ECE 592 – Topics in Data Science

Final – Fall 2017

December 11, 2017

Please remember to justify your answers carefully, and to staple your test sheet and answers together before submitting.

Name: _____ Student ID: _____

Question 1 (Warm-up questions.)
Below are a few quick questions that you should be able to answer quickly.

(a) Consider a graph $G(V, E)$, where V are the vertices and E are edges. Given that G is a forest comprised of k trees, how many edges does it have? (Hint: your answer should relate $|E|$ and $|V|$, the cardinalities of the sets E and V.)

(b) Describe an application where hardware can acquire linear measurements in a compressive way.

(c) Describe a type of signal for which linear approximation is much worse than nonlinear approximation. Make sure to justify your answer.

Question 2 (Bayes classification.)
You are given training vectors labeled into two classes, where each vector is in $p = 2$ dimensions. The training set is comprised of the following points; these will be plotted below.

- Class 1: $\{(11, 11), (13, 11), (8, 10), (9, 9), (7, 7), (7, 5), (15, 3)\}$
- Class 2: $\{(7, 11), (15, 9), (15, 7), (13, 5), (14, 4), (9, 3), (11, 3)\}$

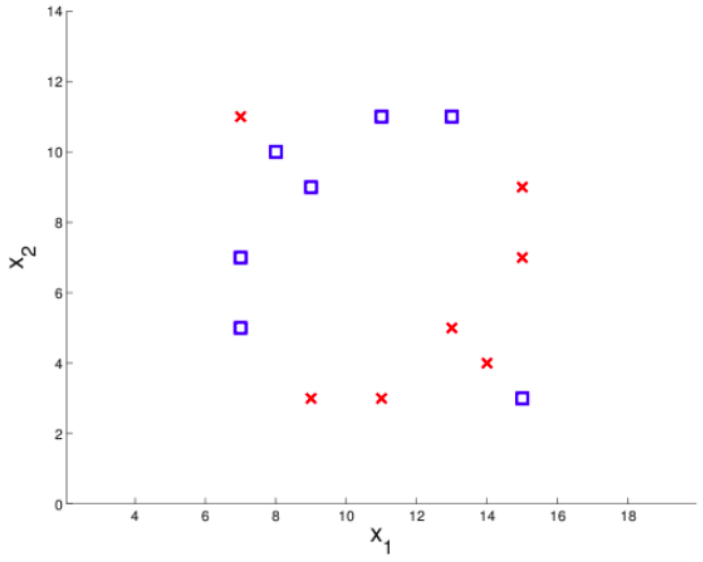
These vectors follow Gaussian distributions, where Class 1 has the following mean vector and covariance matrix,

$$\mu_1 = \begin{bmatrix} 10 \\ 8 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix},$$

and Class 2 has the following mean vector and covariance matrix,

$$\mu_2 = \begin{bmatrix} 12 \\ 6 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}.$$

For ease of visualization, we include a scatter plot of the feature vectors, with Classes 1 and 2 denoted by crosses and squares, respectively.



(a) Are the two classes linearly separable? Please explain.

(b) In the scatter plot, indicate the decision boundary that you would get using a Bayesian classifier; note that both classes use the same diagonal covariance matrix.

(c) In the same scatter plot, sketch the decision boundary that you would get using a nearest neighbor (NN) classifier using $k = 1$.

(d) Classify test samples (6,11) and (14,3) using the Gaussian Bayes classifier. (Hint: you need not go into detailed derivations if you can explain the output of the classifier.)

Question 3 (Dynamic programming (DP).)
A criminal approaches you and wants your advice about producing fake coins. The criminal can produce three types of coins. The values of these coins are $v_1 = 1$, $v_2 = 4$, and $v_3 = 11$, and the number of units of metal needed to produce each such coin is $w_1 = 2$, $w_2 = 5$, and $w_3 = 13$, respectively. The criminal wants to produce N value in coins using as little metal as possible. Use dynamic programming to compute $\Psi(N)$, the minimal number of units of metal needed to produce N value in coins.

Question 4 (Principal component analysis (PCA).)
In a sample of 36 flea beetles, 6 different physical characteristics are measured for each beetle: head length, body length, length of one joint, length of a second joint, total weight, and body temperature. The first four are measured in micrometers, the fifth is measured in milligrams, and the sixth is measured in degrees Celsius. The outcome of a principal component analysis (PCA) and a related plot are given in two figures below. Based on these results, answer the following questions:

(a) The correlation matrix is used here instead of the covariance matrix. Explain why the correlation matrix is a more sensible choice than the covariance matrix for this analysis.

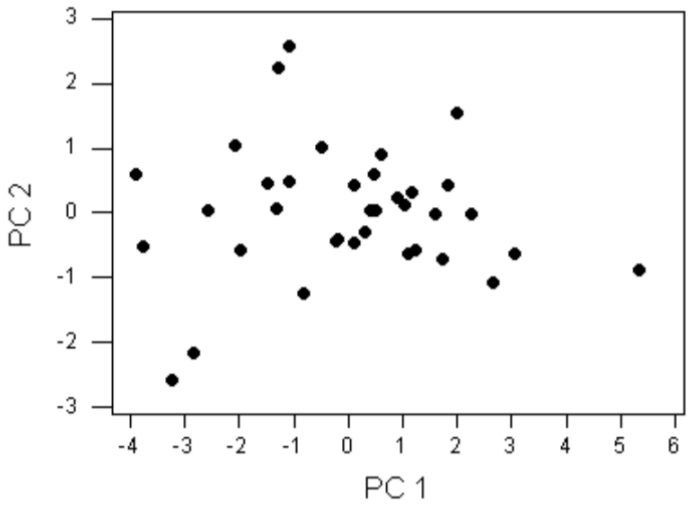
(b) How much of the variation in this dataset is explained by the first principal component? How much is explained by the first and the second components together?

Eigenanalysis of the Correlation Matrix

Eigenvalue	4.0424	1.0419	0.8711	0.0372	0.0064	0.0010
Proportion	0.674	0.174	0.145	0.006	0.001	0.000
Cumulative	0.674	0.847	0.993	0.999	1.000	1.000

Variable	PC1	PC2	PC3
head	0.489	-0.001	-0.077
body	0.495	0.065	0.028
jnt1	0.437	-0.084	0.393
jnt2	0.372	0.083	-0.723
weight	0.496	-0.057	0.023
temp	0.035	0.824	0.562

(c) A plot of the principal component scores for the 36 beetles is shown in the figure below. Imagine that the horizontal and vertical axes are drawn on the plot, dividing it into 4 quadrants: UR (upper right), UL (upper left), LR (lower right), and LL (lower left). Suppose that two new beetles, Big Bob and Tiny Tim, are measured. Bob is much larger than any of the other beetles and is also slightly warmer. Tim, on the other hand, is very small compared with the other beetles but like Bob, Tim is warmer than the others. For both Bob and Tim, which quadrant of the above graph would each lie in? How do you know?



Question 5 (Individual projects.)
From the individual project presentations held in class, select any two (excluding your own). For each one, briefly describe:

(i) The motivation for the project, i.e., what problem was solved and why.

(ii) Which data science techniques were used in the project.