In the name of GOD.

Sharif University of Technology

Machine Learning

Spring 2025

Hamid R. Rabiee, Zahra Dehghanian

Homework 4 — Deadline: 1404/02/07

1. (35) Let $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ be a binary classification dataset where $y^{(i)} \in \{-1, +1\}$, and assume all inputs are distinct and satisfy the condition:

$$\|x^{(i)} - x^{(j)}\| \geq \epsilon > 0 \quad \text{for all } i \neq j.$$

Consider training a Support Vector Machine (SVM) using the Gaussian (RBF) kernel:

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

The SVM decision function in its dual form is given by:

$$f(x) = \sum_{i=1}^{n} \alpha_i y^{(i)} K(x^{(i)}, x) + b$$

   (a) Show that there exists a value of $\tau > 0$ such that the SVM achieves zero training error. *Hint: Use the dual form of the SVM and analyze the behavior of the kernel matrix as $\tau \to 0$.*

   (b) Suppose $\alpha_i = 1$ for all i, and the bias term is set to $b = 0$. Derive an explicit upper bound on τ in terms of ϵ and n that guarantees zero training error.

2. (15) a) Compare the Gini and Entropy criteria in terms of their value ranges. Based on their mathematical definitions, what do you think are the advantages and disadvantages of using each compared to the other?

b) Train a decision tree using the Gini criterion on the following data points with input variables x_1, x_2, and target variable y. Then, predict the value of y for the new data point ($x_1 = 1.5,\ x_2 = 1.5$). Does the result change if the Entropy criterion is used instead?

3. (25) In a small town where communication relies solely on sending encoded strings made up of lowercase English letters (a–z), two friends, Sam and Tina, want to check how similar their encoded messages are. They define a special function, called the **"Secret Code Kernel"**, denoted by $K(s, t)$, which measures the *number of common subsequences* between Sam's message s and Tina's message t.

Formally,

$$K(s, t) = \text{The number of equal subsequences of } s \text{ and } t.$$

For instance, if Sam sends "abc" and Tina replies with "cbc", they calculate $K(s, t) = 4$, since they both share:

x_1	x_2	y
1.5	2.0	1
3.0	2.8	0
1.2	0.7	1
4.2	2.1	1
1.5	1.8	0
0.9	1.2	0
3.1	2.1	1
2.7	2.3	1
1.0	0.8	0
2.8	1.9	1

Table 1: Training data

(a) The letter ‘c’ (once in each string),

(b) Another ‘c’ from a different position match,

(c) The letter ‘b’, and

(d) The sequence ‘bc’.

Your task: Prove that this kernel function is valid (i.e., it satisfies the properties of a kernel), and construct the corresponding feature mapping $\Phi(x)$ that represents each string in a suitable feature space.

4. (25) Complete the attached notebook. 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.