Additional Examples 2020

D Wong

December 2020

The examples below are in addition to the worked examples from the course, to consolidate your understanding. Solutions will follow in due course.

1 Week 3

Q1

Consider two data points with three features:
A (class 0): $x_1 = 1$, $x_2 = 2$, $x_3 = 3$
B (class 1): $x_1 = 4$, $x_2 = 2$, $x_3 = 2$

The Minkowski distance when p=3 is defined as:

$$\sum_{i=1}^{n} (|x_i - y_i|^3)^1/3$$

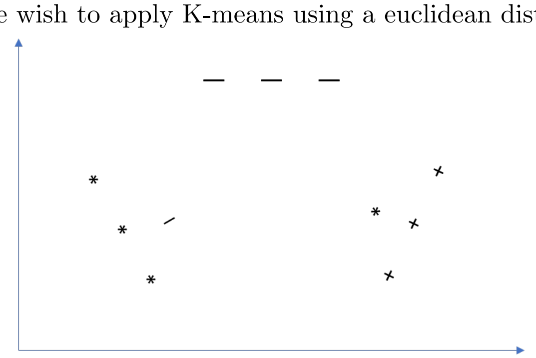
- calculate the Euclidean distance and the Minkowski distance (p=3) between data points A and B
- Consider a new data point,

  C: $x_1 = 2$, $x_2 = 1$, $x_3 = 2$

  By applying K-nearest neighbours, using a euclidean distance metric and K=1, determine which class C belongs to.
- For a training data set containing N examples, in which N/2 belong to class 0 and N/2 belong to class 1, how many distance comparisons must be considered when trying to classify a new test data point?
- How might one reduce the amount of comparisons required? (this is out of scope, but should lead you to investigate extensions to K-NN in more detail

Q2

The following question uses the data set shown below. Data belongs to *three* classes, denoted +, - and *. We wish to apply K-means using a euclidean distance metric.



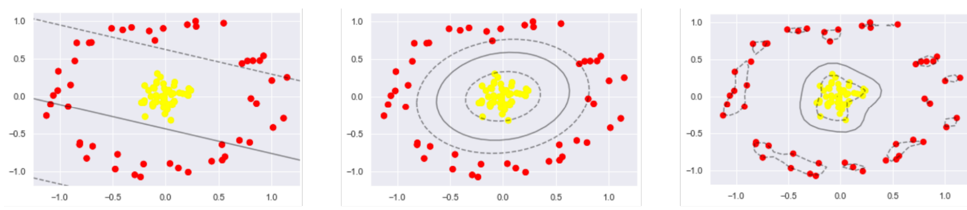
- What value(s) of K minimises the training set error for this data set?
- What value(s) of K minimise the leave-one-out cross validation error for this data set, and what is the value for the error in this case?
- Sketch the decision boundary in the case of K=1

Q3

- In the context of Support Vector Machines, describe a kernel. (n.b. the term 'kernel' has multiple uses in machine learning)
- The quadratic kernel is:

  $$K(x, x') =$$

  Show that this kernel is equivalent to a mapping between the raw feature space $[x_1, x_2]$ to a new implicit feature space with dimensions: $[x_1^2, x_2^2, \sqrt{x_1 x_2}, \sqrt{2}x_1, \sqrt{2}x_2, 1]$
- Which of the following decision boundaries could have been produced using the kernel in the previous question and why?



a b c

- (extension) The rbf kernel corresponds to an infinite dimensional implicit feature space. By considering a series representation of an exponential, show why this is the case.

2 Week 4

Q1

This question uses the data set below to predict whether someone is at high risk (H) or low risk (L) for car insurance, based on Sex (M/F) and car type (Cheap (C)/Medium (M)/Expensive (E))

car type	Sex	Insurance
C	M	H
C	F	L
M	M	H
M	F	L
E	M	H
E	F	H

- What is the entropy $H(High|car\_type)$?
- Draw the full decision tree that would be learned for this dataset.

Q2

A confusion matrix for a binary classifier, A, is presented below.

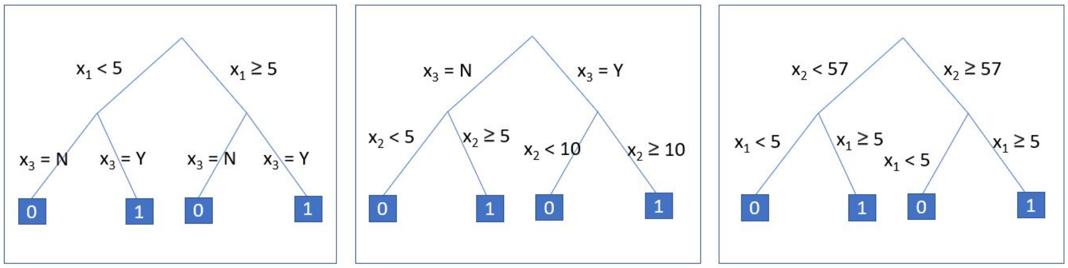
	Actual Pos	Actual Neg
Predicted Pos	450	50
Predicted Neg	0	0

- Is the accuracy a good measure of classifier performance? Why?
- Calculate the sensitivity and specificity of the classifier, and plot this point on an ROC plot
- For a second classifier, B, 90% of the positive class is correctly predicted, 90% of the negative class if correctly predicted. Plot this point on an ROC plot.
- Which classifier is better and why?

3 Week 5

Q1

- A random forest is created by combining the 3 decision trees below.



The random forest is used to classify a new data point into class 0 or class 1:

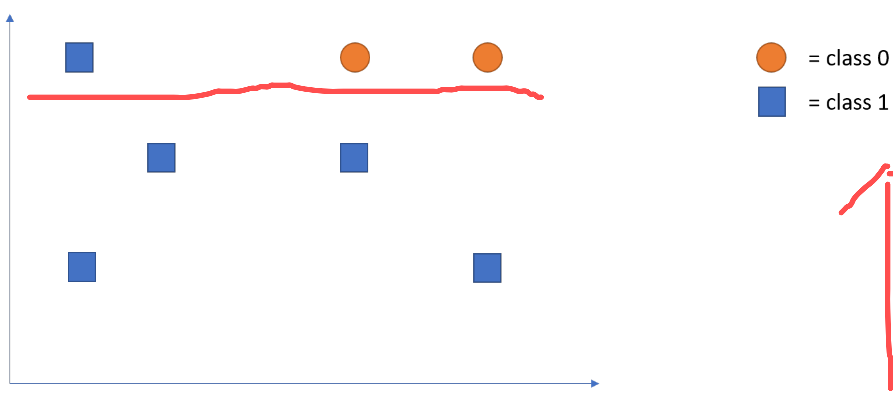
$$(x_1, x_2, x_3) = (10, 15, N)$$

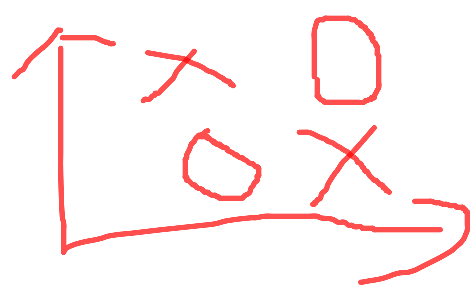
By determining the the predicted class for each tree, determine the predicted class for the random forest.

- How does the random forest differ from the bagging algorithm?
- What is the difference between random forest algorithms versus boosting algorithms?

Q2

- Consider the data set below. If we use Adaboost with a decision stump weak learner, draw the decision boundary in the first iteration, and circle any data points that will be up-weighted in the next iteration.





- Draw a simple dataset with 4 examples for which Adaboost will fail to give zero training error.
- Suppose you are running AdaBoost with 4 training examples. At the start of the current iteration, the four examples have the weights shown in the following table. Another column says if the weak classifier got them correct or incorrect. Determine the new weights for these four examples, and fill in the corresponding entries in the table.

	old weight	correct?	New weight
Example 1	0.16	correct	
Example 2	0.64	correct	
Example 3	0.08	incorrect	
Example 4	0.12	incorrect	

Q3

- Explain why feature selection is important in machine learning
- What is the Pearson correlation coefficient? How can it be used for feature selection?
- Calculate the Pearson correlation coefficient for the dataset below.

Number of visitors v (1000s)	2450	2480	2540	2420	2350	2290	2400	2460
Amount of money spent m (£ millions)	1370	1350	1400	1330	1270	1210	1330	1350