Homework 13

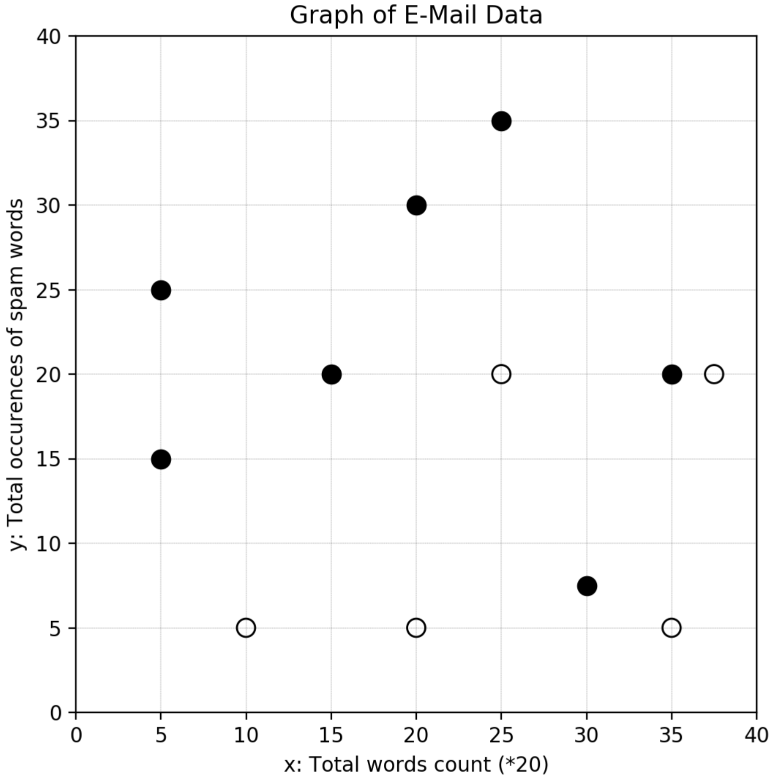
Due: Thursday, May 2, 2024 at 12:00pm **(Noon)**

Written Assignment

Problem 1: PCA & KNN (20 points)

In this question, you will build a spam filter using the K-Nearest Neighbour algorithm like the one you built on Homework 11. This time, you will be using Principal Component Analysis (PCA) to transform the e-mail data from 2 dimensions (and a label) to 1 dimension (and a label). Recall that each e-mail has two features: number of words in the email and total occurrences of the spam words "credit" and "dollars". The following graph shows the training data where filled circles are spam e-mails and unfilled circles are non-spam emails. The x-axis (Total words count) has been scaled for your convenience.

We have found the first principal component: $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Therefore the linear subspace that PCA will project the data to is the line $y = x$. (Assume that we don't need to do any centering of the data.)



a. Draw the first principal component as a line ($y = x$ in this case) on the graph. Also draw the data points projected to this subspace with lines connecting them to the corresponding original data points.

b. How would **1**-Nearest Neighbour algorithm classify an e-mail with 21 spam words in a total of 100 (x=5) words? Explain your answer.

c. How would **1**-Nearest Neighbour algorithm classify the same e-mail given in the previous question after transforming the data to 1-dimension? Explain your answer.

d. Are your answers the same for part b and c? Why or why not?

Problem 2: Exact Recovery of a Linear Compression Scheme (15 points)

In this exercise we show that in the general case, exact recovery of a linear compression scheme is impossible.

a. Let $A \in \mathbb{R}^{n,d}$ be an arbitrary compression matrix where $n \leq d-1$. Show that there exists $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \mathbf{u} \neq \mathbf{v}$, such that $A\mathbf{u} = A\mathbf{v}$.

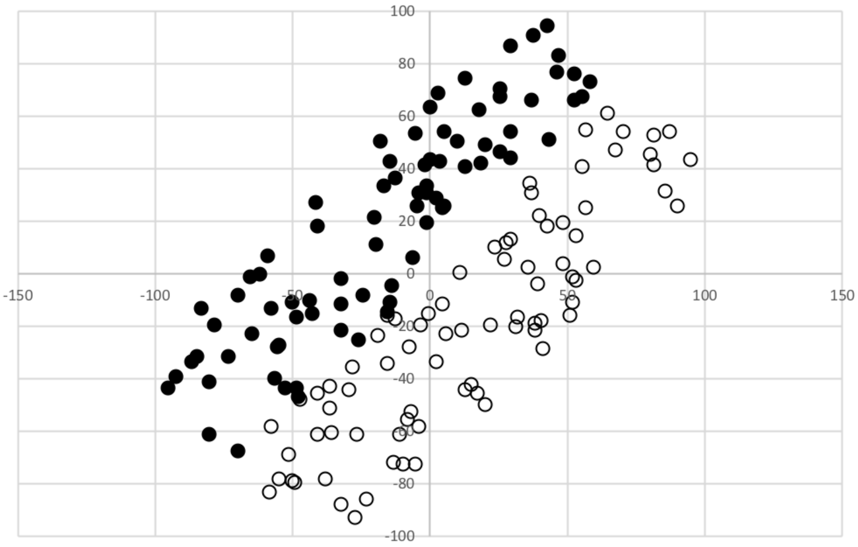
   Hint: Show that there exists $u \neq 0, v = 0$ such that $Au = Av = 0$.

   Hint: Consider using the rank-nullity theorem.

b. Conclude that exact recovery of a linear compression scheme is impossible.

Problem 3: Limitations of PCA (25 points)

Consider the following two-dimensional dataset:



a. Describe one type of machine learning model that would classify this data well. Explain your reasoning.

b. Draw the first principal component on the graph above. (Remember that PCA does not consider the labels as one of the input dimensions.) If PCA was used to reduce this data to one dimension, would the machine learning model from part (a) still classify the data well? Why or why not?

c. Is it possible to project the above data into a one-dimensional linear subspace in which the data remains linearly separable? If so, draw the subspace on the graph above. If not, explain your reasoning.

d. What does this example tell you about the limitations of PCA when used to pre-process data before classification?

Grading Breakdown

The grading breakdown for the assignment is as follows:

Problem 1	33%
Problem 2	25%
Problem 3	42%
Total	100%

Handing In

You will hand in the assignment on Gradescope, uploading it under "Homework 13." For questions, please consult the download/submission guide.

Anonymous Grading

You need to be graded anonymously, so do not write your name anywhere on your handin.

Obligatory Note on Academic Integrity

Plagiarism — don't do it.

As outlined in the Brown Academic Code, attempting to pass off another's work as your own can result in failing the assignment, failing this course, or even dismissal or expulsion from Brown. More than that, you will be missing out on the goal of your education, which is the cultivation of your own mind, thoughts, and abilities. Please review this course's collaboration policy and if you have any questions, please contact a member of the course staff.