**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.
   *Solution:* In this case, the $l$-th component of $z^j$ is defined as

   $$z_l^j = \frac{1}{N^j} \sum_{n \in j} x_l^n.$$

   Because all components $x_l^n$ are nonnegative, their sum and therefore their mean is also nonnegative. It follows that all components of the representatives $z^j$ are 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.
   *Solution:* The argument for why the entries are nonnegative follows exactly as above. The sum of the components of $z_j$ can be written as

   $$\sum_{l=1}^d z^j = \sum_{l=1}^d \left[ \frac{1}{N^j} \sum_{n \in j} x_l^n \right].$$

   By re-ordering the sums, we find

   $$\sum_{l=1}^d z^j = \frac{1}{N^j} \sum_{n \in j} \sum_{l=1}^d x_l^n = \frac{1}{N^j} \sum_{n \in j} 1 = \frac{N^j}{N^j} = 1$$

   (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 $l$th entry of the $j$ group representative.
   *Solution:* From above, we know that the $l$th component of $z^j$ is

   $$z_l^j = \frac{1}{N^j} \sum_{n \in j} x_l^n$$

   In the case where components of $x_l^n$ are Boolean, $z_l^j$ is the fraction of samples in the cluster for which $x_l^n = 1$. We can interpret this as the probability that component $l$ of a member of the cluster is true, or 1.

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 $i$th sample $x^i \in \mathbb{R}^3$, $x^i_l \geq 0$ and $\sum_l^d x^i_l = 1$. 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.

   *Solution:* In this case, the $l$-th, where $l = 1, \ldots, d$, component of $\mu^j$ is defined as

   $$\mu^j_l = \frac{1}{N^j} \sum_{x^n \in j} x^n_l,$$

   where $N^j$ is the number of samples in cluster $j$.

   Since all components $x^n_l$ are nonnegative, their sum and therefore their mean is also nonnegative. It follows that all components of the representatives $\mu^j$ are nonnegative.

   The sum of the components of $\mu^j$ can be written as

   $$\sum_{l=1}^{d} \mu^j = \sum_{l=1}^{d} \left[ \frac{1}{N^j} \sum_{x^n \in j} x^n_l \right].$$

   We can re-order the sums

   $$\sum_{l=1}^{d} \mu^j = \frac{1}{N^j} \sum_{x^n \in j} \sum_{l=1}^{d} x^n_l = \frac{1}{N_j} \sum_{x^n \in j} 1 = \frac{N_j}{N_j} = 1,$$

   which shows that the entries of $\mu^j$ sum also 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.

   *Solution:*

   - Separate samples (rows) in the matrix which are complete from those that are incomplete
   - Apply the $k$-means algorithm to the "complete" subset of the data to find the representatives $z_j$
   - Find the nearest representative $z_j$ to each sample from the "incomplete" subset of the data. To compute the distance of a sample in the "incomplete" subset, use the Euclidean distance corresponding to the dimensions for which there are no missing entries.
   - Set the missing entries of each incomplete sample to the value from its corresponding representative
   - Recombine the "complete" and "incomplete" data subsets

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

   *Solution:*

   Let $z_j$ be the group representative for cluster $j$ and $G_j$ denote the indices of points belonging to cluster $j$. Then, we are considering the $K$ means cost, which is sum of the Euclidean distances of each data point to the representative of the cluster the point belongs to. Hence,

$$J = \frac{1}{N} \sum_{j=1}^{K} \sum_{i \in G_j} \|x^i - z_j\|^2$$

**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?

   *Solution:* The covariance matrix is given by

   $$C = \begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & cov(x_2, x_2) \end{pmatrix},$$

   where $cov(x_i, x_j) = \frac{1}{N-1} \sum_{n=1}^{N} (x_i^n - \mu_i)(x_j^n - \mu_j)$. In the following, we compute $\mu_1$, $\mu_2$ and entry $cov(x_1, x_1)$ of the covariance matrix. The other entries of the covariance matrix are computed analogously.

   $$\mu_1 = \frac{1}{3}(30 + 10 + 20) = 20$$
   $$\mu_2 = \frac{1}{3}(1 + 2.5 + 1.5) = \frac{5}{3}$$
   $$cov(x_1, x_1) = \frac{1}{2}((30-20)(30-20) + (20-20)(20-20) + (10-20)(10-20)) = \frac{200}{2} = 100$$

   Then,

   $$C = \begin{pmatrix} 100 & -7.5 \\ -7.5 & 0.583 \end{pmatrix}$$

   The correlation between "Points obtained in the exam" and "Time spent on youtube" is computed as

   $$cor_{x_1,x_2} = \frac{cov(x_1, x_2)}{\sqrt{cov(x_1, x_1)}\sqrt{cov(x_2, x_2)}} = -\frac{7.5}{\sqrt{100}\sqrt{0.583}} = -0.982,$$

   which is negative. This means that with an increasing time spent on youtube results in a decrease in points in the exam and vice versa. The diagonal entries correspond to variances of feature 1 ("Points obtained in the exam") and feature 2 ("Time spent on youtube"), i.e., $cov(x_1, x_1) = Var(x_1)$ and $cov(x_2, x_2) = Var(x_2)$.

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.

   *Solution:* The mean and standard deviation of feature $i = 1, 2$ is computed as $\mu_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n$ and $\sigma_i = \sqrt{\frac{\sum_{n=1}^{N}(x_i^n - \mu_i)^2}{N-1}}$. In our case, we have two features, so $i = 1, 2$, and $N = 3$ data entries.

$$\mu_1 = \frac{1}{3}(30 + 10 + 20) = 20$$
$$\sigma_1 = \sqrt{\frac{(30-20)^2 + (10-20)^2 + (20-20)^2}{2}} = \sqrt{\frac{200}{2}} = 10$$
$$\mu_2 = \frac{1}{3}(1 + 2.5 + 1.5) = \frac{5}{3}$$
$$\sigma_2 = \sqrt{\frac{(1-5/3)^2 + (2.5-5/3)^2 + (1.5-5/3)^2}{2}} = \sqrt{0.583}.$$

   The resulting standardized matrix $X$ is given by

$$X = \begin{pmatrix} \frac{x_1^1-\mu_1}{\sigma_1} & \frac{x_2^1-\mu_2}{\sigma_2} \\ \frac{x_1^2-\mu_1}{\sigma_1} & \frac{x_2^2-\mu_2}{\sigma_2} \\ \frac{x_1^3-\mu_1}{\sigma_1} & \frac{x_2^3-\mu_2}{\sigma_2} \end{pmatrix} = \begin{pmatrix} 1 & -0.873 \\ -1 & 1.091 \\ 0 & -0.218 \end{pmatrix}.$$

3. Compute the first principal component of $X$.

   *Solution:* First, we compute the eigenvalues of $X^\top X$ by solving the characteristic equation for the eigenvalues:

$$p(\lambda) = det(X^\top X - \lambda I) = 0.$$

$$X^\top X - \lambda I = \begin{pmatrix} 2-\lambda & -1.964 \\ -1.964 & 2-\lambda \end{pmatrix}$$

$$det(X^\top X - \lambda I) = (2-\lambda)(2-\lambda) - (-1.964)^2 = \lambda^2 - 4\lambda + 4 - (-1.964)^2.$$

   Solving the characteristic equation $det(X^\top X - \lambda I) = (2-\lambda)(2-\lambda) - (-1.964)^2 = \lambda^2 - 4\lambda + 4 - (-1.964)^2 = 0$ for $\lambda$ results in the two eigenvalues $\lambda_1 = 3.964$ and $\lambda_2 = 0.036$. The

eigenvector corresponding to $\lambda_1$ must satisfy $X^\top X v_1 = \lambda_1 v_1$, i.e.,

$$\begin{aligned}
&\begin{pmatrix} 2 & -1.964 \\ -1.964 & 2 \end{pmatrix} \begin{pmatrix} v_1^1 \\ v_2^1 \end{pmatrix} = 3.964 \begin{pmatrix} v_1^1 \\ v_2^1 \end{pmatrix} \\
\iff & 2v_1^1 - 1.964 v_2^1 = 3.964 v_1^1 \\
& -1.964 v_1^1 + 2v_2^1 = 3.964 v_2^1 \\
\iff & v_1^1 = \frac{-1.964}{3.964 - 2} v_2^1 \\
& v_1^1 = \frac{3.964 - 2}{-1.964} v_2^1 \\
\iff & v_1^1 = -v_2^1 \\
& v_1^1 = -v_2^1
\end{aligned}$$

$v_1 = (-1, 1)^\top$ satisfies the above equation and is therefore an eigenvector corresponding to eigenvalue $\lambda_1 = 3.964$. An eigenvector $v_2 = (1, 1)^\top$ corresponding to eigenvalue $\lambda_2 = 0.036$ is computed analogously. After normalization the eigenvectors are $v_1 = \frac{1}{\sqrt{2}}(-1, 1)^\top$ and $v_2 = \frac{1}{\sqrt{2}}(1, 1)^\top$. The eigenvectors $v_1$ and $v_2$ are the principal components of $X$. Since $\lambda_1 = 3.964 \geq 0.036 = \lambda_2$, $v_1 = \frac{1}{\sqrt{2}}(-1, 1)^\top$ is the first principal component and $v_2 = \frac{1}{\sqrt{2}}(1, 1)^\top$ is the second principal component.

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?

   *Solution:* Let $i = 1$ and set $\theta_1 = v_1$. We project our data onto the subspace $S = \langle \theta_1 \rangle \subset \mathbb{R}^2$, which is the span of the first eigenvector $v_1$, by computing $A = X\theta_1$.

$$X\theta_1 = \begin{pmatrix} 1 & -0.873 \\ -1 & 1.091 \\ 0 & -0.218. \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1.324 \\ 1.479 \\ -0.154 \end{pmatrix} \in \mathbb{R}^{3\times 1}.$$

   $A$ is the new feature.

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 $X - \hat{X}$?

   *Solution:* The matrix is reconstructed by computing $\hat{X} = X\theta_1\theta_1^\top$ which equals:

$$\hat{X} = X\theta_1\theta_1^\top = \begin{pmatrix} -1.324 \\ 1.479 \\ -0.154 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \end{pmatrix} = \begin{pmatrix} 0.936 & -0.936 \\ -1.046 & 1.046 \\ 0.109 & -0.109. \end{pmatrix}$$

   Compare this with the original matrix and observe that $\hat{X}$ is close to $X$:

$$X = \begin{pmatrix} 1 & -0.873 \\ -1 & 1.091 \\ 0 & -0.218. \end{pmatrix}$$

Furthermore, $\|X - \hat{X}\|_{\mathcal{F}}^2 = \text{trace}((X - \hat{X})^\top (X - \hat{X})) = 0.036$ which is similar to the value of the second eigenvalue of $X^\top X$. This can be seen as information lost in our data matrix by neglecting feature 2 and by projecting our data matrix onto a subspace spanned by the first principal component.

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.

   *Solution:* Following the hint, we compute $X^\top X$ using the SVD composition:

$$X^\top X = (USV^\top)^\top (USV^\top) = VS^\top U^\top USV^\top = VS^\top SV^\top = VS^2V^\top,$$

   where in the third equality we used that $U$ is orthonormal ($U^\top U = I$) and in the fourth equality we used that $S$ is a diagonal matrix ($S^\top S = S^2$). Note that $VS^2V^\top$ is the eigendecomposition of $X^\top X$ and therefore the columns of $V$ are the eigenvectors of $X^T X$ and the diagonal entries of $S^2$ are the eigenvalues of $X^T X$. Thus, the singular values of $X$ are the square roots of the eigenvalues of $X^\top X$.

**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.

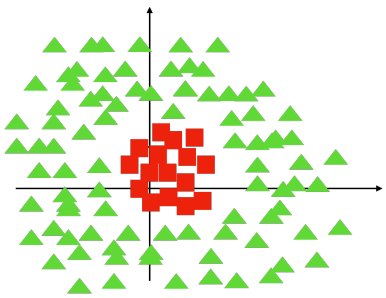

Figure 1: Classification problem training data

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?

   *Solution.* The gini index is $m/(1+m) * 1/(1+m) + 1/(1+m) * m/(1+m) = \frac{2m}{(1+m)^2}$. Since the classifier gets all the squares wrong, its error rate is $\frac{mN_t}{(1+m)N_t} = \frac{m}{1+m}$.

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.

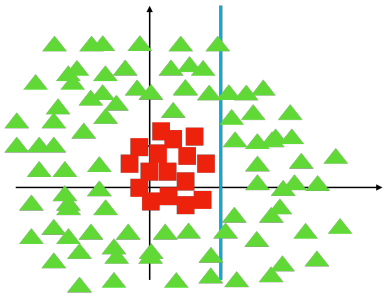

Figure 2: Classification problem with one node of the decision tree

*Solution.* The gini index of the leaf resulting from satisfaction of $x_1 > 1$ is 0. This is because the leaf is pure and all the data points are identified correctly as a triangle. The gini index of the leaf resulting from NOT satisfying $x_1 > 1$ is as follows.
First, there are $(1 - c)N_t$ triangles and $mN_t$ squares for a total of $(1 - c + m)N_t$ data points with $x_1 \leq 1$. Within these data points:
Probability of class triangle is $(1 - c)N_t/(1 - c + m)N_t = \frac{1-c}{1-c+m}$.
Probability (fraction) of class square is $mN_t/(1 - c + m)N_t = \frac{m}{1-c+m}$.
Finally, gini of this leaf is $\frac{1-c}{1-c+m} * \frac{m}{1-c+m} + \frac{m}{1-c+m} * \frac{1-c}{1-c+m} = \frac{2(1-c)m}{(1-c+m)^2}$.

It follows that the gini index of the node with this split is

$$0 \times \frac{c}{1+m} + \frac{2(1-c)m}{(1-c+m)^2} \times \frac{(1-c+m)}{1+m} = \frac{2(1-c)m}{(1+m)(1-c+m)}.$$

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

*Solution.* We need to compare the gini index after the split $\frac{2(1-c)m}{(1+m)(1-c+m)}$ to that before the split $\frac{2m}{(1+m)^2}$. In particular, we should show that $\frac{2m(1-c)}{(1-c+m)(1+m)} < \frac{2m}{(1+m)^2}$. Now, note that:

$$\begin{aligned}
& m > m(1-c) \quad \text{since } 1-c \in (0,1) \\
\iff & 1-c+m > 1-c+m(1-c) \text{ by adding } 1-c \text{ to both sides of the inequality} \\
\iff & \frac{1}{1-c+m} < \frac{1}{(1-c)(1+m)} \text{ taking the inverse of above} \\
\iff & \frac{1-c}{1-c+m} < \frac{1}{1+m} \text{ multiplying both sides by } 1-c \\
\iff & \frac{2m(1-c)}{(1-c+m)(1+m)} < \frac{2m}{(1+m)^2}
\end{aligned}$$

Hence, we arrived at the desired result starting from the fact that $(1 - c) \in (0, 1)$.

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.

   *Solution.* In any split, you will end up with more triangles than squares on *both sides*, so that the final decisions (in the leaves) will always be "triangle", and the accuracy will always be $\frac{1}{1+m}$.

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

   *Solution.* It will look like a square around the red squares.

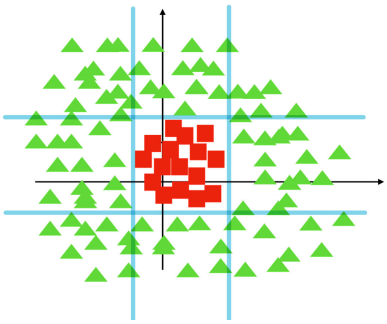

Figure 3: Square boundary corresponding to a decision tree

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

   *Solution.* It is a tree of depth 4, since you need 4 decision boundaries to carve it.

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?

   *Solution.*You would need to learn two parameters, one for each feature, corresponding to a circular decision boundary with center $(0, 0)$. You can construct your predictor as follows:

   (a) You learn the predictor $z$ which can be represented as

   $$z(x_1, x_2) := w_1\Phi_1(x_1, x_2) + w_2\Phi_2(x_1, x_2) = w_1\left(x_1^2 + x_2^2\right) + w_2.$$

   (b) You make your prediction based on $z(x_1, x_2)$ as

   $$\hat{y} = \begin{cases} \text{triangle}, z(x_1, x_2) \geq 0 \\ \text{square}, z(x_1, x_2) < 0. \end{cases} \tag{0.1}$$

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

Note that when $z(x_1, x_2) \geq 0$ which is $x_1^2 + x_2^2 \geq -\frac{w_2}{w_1}$, and this requires $w_1 > 0$ and $w_2 < 0$ in this case. A potential solution based on optimizing for the weights above could look like the following picture.

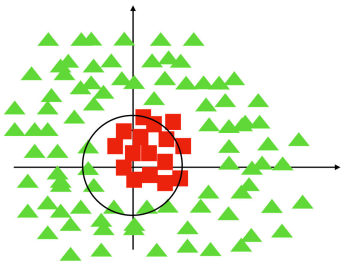

Figure 4: Circular boundary corresponding to logistic regression problem