**Past exam questions. Question 3, classification:**

**2013, January paper**

Question 3:

The dataset below is an extract from a car dataset, containing attributes that describe each car. There is a binary class label, to Buy, yes or no.

Training data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Safety | Value For Money | Boot Size | Buy |
| 1 | High | Good | Large | Yes |
| 2 | Medium | Good | Meduim | Yes |
| 3 | Low | Average | Medium | Yes |
| 4 | Low | Poor | Large | No |
| 5 | High | Poor | Small | No |
| 6 | Medium | Good | Medium | No |

Test row:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Safety | Value For Money | Boot Size | Buy |
|  | Low | Poor | Small | ? |

1. Given the six rows of training data above, explain how ***k***-Nearest Neighbour would classify the row of test data shown above if ***k*** is set to 3. Include all calculations in your answer.

If the actual class label for this row is ‘**no’**, does ***k***-Nearest Neighbour classify it correctly at k=3?

**(12 marks)**

1. Interpret the following confusion matrix from a ***k***-Nearest Neighbour classifier, trained on 40 rows of the car dataset:

|  |  |  |
| --- | --- | --- |
|  | **Predicted Yes** | **Predicted No** |
| **Actual Yes** | 20 | 5 |
| **Actual No** | 10 | 5 |

* 1. What is the overall **accuracy** of the classifier? (**2 marks)**
  2. Calculate the **precision** for each class. Which class has the best precision? (**5 marks)**
  3. Calculate the **recall** for each class. Which class has the best recall? (**5 marks)**

1. Compare k-Nearest Neighbour with one other classification algorithm you have studied in terms of: input and output data types supported; how easy the output is to understand; how well the algorithm can handle missing data; and training time.

(**6 marks)**

***Total: 30 marks***

**2013 Repeat Paper**

The dataset below is based on an online auction site such as e-Bay. The two attributes represent the duration of the auction, and whether or not the starting price was high. The class label, good price, determines if the item was sold for a good (high) price or not.

|  |  |  |  |
| --- | --- | --- | --- |
| **Start Price** | **Length of Auction** | **Good price** | Note:  Entropy(2,1)=0.92  Entropy(2,3)=0.97 |
| High | Short | yes |
| High | Long | yes |
| Low | Long | no |
| Low | Short | no |
| High | Short | no |

1. Explain how an impurity measure such as **entropy** can be used to decide which attribute to select for each node on a decision tree. Use the data given above to illustrate your answer by calculating the entropy for **Start Price** and **Length of Auction**. Based on your calculations, which attribute should be at the root of the tree?

**(14 marks)**

1. Explain what is meant by **pre-pruning** a decision tree. If mining a dataset that is known to be noisy, would you recommend generating a full decision tree or a pruned decision tree? Explain your answer.

**(6 marks)**

1. Interpret the following confusion matrix from a decision tree, trained on 50 rows of the online auction dataset:

|  |  |  |
| --- | --- | --- |
|  | **Predicted Yes** | **Predicted No** |
| **Actual Yes** | 10 | 20 |
| **Actual No** | 0 | 20 |

* 1. What is the overall **accuracy** of the classifier? (**2 marks)**
  2. Calculate the **precision** for each class. Which class has the best precision? (**4 marks)**
  3. Calculate the **recall** for each class. Which class has the best recall? (**4 marks)**

***Total:30 marks***

**2012, January paper:**

1. Explain the difference between an eager and a lazy learner. Give scenarios when each would be a suitable classifier to use.

*(6 marks)*

1. You have been asked to mine a data set with the following characteristics:

There are 2000 rows of data. The fifteen attributes are numeric only, and the class label is binary. The objective is to both understand the patterns in the data, and also build a model that will accurately predict the class label. The patterns in the dataset are not expected to change much over time.

Which classification algorithms are suited to this task? Will you need to use more than one classifier to meet the objectives above? Explain you answer.

*(6 marks)*

1. Given the seven rows of training data below to classify mushrooms, explain how ***k***-Nearest Neighbour would classify the row of test data given above if ***k*** is set to 3. Include all calculations in your answer.

If the actual class label for this row is ‘**edible’**, does ***k***-Nearest Neighbour classify it correctly at k=3?

(*14 marks)*

**Training data: Test data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row Number** | **Cap Shape** | **GillSize** | **Ring Type** | **Type** |
| 1 | Round | Narrow | Oval | poisonous |
| 2 | Round | Narrow | Ring | poisonous |
| 3 | Round | Broad | Ring | poisonous |
| 4 | Flat | Broad | Ring | poisonous |
| 5 | Flat | Narrow | Ring | edible |
| 6 | Flat | Broad | Oval | edible |
| 7 | Round | Broad | Oval | edible |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Round | Narrow | Oval | ? |

1. Discuss how to find the optimal value of ***k*** when using ***k***-Nearest Neighbour.

*(4 marks)*

1. The following confusion matrices present the classification results that were produced using a Decision Tree and Neural Network in the prediction of Iris type for the Iris dataset. Which data mining method performs the best for the particular task? Discuss each evaluation measure you use to assess the performance of each model.

*(10 marks)*

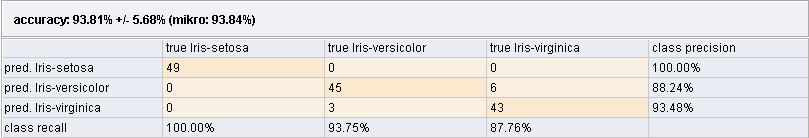


Figure 5 Decision Tree Classification

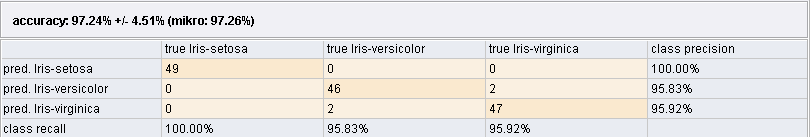


Figure 6 Neural Network Classification

**2012, repeat paper:**

1. Given the classification results in the following confusion matrix, complete the classification *accuracy*, *precision* and *recall* scores.

*(9 marks)*

|  |  |  |
| --- | --- | --- |
| Classified As: | | Correct |
| Positive | Negative |
| 25 | 5 | Positive |
| 2 | 100 | Negative |

1. Why would you use cross validation when training a classification model? In your answer, include how cross validation works, and the purpose of both a training and a test dataset.

*(10 marks)*

1. Explain how an impurity measure such as **entropy** can be used to decide which attribute to select for each node on a decision tree. Use the data given below to illustrate your answer by calculating the entropy for **CapShape** and **GillSize**, and use that to determine which should be used as the root node of a tree to classifying mushrooms as edible or poisonous.



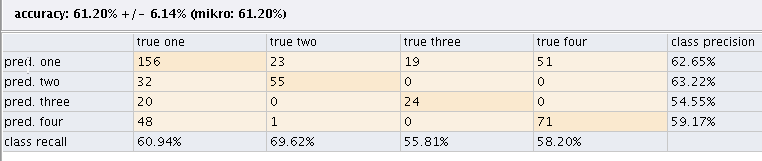
|  |  |  |  |
| --- | --- | --- | --- |
| **CapShape** | **GillSize** | **Edible?** | Note:  Entropy(3,2)=0.97 |
| Round | Narrow | yes |
| Round | Narrow | no |
| Round | Broad | no |
| Flat | Broad | yes |
| Flat | Narrow | yes |
| Flat | Broad | yes |
| Round | Broad | no |
| Round | Narrow | no |

*(14 marks)*

1. Explain what is meant by **pre-pruning** a decision tree. If mining a dataset that is known to be noisy, would you recommend generating a full decision tree or a pruned decision tree? Explain your answer.

*(6 marks)*

**2009, summer paper:**

****

1. Interpret the confusion matrix given above. The class label has four possible values, namely ‘one’, ‘two’, ‘three’ and ‘four’. In your answer explain the individual cell entries, the class **precision** entries and the class **recall** entries. Also explain how these figures could be used to estimate the **cost of the c**lassifier.

**8 marks**

1. Explain how an **impurity measure** can be used to decide on spilt points in a decision tree. **4 marks**
2. Using the data given below, calculate an impurity measure for **student** and **gender**. **8 marks**
3. Based on your results in part (ii) above, advise which attribute should be used as the next split point for the data given.

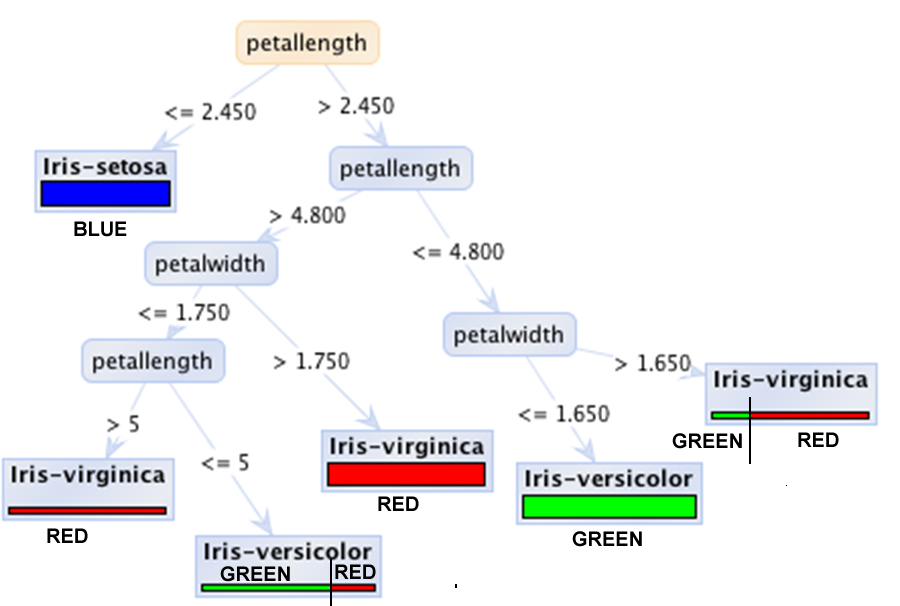
**2 marks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student** | **Gender** | **Label** |  |
| yes | M | high |
| yes | F | low |
| yes | M | low |
| no | M | high |
| no | F | high |
| no | F | high |
| yes | M | low |

1. Explain what is meant by model **over-fitting**. Your answer should discuss causes of model over fitting and how the problem can be addressed when building a **decision tree**.

**8 marks**

**2009, repeat paper:**



1. Answer the following questions based on the graphical representation of a decision tree given above:

* What is represented by leaf nodes?
* What is represented by non-leaf nodes?
* What is represented by branches?
* How does the graphical representation illustrate the number of examples incorrectly predicted? Explain your answer.

**5 marks**

1. For each of the following, state how the row of data would be classified by the decision tree above, and how confident you are in the prediction.

|  |  |
| --- | --- |
| **Petal Length** | **Petal Width** |
| 5.2 | 1.7 |
| 4.7 | 1.8 |

**2 marks**

1. Which is the most predictive attribute in the tree above? In your answer, give an overview of how an **impurity measure** selects the most predictive attribute. Illustrate your answer with an example.

**5 marks**

1. Discuss why one might consider **pruning** a decision tree, and what is the benefit of pruning.

**5 marks**

1. Explain how a **nearest neighbour** classifier works.

**5 marks**