Machine Learning Basics

Instructors: Mohammadreza A. Dehaqani, Babak n. Arabi, Mostafa Tavassolipour

Amir Naddaf Fahmideh, Kasra Hajiheidari

Fall 2025

Homework 4

Question 1: One Full Iteration of K-Means

Consider the dataset D consisting of five points. You are asked to perform one full iteration of k-means clustering on this dataset with $k = 2$, using Euclidean distance.

$$D = \begin{bmatrix} 5.5 & 3.1 \\ 5.1 & 4.8 \\ 6.6 & 3.0 \\ 5.5 & 4.6 \\ 6.8 & 3.8 \end{bmatrix}$$

The initial cluster centers (centroids) are:

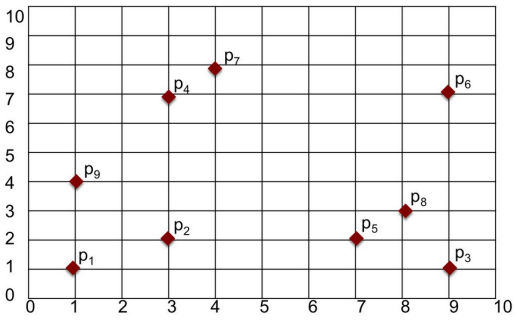
- $\mu_0 = (5.3, 3.5)$
- $\mu_1 = (5.1, 4.2)$

1. Assign each data point to the nearest cluster center.
2. Compute the updated cluster centers as the mean of the points assigned to each cluster.
3. Report the new center for Cluster 0 and the new center for Cluster 1.
4. State the final cardinality (number of points) of Cluster 0 after the update.

Question 2: Linkage Strategies in Agglomerative Clustering

In bottom-up agglomerative clustering, two common strategies for determining the distance between clusters are **single linkage** and **complete linkage**, which consider the minimum and maximum distance between cluster members as the merger criterion, respectively.

1. Compare the computational complexity of these two methods.
2. Which method is more robust to *outliers*? Justify your answer.
3. Using the **complete linkage** method, perform clustering on the following dataset and then draw the resulting **dendrogram**.



Point	x	y
p_1	1	1
p_2	3	2
p_3	9	1
p_4	3	7
p_5	7	2
p_6	9	7
p_7	4	8
p_8	8	3
p_9	1	4

Question 3: Dimensionality Reduction and Data Geometry

Consider several datasets with different geometric and statistical properties. These include linearly correlated Gaussian data, two concentric circles, high-dimensional data with strong class overlap, and data containing many irrelevant or noisy features.

For each case, determine whether PCA, Kernel PCA, or LDA is the most appropriate dimensionality reduction method. Justify your choice by referring to the objective function of the method, the geometry of the data, and the assumptions underlying each technique. Your explanation should clearly demonstrate an understanding of why certain methods succeed or fail depending on the data distribution.

Question 4: Formal Theory of LDA and Kernel PCA

Prove that for a classification problem with C classes, Linear Discriminant Analysis can produce at most $C-1$ nonzero discriminant directions. Your proof should rely on rank arguments involving the betweenclass scatter matrix.

Then consider Kernel PCA. Starting from the covariance operator in feature space, explain why direct eigendecomposition is infeasible. Show that every principal component in feature space can be written as a linear combination of mapped training samples. Derive the kernel eigenvalue problem and obtain the expression for projecting a new data point using only kernel evaluations.

Question 5: Manual Computation of LDA

Apply Linear Discriminant Analysis by hand to the following synthetic dataset consisting of two classes

$$Class1 : (4,2), (2,4), (2,3), (3,6), (4,4)$$

$$Class2 : (9,10), (6,8), (9,5), (8,7), (10,8)$$

Compute the sample mean of each class and the overall mean. Derive the within-class scatter matrix and the between-class scatter matrix. Solve the generalized eigenvalue problem associated with Fisher's criterion and determine the optimal discriminant direction. Finally, project all data points onto this direction and report their one-dimensional coordinates.

Question 6: Error Types

Feature selection methods can be broadly categorized into filter, wrapper, and embedded approaches. Among wrapper methods, Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and bidirectional (stepwise) selection are widely used due to their conceptual simplicity and model-driven nature.

- Explain in detail how Sequential Forward Selection and Sequential Backward Selection operate, including their initialization, update steps, and stopping criteria. Then describe how bidirectional selection combines elements of both approaches. In your explanation, clearly discuss why these methods are considered greedy algorithms and what is meant by "greedy" in this context.
- Analyze the consequences of this greedy behavior. Explain why SFS, SBS, and bidirectional selection are not guaranteed to find the globally optimal subset of features, even when the underlying predictive model is convex. Provide a concrete example or scenario in which an early selection or removal decision leads to a suboptimal final feature subset.
- Compare sequential selection methods with two alternative approaches: Recursive Feature Elimination (RFE) and feature selection based on L1 regularization (such as Lasso). Your comparison should address computational complexity, sensitivity to feature correlations, ability to capture feature interactions, and suitability for high-dimensional datasets.

Question 7: PCA and Kernel PCA on Structured Data

In this assignment, you will first analyze PCA on a real low-dimensional dataset and then investigate the limitations of linear methods on nonlinear data.

a) Begin by loading the Iris dataset using the function sklearn.datasets.load_iris. After standardizing the data, apply Principal Component Analysis to reduce the dimensionality of the dataset from four features to two principal components. Compute the eigenvalues of the covariance matrix and report the explained variance ratio associated with each principal component. Visualize the data in the twodimensional PCA space, using different colors to represent the three classes.

After producing the visualization, analyze the result. Explain which principal component captures the largest amount of variance and provide an intuitive reason for this behavior. Discuss whether the classes appear to be well separated in the reduced space and explain why this is expected given the objective function of PCA. Finally, comment on whether PCA is an appropriate dimensionality reduction technique if the ultimate goal is classification, and justify your answer based on the properties of PCA.

Next, consider how the result would change if Linear Discriminant Analysis were applied instead of PCA. Without implementing LDA, explain conceptually how the projection would differ. Your answer should explicitly refer to the objective functions of PCA and LDA, the use of class labels, and the expected effect on class separability in the projected space.

b) In the second part of this assignment, you will examine a nonlinear dataset. Generate a twodimensional dataset consisting of two nested circles using the function sklearn.datasets.make_circles. This dataset is not linearly separable. Apply standard PCA to the data and visualize the result in two dimensions. Then apply Kernel PCA using a Gaussian (RBF) kernel and project the data into two dimensions. Plot both projections and color the points according to their class labels.

Compare the results obtained using PCA and Kernel PCA. Explain why standard PCA fails to separate the two classes, while Kernel PCA is able to do so. In your discussion, describe the role of nonlinear feature mappings and explain how the kernel trick allows Kernel PCA to capture the underlying structure of the data.

Question 8: Feature Selection and Visualization on a Real Dataset

In this assignment, you will perform feature selection using multiple techniques and compare their behavior, stability, and interpretability on a real-world dataset.

Use the Heart Disease Dataset. Treat the task as a binary classification problem,

where the goal is to predict whether a patient has heart disease based on clinical features.

Begin by performing an exploratory analysis of the dataset. Visualize feature distributions and compute the correlation matrix between input features. Based on this analysis, apply a correlation-based filter method to identify and remove redundant features. If no features are removed, explain why correlation-based filtering may be ineffective for this dataset.

Next, apply a univariate feature selection method (such as mutual information) to rank features according to their relevance to the target variable. Select the top-k features and evaluate the performance of a classifier trained using only these features. Compare the results with a model trained on the full feature set. Use the same classifier and evaluation protocol for both models.

Then apply Recursive Feature Elimination (RFE) using an appropriate classifier (for example, logistic regression or a linear support vector machine). Report the selected features and compare them with those obtained from the univariate method, both in terms of feature overlap and classification performance.

Extend this analysis by applying Recursive Feature Elimination with Cross-Validation (RFECV). Discuss how incorporating cross-validation affects the robustness and stability of the selected feature subset. If RFECV selects most or all features, explain what this implies about feature redundancy, sample size, and model bias.

Afterward, apply an embedded feature selection method using L1-regularized logistic regression (Lasso) and analyze feature importance based on coefficient sparsity. Compare the importance ranking produced by model with the features selected by RFE and univariate methods. Discuss similarities and differences, particularly in the presence of correlated features.

Finally, apply Principal Component Analysis (PCA) to the dataset as an exploratory tool. Analyze how many components are required to explain a large proportion of variance, and explain why PCA-based dimensionality reduction is conceptually different from feature selection in terms of interpretability and model behavior. Ultimately, compare the number of components with the number of features selected by RFE and RFECV.

Your final report should compare all feature selection approaches in terms of predictive performance, interpretability, sensitivity to feature correlations, and computational cost.

Question 9: Image Segmentation (Bonus)

Using the K-Means clustering algorithm with minimum Euclidean-distance-based assignments of samples to cluster centroids, segment the two attached color images into $K \in \{2, 3, 4, 5\}$ segments. As the feature vector for each pixel use a 5-dimensional feature vector consisting of normalized vertical and horizontal coordinates of the pixel relative to the top-left corner of the image, as well as normalized red, green, and blue values of the image color at that pixel. Normalize each feature by linearly shifting and scaling the values to the interval $[0, 1]$, such that the set of 5-dimensional normalized feature vectors representing each pixel are in the unit-hypercube $[0, 1]^5$.

For each $K \in \{2, 3, 4, 5\}$, let the algorithm assign labels to each pixel; specifically, label $l_{rc} \in \{1, \ldots, K\}$ to the pixel located at row r and column c. Present your clustering results in the form of an image of these label values. Make sure you improve this segmentation outcome visualization by using a contrast enhancement method; for instance, assign a unique color value to each label and make your label image colored, or assign visually distinct grayscale value levels to each label value to make best use of the range of gray values at your disposal for visualization.

Repeat this segmentation exercise using GMM-based clustering. For each specific K, use the EM algorithm to fit a GMM with K components, and then use that GMM to do MAP-classification style cluster label assignments to pixels. Display results similarly for this alternative clustering method. Briefly comment on the reasons of any differences, if any.