Problem 1. (K-means)

1. Problem 4.2 from Chapter 4 of LinAlgebra:

   k-means with nonnegative, proportions, or Boolean vectors. Suppose that the vectors $\{x^i\}_{i=1}^N \in \mathbb{R}^d$ are clustered using k-means, with group representatives $\{z^j\}_{j=1}^k \in \mathbb{R}^d$. Recall the definition of representative z^j as the average of the vectors that belong to cluster j

   $$z_j = \frac{1}{N^j}\sum_{n\in j} x^n,$$

   where N^j is the number of vectors that make up the cluster with index j, and n are the indices of the vectors $\{x^n\}$ belonging to cluster j.

   (a) Suppose that the original vectors $\{x^i\}$ are nonnegative, *i.e.*, their entries $\{x_l^i\}_{l=1}^d \geq 0$. Explain why the representatives $\{z^j\}$ are also nonnegative.

   (b) Suppose that the original vectors $\{x^i\}$ represent proportions, *i.e.*, their entries are nonnegative and sum to one. (This is the case when x^i are word count histograms, for example.) Explain why the representatives $\{z^j\}$ also represent proportions, *i.e.*, their entries are nonnegative and sum to one.

   (c) Suppose the original vectors $\{x^i\}$ are Boolean, *i.e.*, their entries are either 0 or 1. Give an interpretation of z_l^j, the lth entry of the j group representative.

2. A data set $X \in \mathbb{R}^{N\times d}$ is clustered using k-means with mean points for the clusters (group representatives) $M \in \mathbb{R}^{k\times d}$. Suppose that the original data represent proportions, *i.e.*, their entries are non-zero and sum to one. Taking the ith sample $x^i \in \mathbb{R}^3$, $x_l^i \geq 0$ and $\sum_l^d x_l^i = 1$. Explain why the group representatives $\mu^j \in \mathbb{R}^d$ also represent proportions, *i.e.*, their entries are non-negative and sum to one.

3. Read the section "Guessing missing entries" from Section 4.5 of the book. Based on this, describe how you would fill missing entries in a data matrix.

4. For "Choosing k" you can read Section 4.3 of the book. What is the cost function that is being optimized.

Problem 2. (PCA)

We are making measurements of "Points obtained in the exam" and "Time spent on youtube". Let $X_r \in \mathbb{R}^{3\times 2}$ be our data matrix with 3 data entries and two features given by:

$$X_r = \begin{pmatrix} x_1^1 & x_2^1 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \end{pmatrix} = \begin{pmatrix} 30 & 1 \\ 10 & 2.5 \\ 20 & 1.5 \end{pmatrix}$$

1. Compute the covariance matrix of $X_r \in \mathbb{R}^{3\times 2}$. Is there a positive or negative correlation between "Points obtained in the exam" and "Time spent on youtube"? What is the interpretation of the diagonal elements?

2. Standardize the data matrix. Recall, for this you need to subtract the mean of each feature vector and divide by standard deviation of each feature vector. Call the resulting matrix X.

   Note: standardization and normalization terms are sometimes used interchangeably.

3. Compute the first principal component of X.

4. Using the first principal component, define the new features $A \in \mathbb{R}^3$ based on the original data matrix $X \in \mathbb{R}^{3\times 2}$. Which linear combination of the original data gives rise to these new features?

5. Reconstruct an approximation $\hat{X} \in \mathbb{R}^{3\times 2}$ to the original matrix using the first principal component. What is the Frobenius norm of the matrix $X - \hat{X}$?

6. The singular value decomposition of a matrix $X \in \mathbb{R}^{N\times d}$ is given by $X = USV^\top$, where $U \in \mathbb{R}^{N\times N}, S \in \mathbb{R}^{N\times d}, V \in \mathbb{R}^{d\times d}$ and U, V are orthogonal matrices. The singular values are the non-zero diagonal entries of S. Verify that V in this decomposition is the matrix whose columns are the eigenvectors of $X^T X$ and the singular values are the square root of the eigenvalues of $X^\top X$.

   Hint: Simply perform $X^T X$ using the SVD decomposition and use the orthogonal properties of the matrices.

Problem 3. (Decision trees)
Consider a classification problem with $x \in \mathbb{R}^2$ and $y \in \{\text{square}, \text{triangle}\}$. The training data is shown in Figure 1 below. There are N_t triangles and N_s squares in the training data, where $N_s = mN_t$ with $m \in (0, 1)$. So, for example, if there are 100 triangles, and $m = 0.1$. then there are 10 red squares and a total of 110 data points.

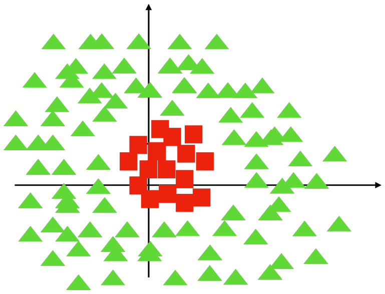

Figure 1: Classification problem training data

1. A so-called null classifier gives the majority label of the training data to any test point $x \in \mathbb{R}^2$. Hence, it considers that x has no effect on the label. Since we have $N_t = (1/m)N_s > N_s$ the majority label is triangle and the null-classifier labels any test point x as a triangle. What is the gini index of this classifier? What is the error rate of this classifier on the training data?

2. Now, consider feature 1 and the threshold at $x_1 = 1$ shown in Figure 2 below as a candidate for forming a split in a first node of a decision tree to be constructed for classification. So, the split criteria is whether $x_1 > 1$. Suppose that a fraction of $c \in (0, 1)$ number of triangles falls to the right of the line at $x_1 = 1$ shown in the figure. In other words, cN_t of triangles have $x_1 > 1$. Hence, $(1 - c)N_t$ are the number of triangles to the left of the line. Write the gini index of the two leaves and of the node according to this split.

3. Show that the gini index after the split is smaller than the gini index of the null classifier.

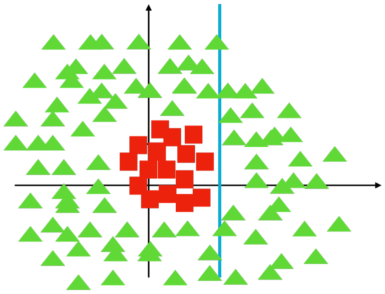

Figure 2: Classification problem with one node of the decision tree

4. Observe that anywhere you put a line, the number of “triangles” is more than the number of squares. Thus, show that no matter where you put the blue line, the accuracy of the classifier does not improve, even though its gini index can improve[1]

   Remark: note that this is the case also if you put the lines horizontally or vertically. In particular, this shows that gini index could be potentially a more useful criteria for forming the threshold than accuracy. Note that the performance of the final classifier is measured in terms of accuracy regardless of the criteria used.

5. Draw the boundaries corresponding to a decision tree that could separate the two classes.

6. What is the depth of the decision-tree that separates these two classes?

7. Your friend suggests to you to use a logistic regression for this classification problem. She thinks that it is sufficient to consider two feature as $\Phi_1(x_1, x_2) = x_1^2 + x_2^2$ and $\Phi_2(x_1, x_2) = 1$ for the logistic regression problem. How many parameters you would need to learn for the logistic regression model? What would the decision boundaries look like in this case?

[1] This is one of the motivations of using other criteria than accuracy in defining the decision-trees.