Example Midterm Exam Data Mining, IBI008

- Write your name and student number on the answer sheet.
- The exam is multiple choice.
- All questions have four possible answers, marked by the letters **A**, **B**, **C**, and **D**.
- NB: in some questions you are asked to identify the *correct* statement and in others the *incorrect* statement.
- All questions are weighted equally.
- Good luck!

n	1	2	3	4	5	6	7	8
x_n	-3	-5	-8	0	75	-4	-1	-6

Table 1: A simple data set with eight data objects each with one attribute.

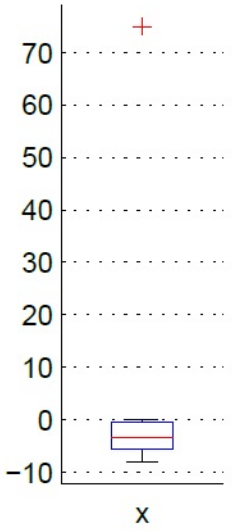


Figure 1: Boxplot corresponding to the data in Table 1.

1. Consider the data in Table 1 illustrated by the boxplot in Figure 1. Which statement is *incorrect*?

   A. The plus sign indicates an outlier.

   B. The line in the middle of the box indicates the median.

   C. The height of the box indicates the standard deviation.

   D. The upper and lower whiskers indicate the most extreme data that are not outliers.

Solution: The upper and lower edges of the box indicate the 25th and 75th percentiles, also called 1st and 3rd quartiles. Thus, the total height of the box is the inter quartile range and not the standard deviation.

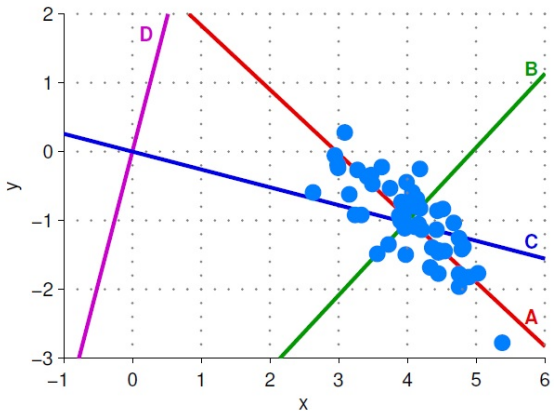


Figure 2: A scatter plot.

2. Consider the data illustrated in the scatter plot in Figure 2. Each data object has two attributes, x and y. Which axes correspond to the two principal components?

   A. A and B.

   B. C and D.

   C. C and B.

   D. A and C.

Solution: In PCA, the mean of the data is removed. The first PCA dimension is chosen to capture as much of the variability as possible. The second dimension is orthogonal to the first, and, subject to that constraint, captures as much of the remaining variability as possible, and so on. Thus, the first principal axis is A and the second, which must be orthogonal, is B.

3. Consider the data illustrated in the scatter plot in Figure 2. Which statement is *correct*?

   A. The singular value corresponding to the first principal component is larger than the singular value corresponding to the second principal component.

   B. The singular value corresponding to the first principal component is the same as the singular value corresponding to the second principal component.

   C. The singular value corresponding to the first principal component is smaller than the singular value corresponding to the second principal component.

   D. There is insufficient information in the plot to determine the relative magnitude of the singular values.

Solution: By definition, the singular values are in non-increasing order, which rules out C. We can judge the relative size of the singular values from the plot, since they relate to the variance in the corresponding principal component direction. It is clear from the graph, that the variance along the first principal component is greater than along the second.

4. When training a decision tree, we use the classification error as impurity measure $I(t)$ given by

$$I(t) = 1 - \max_i [p(i|t)] \,,$$

with $p(i|t)$ the fraction of data objects belonging to class i at a given node t and c is the number of classes. The gain Δ compares the classification error before a split with the weighted sum of classification errors after the split:

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(\nu_j)}{N} I(\nu_j) \,,$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(\nu_j)$ is the number of data objects associated with the child node ν_j.

We will consider classification of wine into Red and White wine. At a potential split we have:

- Before the split: 5 Red and 10 White.
- After the split:
  - 3 Red and 2 White in the left node.
  - 2 Red and 8 White in the right node.

Which statement is *correct*?

A. $\Delta = -4/15$.

B. $\Delta = 1/15$**.**

C. $\Delta = 1/3$.

D. $\Delta = 2/3$.

Solution: We have

$$I(\text{parent}) = 1 - \frac{10}{15} = \frac{1}{3} \,, \quad I(\text{left}) = 1 - \frac{3}{5} = \frac{2}{5} \,, \text{ and } I(\text{right}) = 1 - \frac{8}{10} = \frac{1}{5} \,,$$

yielding

$$\Delta = \frac{1}{3} - \frac{5}{15}\frac{2}{5} - \frac{10}{15}\frac{1}{5} = \frac{1}{15} \,.$$

5. In the analysis of car prices the following attributes were collected for each car: the number of kilometers the car had driven (denoted *Mileage*), the make (brand) of the car (denoted *Make*), and the year the car was produced (denoted *Year*). Which statement about the three attributes is *correct*?

   A. *Mileage* is ratio, the *Make* of car is interval and the *Year* is ratio.

   B. *Mileage* is interval, the *Make* of car is ordinal and the *Year* is nominal.

   C. *Mileage* is ratio, the *Make* of car is nominal and the *Year* is interval.

   D. *Mileage* is interval, the *Make* of car is ratio and the *Year* is ordinal.

Solution: *Mileage* is ratio as zero means absence of what is measured, i.e. no mileage on the car and we can talk about a car having driven twice as far as another car. *Make* of car is nominal as this is a categorical variable however we can not in general talk about one make of car being less than or equal to another make of car. *Year* is interval as distance between the year attribute can be measured.

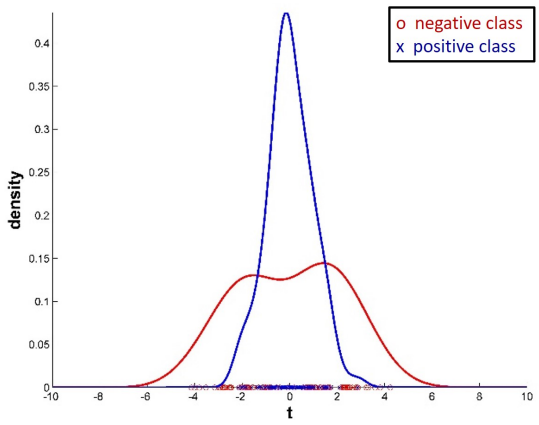
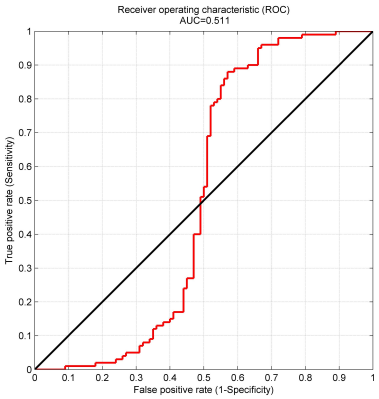
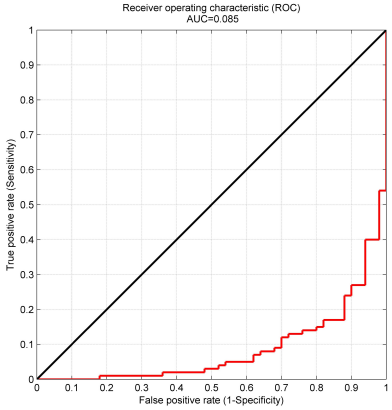


Figure 3: A classifier has given the score t to observations belonging to the two classes "negative" (red) and "positive" (blue).

ROC Curve A



ROC Curve B



ROC Curve C

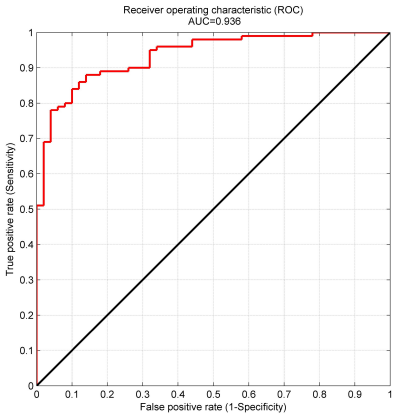


Figure 4: Three Receiver Operator Characteristic (ROC) curves.

6. We will consider a two-class classification problem. A classifier has been trained that gives a score t to each observation. Figure 3 shows the value of the score for each observation, as well as the density of the two classes as a function of the score t estimated using a kernel density estimator (a sort of smoothed histogram). The higher the score t, the more likely the classifier considers the class to be positive.

   Figure 4 shows the ROC curves for three different classifiers. ROC curves plot the false positive rate (number of false positives divided by the total number of positive examples) against the true positive rate (number of true positives divided by the total number of positive examples).

Which of the three ROC curves in Figure 4 corresponds to the classifier that scores the observations according to Figure 3?

A. ROC curve A corresponds to the classifier.

B. ROC curve B corresponds to the classifier.

C. ROC curve C corresponds to the classifier.

D. It is not possible to generate a ROC curve for the classifier since this is a non-linear classification problem.

Solution: Everything above a threshold is classified as being positive, everything below as negative. For an extremely high threshold, all observations are thus classified as negative, corresponding to a TP rate of 0 and an FP rate of 0, i.e., the origin in the ROC plot. Lowering the threshold, we will first get many false positives: incorrectly predicting the red, negative samples with a high value for t to be positive. The TP rate will stay more or less the same. This rules out ROC curve C. When the threshold is lowered to 2 and further, it will start to correctly classify the blue, positive observations, leading to a sharp increase in the TP rate for relatively small increases in the FP rate. Lowering the threshold even further, below -2, there are hardly any blue (positive) observations left, so the TP rate will stay more or less constant, whereas the FP rate will keep on increasing. This is precisely the behavior seen in ROC curve A.

7. Which of the following statements about decision trees is *correct*?

   A. Post-pruning helps to prevent overfitting, whereas pre-pruning does not.

   B. Information gain ratio (GainRATIO) makes the decision tree favor splits of attributes with a large number of distinct values.

   C. Decision trees can be constructed for data with many attributes.

   D. Decision trees are particularly suited for continuous attributes.

Solution: Both post-pruning and pre-pruning help to prevent overfitting, even though post-pruning typically works better. Information gain ratio biases *against* attributes with a large number of distinct values. Decision trees are particularly suitable for discrete attributes. When decision trees are used with continuous attributes, these must be converted to discrete variables, e.g., by dividing them into discrete intervals. Indeed, decision trees are perfectly capable to handle data with many attributes.

8. Let

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} .$$

We would like to estimate the similarity between $\mathbf{x}$ and $\mathbf{y}$. We will do this in a robust manner and we will therefore estimate the similarity as the average of the Simple Matching Coefficient,

$$\text{SMC}(\mathbf{x}, \mathbf{y}) = \frac{\text{Number of matching attribute values}}{\text{Number of attributes}} ,$$

the Jaccard similarity,

$$J(\mathbf{x}, \mathbf{y}) = \frac{\text{Number of matching presences}}{\text{Number of attributes not involved in 00 matches}}$$

and the cosine similarity

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} ,$$

i.e.,

$$t = \frac{1}{3} \left[\text{SMC}(\mathbf{x}, \mathbf{y}) + J(\mathbf{x}, \mathbf{y}) + \cos(\mathbf{x}, \mathbf{y}) \right] .$$

What is the value of t?

A. $t = \frac{9}{24}$.
B. $t = \frac{19}{45}$.
C. $t = \frac{37}{60}$.
D. $t = \frac{121}{180}$**.**

Solution: We get

$$t = \frac{1}{3} \left[\text{SMC}(\mathbf{x}, \mathbf{y}) + J(\mathbf{x}, \mathbf{y}) + \cos(\mathbf{x}, \mathbf{y}) \right] = \frac{1}{3} \left[\frac{4}{6} + \frac{3}{6-1} + \frac{3}{2 \cdot 2} \right] = \frac{121}{180} .$$

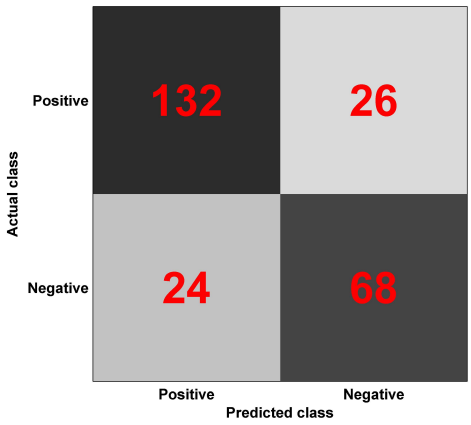


Figure 5: A confusion matrix.

9. A classifier takes as attributes some measurements derived from blood samples and classifies based on these features whether the tested subjects have a virus infection (test is positive) or not (test is negative). The confusion matrix of the classifier is given in Figure 5. The precision specifies the fraction of true positives among the instances that the classifier predicts to be positive. What are the precision and the error rate of the classifier?

   A. precision $= \frac{33}{50}$, error rate $= \frac{4}{5}$.

   B. precision $= \frac{33}{50}$, error rate $= \frac{1}{5}$.

   C. precision $= \frac{11}{13}$, error rate $= \frac{1}{5}$.

   D. precision $= \frac{11}{13}$, error rate $= \frac{4}{5}$.

Solution: We get

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{132}{132 + 24} = \frac{11}{13} ,$$

and

$$\text{error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{50}{250} = \frac{1}{5} .$$

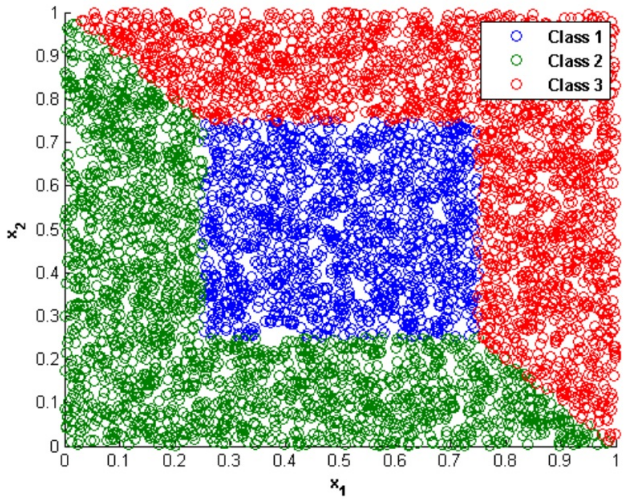


Figure 6: A classification problem.

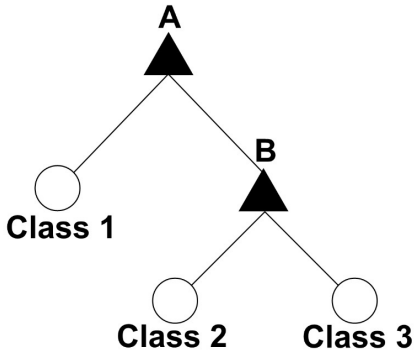


Figure 7: Decision tree corresponding to the data in Figure 6.

10. Consider the classification problem given in Figure 6 and the decision tree in Figure 7 with two decisions denoted **A** and **B**. The convention is that "True" goes left and "False" goes right. Objects are two-dimensional vectors, written $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Let

$$L_p(\mathbf{x}) = \|\mathbf{x}\|_p = \left(\sum_{k=1}^{2} |x_k|^p \right)^{1/p}$$

denote the so-called L_p-norm. Which one of the following classification rules would lead to a correct classification of the data?

A. **A** : $\left\|\mathbf{x} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right\|_1 \leq 0.25$
B : $\|\mathbf{x}\|_\infty \leq 1$

B. **A** : $\|\mathbf{x}\|_1 \leq 1$
B : $\left\|\mathbf{x} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right\|_1 \leq \infty$

C. **A** : $\left\|\mathbf{x} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right\|_2 \leq 0.25$
B : $\|\mathbf{x}\|_\infty \leq 1$

D. **A** : $\left\|\mathbf{x} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right\|_\infty \leq 0.25$
B : $\|\mathbf{x}\|_1 \leq 1$

Solution: The decision **A**, $\left\|\mathbf{x} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right\|_\infty \leq 0.25$, will classify the data in the inner rectangular region to class 1. **B**, $\|\mathbf{x}\|_1 \leq 1$, implies $|x_1| + |x_2| \leq 1$, which will split class 2 from class 3.

11. We would like to determine how well a decision tree can distinguish persons with a virus infection from persons who do not have a virus infection based on properties of their blood. We would like to estimate the generalization performance, i.e., how well the decision tree will do on persons never seen before. To prevent overfitting, we consider trees with different depths. Since we have quite some data available for training and testing, we decide to use a hold-out procedure, leaving independent data aside for measuring performance in the so-called test set.

    Which of the following statements is *correct*?

    A. We can optimize the depth of the tree based on the training error and then report the performance of the tree with the optimal depth on the test set.

    B. We can train trees with many different depths, and then take the lowest error on the test set obtained with any depth as our estimate of the generalization performance.

    C. Using a single test set is fine, as long as both training and test set contain at least 10 observations.

    D. None of the above.

Solution: Optimizing the depth of a tree based on training error makes no sense: it will always take the largest depth (A incorrect). Picking the best depth based on the test set and then reporting generalization leads to overfitting of the hyper-parameter, here the depth of the tree, on the test set. The correct approach is, for example, to use inner cross-validation to find the optimal depth and then report the error on the test set (B incorrect). You need many more observations in both training and test set for the single hold out procedure to yield reliable results (C incorrect). So D is correct.

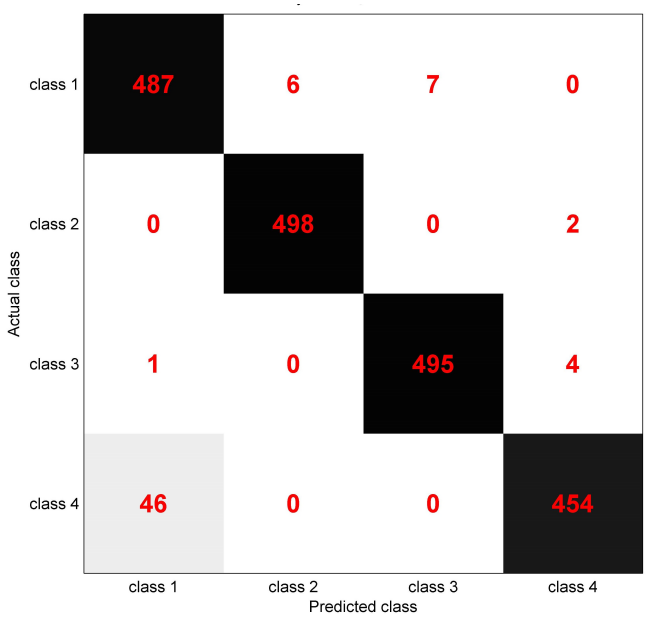


Figure 8: A confusion matrix.

12. Figure 8 gives the confusion matrix for a particular classifier. Which of the following statements is *incorrect*?

    A. The error rate of the classifier is 66/2000.

    B. The classifier is performing significantly better than random guessing.

    C. The main confusion of the classifier is given by observations in class 4 being classified as class 1.

    D. The classification problem has class-imbalance issues that should be addressed.

Solution: The classification problem does not have class-imbalance issues as there is the same number of observations in each class, i.e., 500.

13. A singular value decomposition is carried out on a dataset comprised of three data points $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ collected in an $N \times M$ matrix $\mathbf{X}$ such that each row of the matrix is a data point. Suppose the matrix $\tilde{\mathbf{X}}$ corresponds to $\mathbf{X}$ with the mean of each column subtracted, i.e.,

$$\mathbf{X} = \begin{pmatrix} 3.00 & 2.00 & 1.00 \\ 4.00 & 1.00 & 2.00 \\ 0.00 & 1.00 & 2.00 \end{pmatrix} \quad \text{with} \quad \tilde{X}_{ij} = X_{ij} - \frac{1}{N}\sum_{k=1}^{N} X_{kj} \, .$$

$\tilde{\mathbf{X}}$ has the singular value decomposition $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, with

$$\mathbf{U} = \begin{pmatrix} -0.26 & 0.77 & 0.58 \\ -0.54 & -0.61 & 0.58 \\ 0.80 & -0.16 & 0.58 \end{pmatrix} \, , \quad \mathbf{\Sigma} = \begin{pmatrix} 2.96 & 0.00 & 0.00 \\ 0.00 & 1.10 & 0.00 \\ 0.00 & 0.00 & 0.00 \end{pmatrix} \, ,$$

$$\text{and} \quad \mathbf{V} = \begin{pmatrix} -0.99 & -0.13 & -0.00 \\ -0.09 & 0.70 & -0.71 \\ 0.09 & -0.70 & -0.71 \end{pmatrix} \, .$$

Here $\mathbf{V}$ contains in its columns the so-called right singular vectors, which correspond to the principal directions. What are the (rounded to two significant digits) coordinates of the (mean-centered) first observation $\tilde{\mathbf{x}}_1$ projected onto the two-dimensional subspace containing the maximal variation?

A. (-3.06 0.31).

B. (-0.78 0.85).

C. (-1.07 0.21).

D. (-3.16 0.23).

Solution: The projection can be found by subtracting the mean from $\mathbf{X}$ and projecting onto the first two columns of $\mathbf{V}$ (see, e.g., appendix B of TSK). The first point with the mean subtracted has coordinates $(3-7/3, 2-4/3, 1-5/3)$. To compute the first component, we should take the inner product with (the transpose of) the first column of $\mathbf{V}$ or, equivalently, right multiply with the first column of $\mathbf{V}$, and the same for the second column. So, doing this in one go, we get:

$$\begin{pmatrix} 3-7/3 \\ 2-4/3 \\ 1-5/3 \end{pmatrix}^T \begin{pmatrix} -0.99 & -0.13 \\ -0.09 & 0.7 \\ 0.09 & 0.7 \end{pmatrix} = (-0.78 \;\; 0.85) \, ,$$

corresponding to option B.