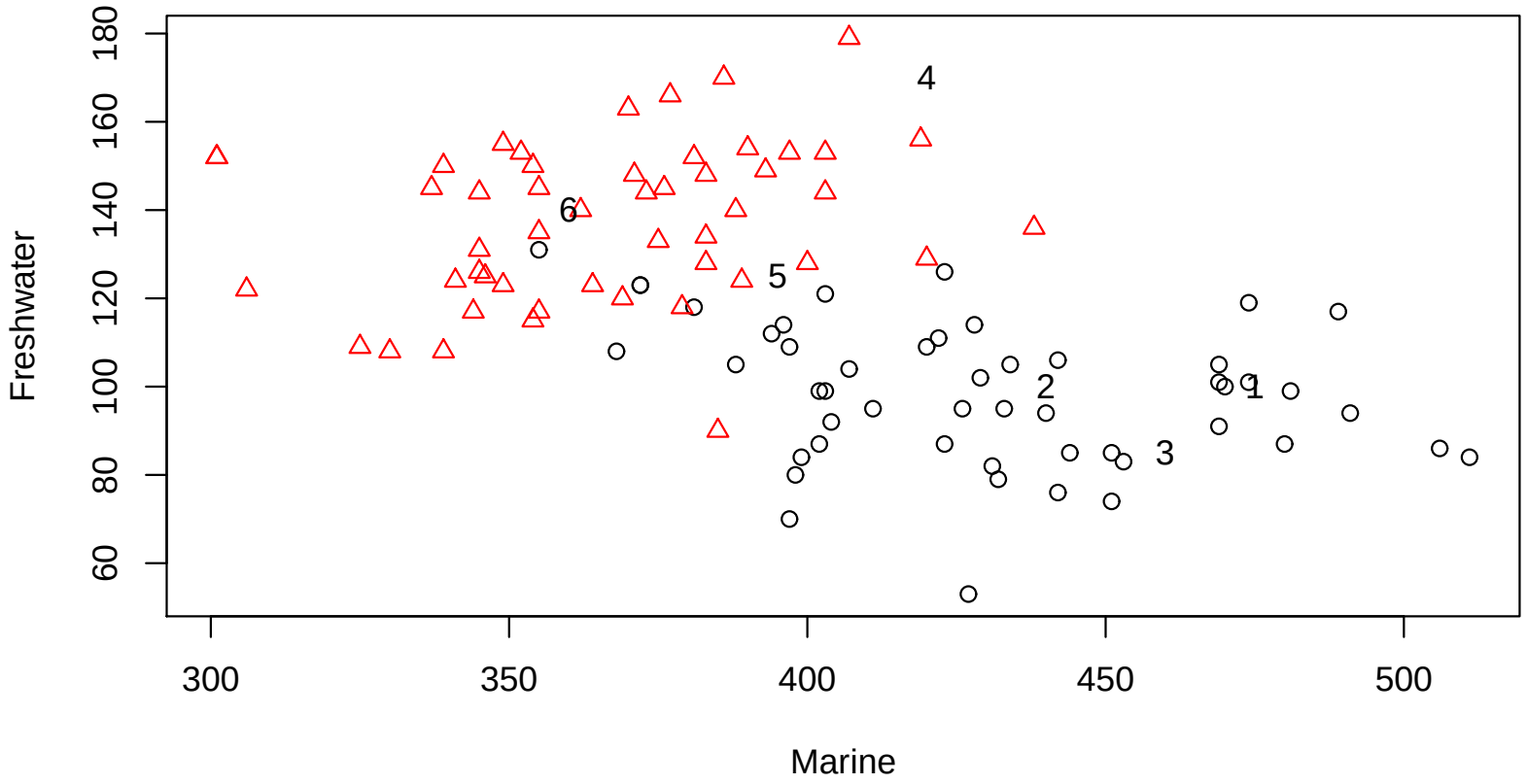
- The salmon fishery is a valuable resource for both the US and Canada. As it is a limited resource it must be managed efficiently. Also, as more then one country is involved, problems must be solved equitably, so Alaskan fisherman cannot catch too many Canadian salmon and vice versa.
- These fish are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

k-Nearest Neighbours

- We can employ the 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 k closest neighbours.
- Then the k-nearest neighbours method classifies the new observation as belonging to the class that was most prevalent in those k labelled neigbours.

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:

	k	p1	p2	p3	p4	p5	p6
[1,]	1	1	1	1	2	2	2
[2,]	2	1	1	1	2	2	2
[3,]	3	1	1	1	2	2	2
[4,]	4	1	1	1	2	1	2
[5,]	5	1	1	1	2	2	2
[6,]	6	1	1	1	2	1	2
[7,]	7	1	1	1	2	2	2
[8,]	8	1	1	1	2	2	2
[9,]	9	1	1	1	2	1	2
[10,]	10	1	1	1	2	1	2

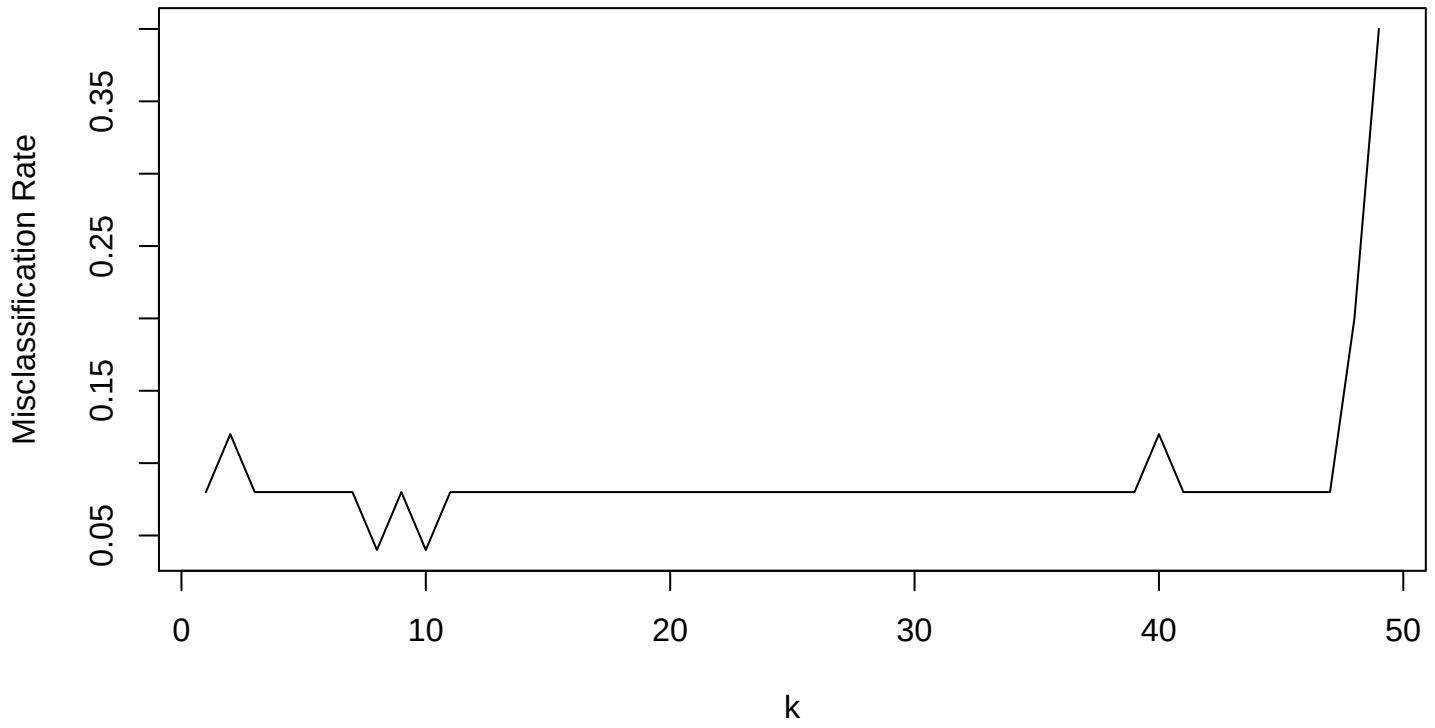
- Alaska=1 and Canada=2.

Choosing k

- The classification varies with k, *e.g.*, Salmon 5.
- One possible approach for choosing k is to split the labelled data into three parts:
  - **Training:** Points whose labels are used to classify unlabelled points.
  - **Test:** Points we know the labels for but which are considered unlabelled in order to find the value of k that is best at classifying them.
  - **Validation:** Remaining labelled points that are considered unlabelled in order to estimate the classification error for the best k from the Test step.
- We will apply this rule to the Salmon data using a $(50\%, 25\%, 25\%)$ split.

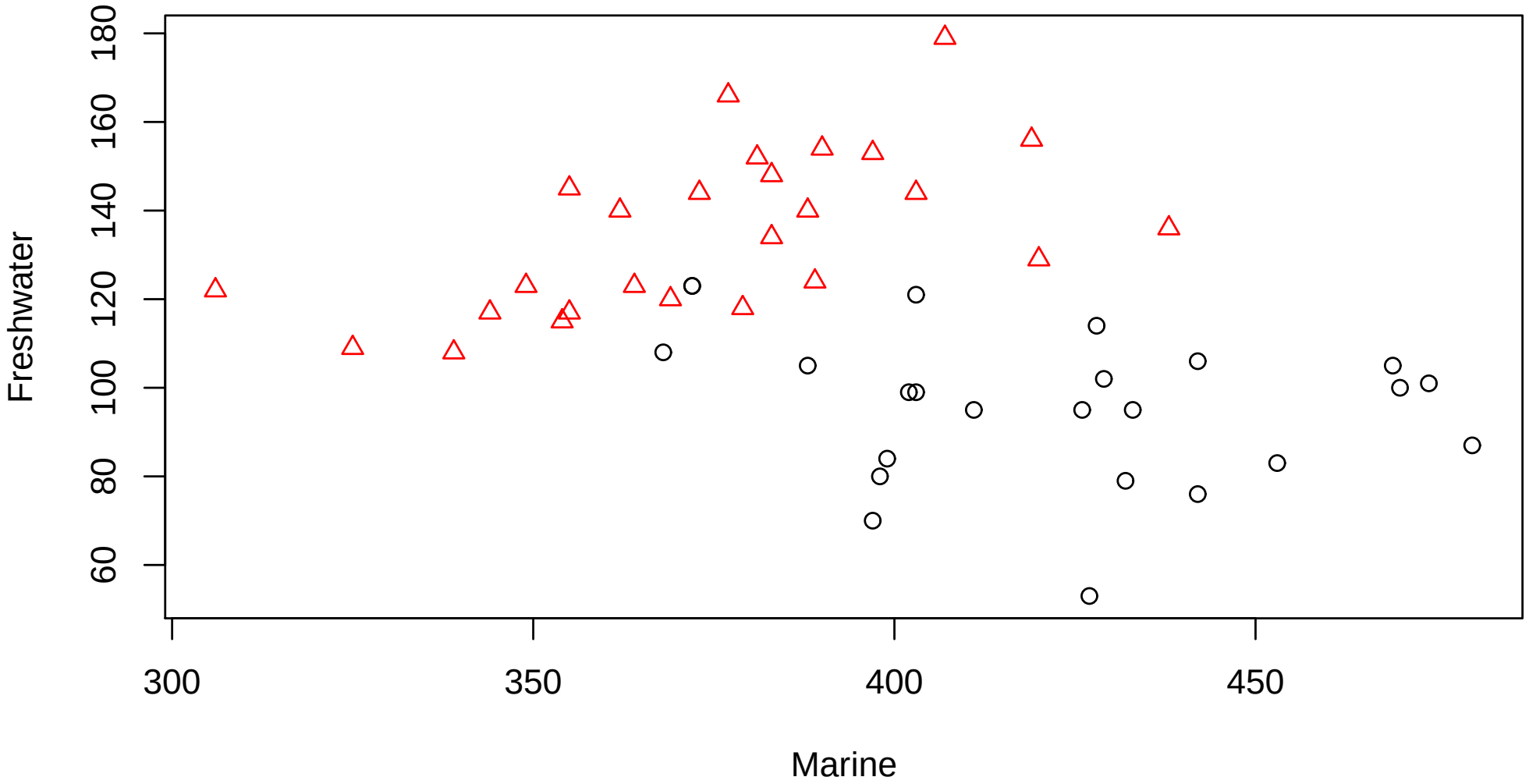
Choosing k

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

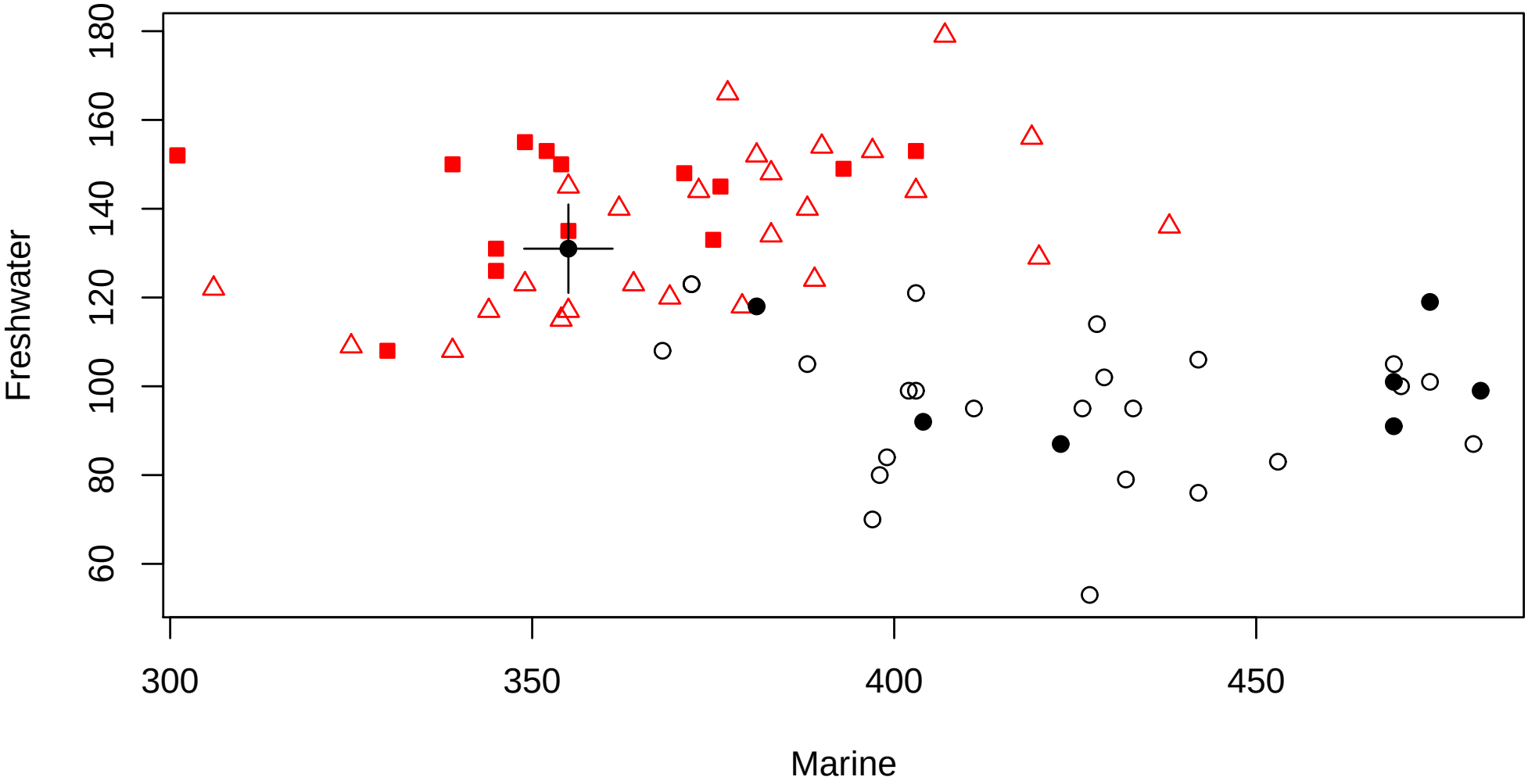


- The value $k = 8$ achieves the best rate and in the validation phase this achieves a 96% correct classification rate.

Training Data

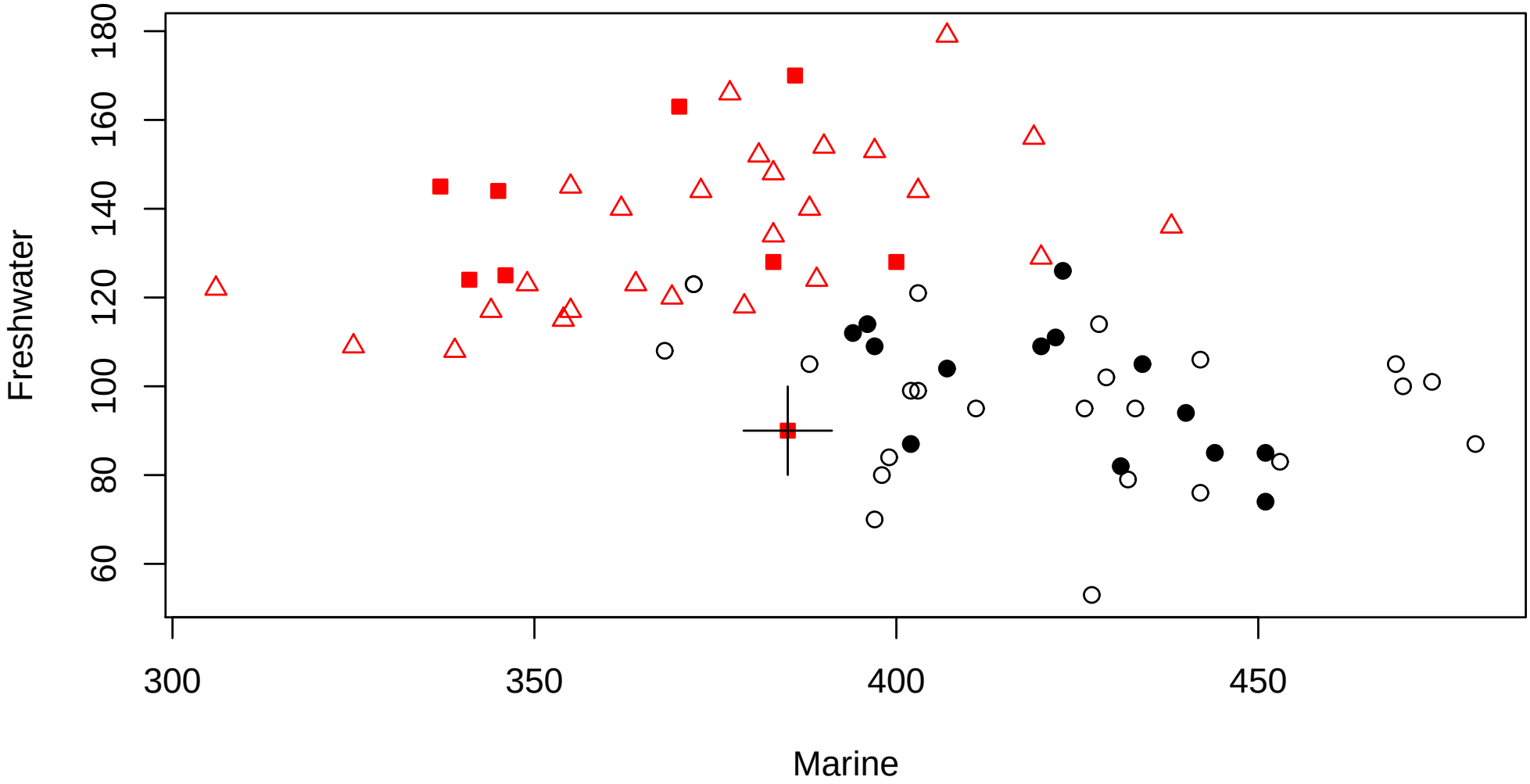


Training & Test Data



- The cross-hair indicates a misclassified point.

Training & Validation Data



- The cross-hair indicates a misclassified point.

Why validate?

- The correct classification rate for test data typically overestimates the percentage correct classifications in validation data. This is because the value of k is chosen specifically for the test set, and may not be representative for another unlabelled sample.
- The validation data 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

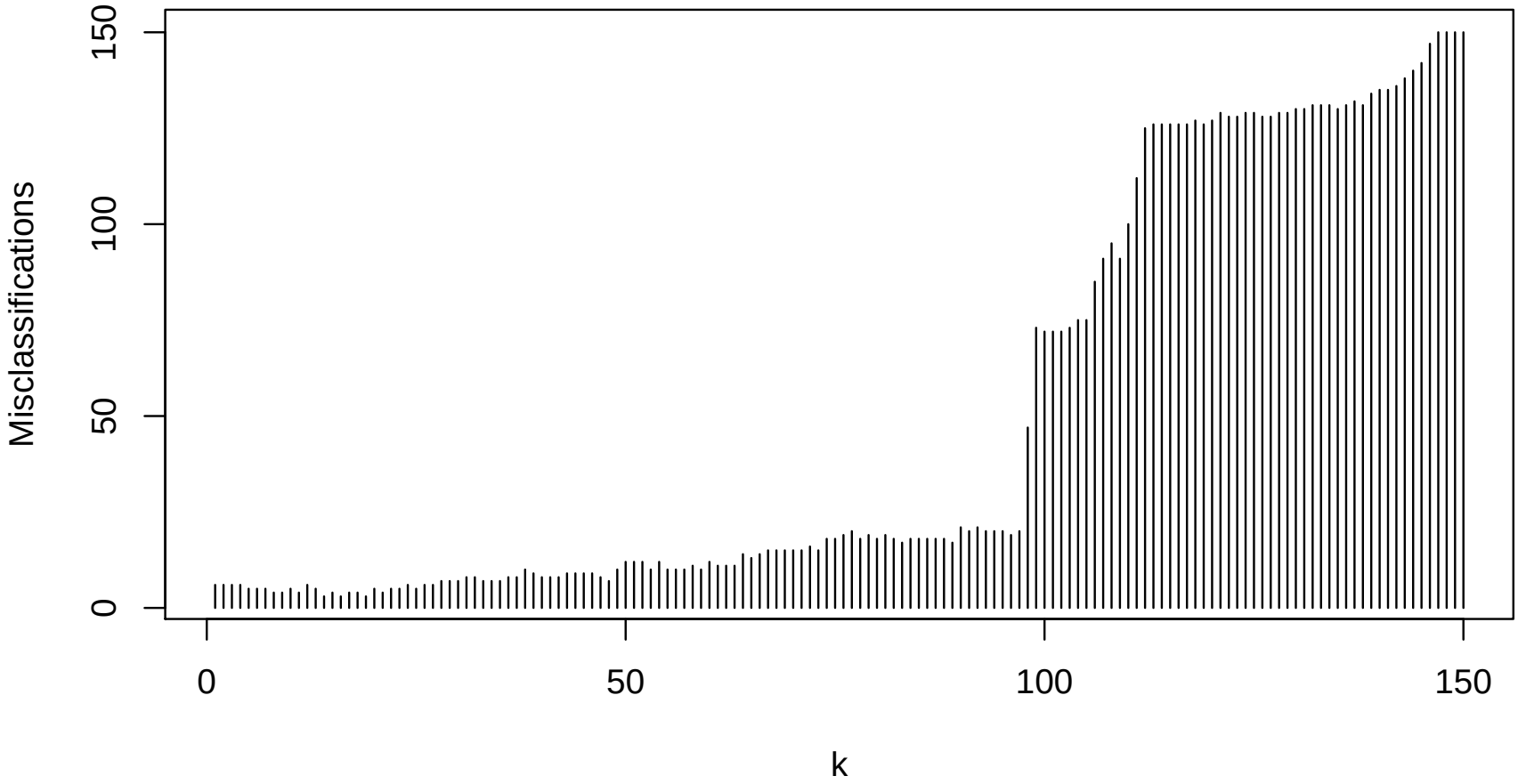
- The idea of breaking a data set into three parts to find the best value of k and to also estimate the correct classification rate is a general statistical technique.
- In many statistical modelling situations the following methodology is applied:
  - **Training data:** Data used to fit a model.
  - **Test data:** Data used to compare competing models.
  - **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

- An alternative approach to choosing k is to use the statistical technique of *cross-validation*.
- In this problem (leave-one-out) cross-validation would be used as follows:
  - For each value of k remove each data point and determine if that data point would be correctly classified knowing the labels of all other data points. If there are 100 data points this means making 100 classifications of 1 point based on the labelling of the other 99.
  - For example, if there are 3 data points $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$. Then we see if we correctly classify $\mathbf{x}_1$ given labelling for $\mathbf{x}_2$ and $\mathbf{x}_3$, correctly classify $\mathbf{x}_2$ given labelling for $\mathbf{x}_1$ and $\mathbf{x}_3$, and if we correctly classify $\mathbf{x}_3$ given labelling for $\mathbf{x}_1$ and $\mathbf{x}_2$. Each value of k will either get none correct, one correct, two correct, or all correct.
  - Select the value of k that has the best classification rate.

Cross-Validation

- Cross-validation on the Iris data leads to the following:



- The best results are for $k = 14$.