Problem 1. (kNN)

1. A company wants to detect whether its product is faulty ($y = 1$) or intact ($y = 0$) based on the product's weight (x_1) and length (x_2). The company asks you to use the k-nearest neighbor approach with $k = 1$. For a product with $x^{test} = (2, 4)$, you are given two datapoints $x^1 = (2.2, 3.8)$ with $y^1 = 1$ and $x^2 = (2.3, 4.0)$ with $y^2 = 0$. Determine the label of the test point x^{test} based on the Manhattan distance.

   Solution: We compute the Manhattan distance as

   $$d(x^{test}, x^1) = |2 - 2.2| + |4 - 3.8| = 0.4$$
   $$d(x^{test}, x^2) = |2 - 2.3| + |4 - 4| = 0.3$$

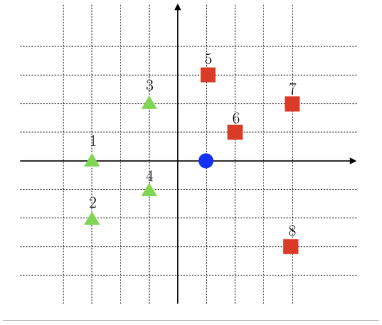
   The $1-$nearest neighbor method with Manhattan distance labels the product with $x^{test} = (2, 4)$ as belonging to the same class as x^2, so to class $Y = 0$ and thus intact.

2. Now, consider kNN for regression, where you want to predict the lifetime of the product in hours: $y \in \mathbb{R}$. Compute the label of x^{test} given that the labels of the $k = 3$ closest data points are $y^1 = 2.3$, $y^2 = 2.5$, and $y^3 = 1.8$.

   Solution: We simply average over the labels of the 3 closest data points and obtain

   $$y^{test} = \frac{1}{3}(2.3 + 2.5 + 1.8) = 2.2.$$

3. In the figure below, classify the new point (circle) using a kNN classifier. Choose the closest $k = 3$ neighbors among the data points 1 to 8 and their class label (△ or □) for different distance metrics.



	Closest $k = 3$ points	Label
L1 (Manhattan) distance	1, 2, 3, 4, 5, 6, 7, 8	□, △
L2 (Euclidean) distance	1, 2, 3, 4, 5, 6, 7, 8	□, △

Solution:

	Closest $k = 3$ points	Label
L1 (Manhattan) distance	4,5,6	□
L2 (Euclidean) distance	3,4,6	△

Problem 2. (k-means)

1. Problem 4.2 from Chapter 4 of Introduction to Applied Linear Algebra:

   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 lth entry of the j group representative.
   Solution: From above, we know that the lth 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 ith 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 μ^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 μ^j are nonnegative.

   The sum of the components of μ^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 μ^j sum also to one.

3. 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 3. (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:

$$X_r = \begin{pmatrix} x^1_1 & x^1_2 \\ x^2_1 & x^2_2 \\ x^3_1 & x^3_2 \end{pmatrix} = \begin{pmatrix} 30 & 1 \\ 10 & 2.5 \\ 20 & 1.5 \end{pmatrix}$$

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 μ_1, μ_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 λ results in the two eigenvalues $\lambda_1 = 3.964$ and $\lambda_2 = 0.036$. The eigenvector corresponding to λ_1 must satisfy $X^\top X v_1 = \lambda_1 v_1$, i.e.,

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

$v_1 = (-1, 1)^\top$ 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^TX and the singular values are the square root of the eigenvalues of $X^\top X$.

   Hint: Simply compute X^TX using the SVD decomposition and use the orthogonality of U and V to simplify.

   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^TX and the diagonal entries of S^2 are the eigenvalues of X^TX. Thus, the singular values of X are the square roots of the eigenvalues of $X^\top X$.