Midterm

DSA1101
Introduction to Data Science

Semester 1, 2018

Name:

Matriculation card number:

Problem	**Score**
1	
2	
3	
4	
Total:	

Problem 1 (35 points). Suppose we have two data vectors $x = c(x_1, x_2, x_3) = c(1, 2, 4)$ and $y = c(y_1, y_2, y_3) = (0, 8, 10)$.

(a) (5 points) Assume the following linear relationship between y and x:

$$y \approx \beta_0 + \beta_1 x.$$

Complete the following table based on the 3 data points, leaving your answers in terms of β_0 and β_1:

i	x_i	y_i	$\beta_0 + \beta_1 x_i$	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$
1				
2				
3				

(b) (5 points) Write down an expression for the Residual Sum of Squares (RSS) for the model in (a), leaving your answer in terms of β_0 and β_1:

(c) (15 points) Based on the RSS given in (b), derive and write down the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) (5 points) Instead of the model in (a), suppose we postulate the following relationship between y and x:

$$y \approx \alpha_0,$$

i.e. the model only involves the intercept parameter α_0. Write down an expression for the Residual Sum of Squares of this model in terms of α_0 and derive the least squares estimate $\hat{\alpha}_0$.

(e) (5 points) Hence, calculate the R^2 statistic for the model in (a).

Problem 2 (15 points). *The k-means clustering algorithm*
Suppose we have data for four objects on two features:

object	x_1	x_2
A	2	3
B	6	3
C	2	5
D	6	5

We set $k = 2$ to cluster the four data points into two clusters, $\mathcal{P}$ and $\mathcal{Q}$

(a) (5 points) Based on the centroids $(x_{1,\mathcal{P}}, x_{2,\mathcal{P}}) = (2, 4)$ and $(x_{1,\mathcal{Q}}, x_{2,\mathcal{Q}}) = (6, 4)$, assign the four points into the two clusters. Compute the *Within Sum of Squares* (WSS) for this clustering assignment.

(b) (5 points) Based on the centroids $(x_{1,\mathcal{P}}, x_{2,\mathcal{P}}) = (4, 3)$ and $(x_{1,\mathcal{Q}}, x_{2,\mathcal{Q}}) = (4, 5)$, assign the four points into the two clusters. Compute the *Within Sum of Squares* (WSS) for this clustering assignment.

(c) (5 points) Which of the clustering assignments in (a) and (b) is better? Justify your answer in terms of WSS.

Problem 3 (25 points). Suppose we have two data vectors $x = c(x_1, x_2, ..., x_n)$ and $y = c(y_1, y_2, ..., y_n)$, both of length n. In addition, a and b are two positive constants. Consider the two new data vectors $ax = c(ax_1, ax_2, ..., ax_n)$ and $by = c(by_1, by_2, ..., by_n)$

(a) (5 points) Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ be the sample mean of x. Show that $\overline{ax} = a\bar{x}$.

(b) (5 points) Let $\text{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ be the sample variance of x. Show that $\text{var}(x) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$.

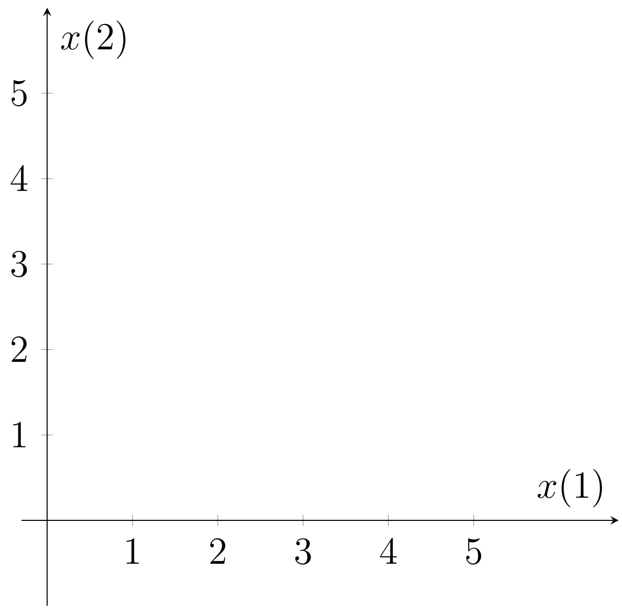
(c) (5 points) Let $\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ be the sample covariance between x and y. Show that $\text{cov}(x, y) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right)$.

(d) (5 points) Using the results from (b) and (c), show that $\text{cov}(ax, by) = ab \times \text{cov}(x, y)$ and $\text{var}(ax) = a^2 \times \text{var}(x)$.

(e) (5 points) Hence, show that $r_{ax,by} = r_{x,y}$, where $r_{x,y}$ is the sample correlation coefficient between x and y.

Problem 4 (25 points). Suppose we have a training set of six data points in two features $x(1) = c(1, 2, 1, 3, 3, 2)$ and $x(2) = c(3, 2, 1, 3, 1, 0)$, as well as the corresponding binary label values $y = c(0, 0, 1, 1, 1, 0)$

(a) (5 points) Plot the six training data points in the 2-dimensional feature space below, using the symbols $\bullet$ for points with $y = 1$ and $\circ$ for points with $y = 0$.



(b) (5 points) Plot the following test data points on the same graph in (a) using the symbol $\times$. Based on the graph, predict the label values of y for the test objects A,B and C using the 3-nearest neighbors classifier with the majority rule.

test object	$x(1)$	$x(2)$
A	1	2
B	3	2
C	2	0.5

(c) (10 points) Suppose the actual label values for the test objects are as follows: $y_A = 0$, $y_B = 1$ and $y_C = 0$. Compute the *accuracy*, *true positive rate*, *false positive rate*, *false negative rate* and *precision* of the classifier based on the actual and predicted values of y for the three test objects. The definitions are as follows:

		Predicted y	
		1	0
Actual y	1	True Positives	False Negatives
	0	False Positives	True Negatives

(d) (5 points) Instead of the majority rule, suppose the classifier predict the label value to be 1 when the fitted outcome value $\hat{y}(x) > 0.8$. If the *false positive rate* changes, state whether you expect it to increase or decrease. Justify your answer.