In the name of GOD.

Sharif University of Technology

Machine Learning

Spring 2025
Hamid R. Rabiee, Zahra Dehghanian

Homework 5 Deadline: 1404/02/28

1. (15)

Consider the K-Nearest Neighbors (KNN) algorithm and answer the following parts:

(A) – Distance Metrics

Suppose we have two points $A = (a_1, a_2)$ and $B = (b_1, b_2)$ in $\mathbb{R}^2$ (with $a_1 < b_1$, $a_2 < b_2$, so they are not collinear). Derive the locus of points (x, y) that are equidistant to A and B under each metric below:

(i) **Euclidean (L2) distance:**

$$d((x, y), A) = d((x, y), B)$$

Simplify to find the equation of the boundary line.

(ii) **Manhattan (L1) distance:**

$$|x - a_1| + |y - a_2| = |x - b_1| + |y - b_2|$$

Describe the shape of this boundary (it will be piecewise-linear rather than a single straight line).

Explain qualitatively how the choice of L2 vs L1 changes the decision boundary of a 1-NN classifier in this simple case.

(B) Bias-Variance Tradeoff

Explain how the choice of k (the number of neighbors) affects the bias and variance of a KNN model (for either classification or regression). In particular, discuss what happens in the limits $k = 1$ and $k = n$ (where n is the training set size). Relate your answer to the bias–variance decomposition of prediction error. You may also mention asymptotic consistency results for KNN (e.g., how KNN error approaches the Bayes error for large n with appropriate k).

(C) Kernel Weighting

In KNN regression, one can assign weights to neighbors based on distance. Let the weight of neighbor i be

$$w_i = K(d(x, x_i))$$

where $d(\cdot, \cdot)$ is the chosen distance.

(a) Write the formula for the kernel-weighted KNN regression prediction $\hat{f}(x)$ in terms of w_i and y_i.

(b) Now consider using a Gaussian kernel

$$K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{\sigma^2}\right).$$

Describe qualitatively what happens as $\sigma \to 0$ and as $\sigma \to \infty$.

2. (25) Prove that in Principal Component Analysis (PCA), the reconstruction error after projecting onto the first k principal components is equal to the sum of the discarded eigenvalues, where the reconstruction error is computed using the Frobenius norm. Specifically, show:

$$\frac{1}{n}\|X - \hat{X}\|_F^2 = \sum_{i=k+1}^{d} \lambda_i$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of the data covariance matrix and the data matrix X is zero-centered (i.e., each column has mean zero).

3. (15) Suppose you have recently been hired as a machine learning specialist at a financial services company. One of the projects assigned to you is to develop a model to identify suspicious behavior in customer data. The available dataset includes features of transactions and user profiles. Each sample also has a label indicating whether the behavior was suspicious (e.g., fraudulent) or not.

(a) When the goal is to accurately detect suspicious behaviors, which method do you prefer between Bagging and Boosting? Justify your answer in terms of bias and variance reduction.

(b) If your dataset contains noisy labels (e.g., incorrect class labels) or outliers, which method would you choose? Why?

(c) Given limited computational resources and the need to deploy a model quickly, which method is a better choice in terms of training time and parallelization?

(d) One of the students has proposed the following to improve the performance of Boosting: Instead of using a single Boosting model, train several Boosting models in parallel and independently, each on a different bootstrap sample of the data. The final output is determined by majority voting among them. Conceptually, this is similar to applying Bagging on Boosting models.

Explain the advantages and disadvantages of this proposed method compared to standard Boosting, especially in the presence of noise/outliers and in terms of execution time.

4. (15) Suppose in an ensemble method, K identical base models are independently trained on data obtained using bootstrap sampling (random sampling with replacement). The final prediction is obtained by averaging the outputs of these models.

Prove that as the number of models K increases, the final prediction error (mean squared error) does **not worsen**.

Use the decomposition of prediction error into three components:

- **Bias**
- **Variance**

- **Irreducible Noise**

5. (15 + 15) Complete the attached notebooks. You are permitted to use chatbots or other resources for assistance, but you must ensure that you fully understand the code and implementations. During the online sessions, you may be asked to explain specific functions, code lines, and your overall approach. Be prepared to demonstrate your understanding in detail.