Theoretical Assignment 1

Intro to Machine Learning, TAU

November 18, 2024

Question 1

Imagine you are classifying flowers as either *Iris-setosa* or *Iris-versicolor* based on two features: petal length and petal width. You have the following information:

Class	**Mean Petal Length**	**Std Dev Petal Length**	**Mean Petal Width**	**Std Dev Petal Width**
Iris-setosa	1.5	0.2	0.2	0.1
Iris-versicolor	4.7	0.3	1.4	0.2

Table 1: Summary statistics for petal length and width by flower class.

Assume that the features are independent and normally distributed.

1. Calculate the likelihoods $P(\text{petal length} = 1.6 \mid \text{Iris-setosa})$ and $P(\text{petal width} = 0.25 \mid \text{Iris-setosa})$ using the probability density function of a normal distribution, and plot them.

2. Based on these likelihoods, design a classifier that receives the petal width and petal length and predicts the the class.

3. Using the classifier from the previous section, determine which class (Iris-setosa or Iris-versicolor) is more likely for a flower with a petal length of 1.6 and a petal width of 0.25. Assume equal priors for both classes.

Question 2

A medical test is used to classify patients into two categories: *Disease* or *No Disease*. However, due to the high costs associated with misclassification, there is also an option to reject a classification if confidence is too low. You are given the following information:

- $P(\text{Disease}) = 0.2$

- $P(\text{No Disease}) = 0.8$
- The test has a sensitivity (true positive rate) of 90%, meaning $P(\text{Positive Test} \mid \text{Disease}) = 0.9$.
- The test has a specificity (true negative rate) of 85%, meaning $P(\text{Negative Test} \mid \text{No Disease}) = 0.85$.

A patient receives a positive test result.
Cost Structure:

- The cost of a false positive (FP) error (i.e., diagnosing a healthy patient as having the disease) is \$1,000.
- The cost of a false negative (FN) error (i.e., failing to diagnose a patient with the disease) is \$5,000.
- The cost of rejecting the classification for further testing is \$2,000.

Tasks:

1. Calculate the probability that this patient actually has the disease given a positive test result, $P(\text{Disease} \mid \text{Positive Test})$.
2. Based on this probability and the cost structure, decide whether to classify the patient as *Disease*, *No Disease*, or *Reject*. Choose the option with the lowest expected cost. Use a confidence threshold of 70% for deciding whether to classify the patient as having the disease. If the probability of disease is below 70% but above 30%, classify as *Reject*; otherwise, classify as *No Disease*.

Question 3

Given a real number $R \geq 0$, define the hypothesis $h_R : \mathbb{R}^d \to \{0, 1\}$ as follows:

$$h_R(x) = \begin{cases} 1 & \text{if } \|x\|_2 \leq R, \\ 0 & \text{otherwise.} \end{cases}$$

Consider the hypothesis class $H_{\text{ball}} = \{h_R \mid R \geq 0\}$. Prove directly (without using the Fundamental Theorem of PAC Learning) that H_{ball} is PAC learnable in the realizable case. Assume for simplicity that the marginal distribution of X is continuous. How does the sample complexity depend on the dimension d? Explain.

Question 4

Given a polynomial $P : \mathbb{R} \to \mathbb{R}$, define the hypothesis $h_P : \mathbb{R}^2 \to \{0, 1\}$ as follows:

$$h_P(x_1, x_2) = \begin{cases} 1 & \text{if } P(x_1) \geq x_2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the VC-dimension of $H_{\text{poly}} = \{h_P \mid P \text{ is a polynomial}\}$. You can use the fact that given n distinct values $x_1, \ldots, x_n \in \mathbb{R}$ and $z_1, \ldots, z_n \in \mathbb{R}$, there exists a polynomial P of degree $n - 1$ such that $P(x_i) = z_i$ for every $1 \leq i \leq n$.