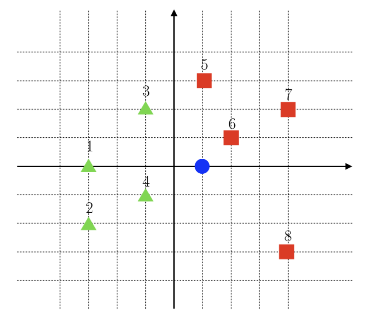
Problem 1. (kNN)

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.

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 $y^3 = 1.8$.

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 ($\triangle$ or $\square$) for different distance metrics.



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

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.

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

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:

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

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 compute $X^T X$ using the SVD decomposition and use the orthogonality of U and V to simplify.