Statistical Foundations of Learning - CIT4230004
Assignment 1 Solutions

Summer Semester 2024

Overview

This assignment covers the following topics:

- Bayes Risk and Bayes Classifier
- VC Dimension
- Universal Consistency of ϵ-neighbourhood classifiers

Each problem involves calculating theoretical properties and demonstrating proofs of given statements.

Problem 1: Bayes Risk I

Given: $X = Y = \{1, 2, 3\}$. Label distribution:

$$P(Y = j) = \begin{cases} 1/4 & \text{if } j = 1, 2 \\ 1/2 & \text{if } j = 3 \end{cases}$$

Conditional feature distributions:

$$P(X = i|Y = 1) = \begin{cases} 1/3 & \text{if } i = 2 \\ 2/3 & \text{if } i = 3 \end{cases}$$

$$P(X = i|Y = 2) = \begin{cases} 1/2 & \text{if } i = 1 \\ 1/2 & \text{if } i = 3 \end{cases}$$

$$P(X = i|Y = 3) = \begin{cases} 2/3 & \text{if } i = 1 \\ 1/3 & \text{if } i = 2 \end{cases}$$

(a) Compute the Bayes classifier

Solution: The Bayes classifier h^* maximizes $P(Y = y|X = x)$.

For $x = 1$:

$$P(Y = 1|X = 1) = 0, \quad P(Y = 2|X = 1) = \frac{1/2 \cdot 1/4}{P(X = 1)} = \frac{1/8}{P(X = 1)}, \quad P(Y = 3|X = 1) = \frac{2/3 \cdot 1/2}{P(X = 1)} = P$$

Thus, $h^*(1) = 3$.

For $x = 2$:

$$P(Y = 1|X = 2) = \frac{1/3 \cdot 1/4}{P(X = 2)} = \frac{1/12}{P(X = 2)}, \quad P(Y = 2|X = 2) = 0, \quad P(Y = 3|X = 2) = \frac{1/3 \cdot 1/2}{P(X = 2)} = P$$

Thus, $h^*(2) = 3$.

For $x = 3$:

$$P(Y = 1|X = 3) = \frac{2/3 \cdot 1/4}{P(X = 3)} = \frac{1/6}{P(X = 3)}, \quad P(Y = 2|X = 3) = \frac{1/2 \cdot 1/4}{P(X = 3)} = \frac{1/8}{P(X = 3)}, \quad P(Y = 3|X =$$

Thus, $h^*(3) = 1$.

The Bayes classifier h^* is:

$$h^*(x) = \begin{cases} 3 & \text{if } x = 1 \\ 3 & \text{if } x = 2 \\ 1 & \text{if } x = 3 \end{cases}$$

(b) Compute the Bayes risk

Solution: The Bayes risk R^* is the expected loss of the Bayes classifier.

$$R^* = \sum_x \min_y P(Y = y|X = x)P(X = x)$$

For $x = 1$:

$$\min(P(Y = 1|X = 1), P(Y = 2|X = 1), P(Y = 3|X = 1)) = \min(0, \frac{1/8}{P(X = 1)}, \frac{1/3}{P(X = 1)}) = 0$$

For $x = 2$:

$$\min(P(Y = 1|X = 2), P(Y = 2|X = 2), P(Y = 3|X = 2)) = \min(\frac{1/12}{P(X = 2)}, 0, \frac{1/6}{P(X = 2)}) = 0$$

For $x = 3$:

$$\min(P(Y = 1|X = 3), P(Y = 2|X = 3), P(Y = 3|X = 3)) = \min(\frac{1/6}{P(X = 3)}, \frac{1/8}{P(X = 3)}, 0) = 0$$

Thus, the Bayes risk R^* is:

$$R^* = 0$$

Problem 2: Bayes Risk II

Given: $X = \{1, 2, \ldots, 30\}$, $Y = \{\pm 1\}$ and class probability:

$$\eta(x) = P(y = +1|x) = \begin{cases} 1 - \alpha & \text{if } x \in \{11, 12, \ldots, 20\} \\ \alpha & \text{otherwise} \end{cases}$$

(a) Compute the Bayes risk and classifier

Solution: The Bayes classifier h^* maximizes $P(Y = y|X = x)$.
The Bayes classifier h^* is:

$$h^*(x) = \begin{cases} +1 & \text{if } x \in \{11, 12, \ldots, 20\} \\ -1 & \text{otherwise} \end{cases}$$

The Bayes risk R^* is:

$$R^* = \sum_x \min(\eta(x), 1 - \eta(x)) P(X = x)$$

Define $q_1 = \sum_{i=1}^{10} p_i$, $q_2 = \sum_{i=11}^{20} p_i$, $q_3 = \sum_{i=21}^{30} p_i$.

$$R^* = \alpha(q_1 + q_3) + (1 - \alpha) q_2$$

Problem 3: Universal Consistency of ϵ-neighbourhood classifiers

Given: Domain $X \subseteq \mathbb{R}$. Training sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset X \times \{\pm 1\}$, $\epsilon > 0$.

(a) Express $h_{S,\epsilon}$ as a plug-in classifier with a weighted average estimator $\hat{\eta}$

Solution:

$$\hat{\eta}(x) = \frac{1}{|\{i : |x_i - x| \leq \epsilon\}|} \sum_{i : |x_i - x| \leq \epsilon} y_i$$

$$h_{S,\epsilon}(x) = \text{sign}(\hat{\eta}(x))$$

(b) Simplify $\hat{\eta}$ for $X = \{0, 1\}$ and $\epsilon < 1$. Show $\hat{\eta}(x)$ converges to $\eta(x)$ in probability as $m \to \infty$

Solution: For $X = \{0, 1\}$:

$$\hat{\eta}(0) = \frac{\sum_{i : x_i = 0} y_i}{|\{i : x_i = 0\}|}$$

$$\hat{\eta}(1) = \frac{\sum_{i:x_i=1} y_i}{|\{i : x_i = 1\}|}$$

Both are binomial averages.

Using the Law of Large Numbers for binomially distributed variables:

$$\hat{\eta}(0) \to \eta(0) \quad \text{and} \quad \hat{\eta}(1) \to \eta(1) \quad \text{in probability as} \quad m \ \to \infty$$

This means that the proportion of y_i values correctly estimating $\eta(x)$ converges as the sample size increases.

(c) Show ϵ-neighbourhood classifier is universally consistent on $X = \{0, 1\}$ for any $\epsilon < 1$ without using Stone's theorem

Solution: A classifier is universally consistent if its risk converges to the Bayes risk as the sample size $m \to \infty$.

For a plug-in classifier with estimator $\hat{\eta}$:

$$R(h_{S,\epsilon}) - R^* \leq 2\mathbb{E}[|\hat{\eta}(x) - \eta(x)|]$$

As $m \to \infty$, $\hat{\eta}(x) \to \eta(x)$ in probability, so the risk difference converges to zero.

Since $\hat{\eta}(x) \to \eta(x)$ in probability, the expected absolute difference $\mathbb{E}[|\hat{\eta}(x) - \eta(x)|] \to 0$.

Hence, the ϵ-neighbourhood classifier's risk converges to the Bayes risk, proving universal consistency for any $\epsilon < 1$.

Problem 4: VC Dimension

Given: $v_1, \ldots, v_n \in \mathbb{R}^d$ for some $n < d$. Define the hypothesis class:

$$\mathcal{H} = \left\{ x \mapsto \text{sign}\left(\sum_{i=1}^{n} \alpha_i \langle v_i, x \rangle + b \right) \mid \alpha_1, \ldots, \alpha_n, b \in \mathbb{R} \right\}$$

(a) Show that VCdim($\mathcal{H}$) $\leq n + 1$

Solution:

Definitions and Preliminary Steps

1. **VC Dimension Definition:** The Vapnik-Chervonenkis (VC) dimension of a hypothesis class $\mathcal{H}$ is the largest number of points that can be shattered by $\mathcal{H}$. A set of points is said to be shattered by $\mathcal{H}$ if, for every possible way of labeling these points, there exists a hypothesis in $\mathcal{H}$ that correctly classifies the points according to the labels.

2. **Hypothesis Class:**

$$\mathcal{H} = \left\{ x \mapsto \text{sign}\left(\sum_{i=1}^{n} \alpha_i \langle v_i, x \rangle + b \right) \mid \alpha_1, \ldots, \alpha_n, b \in \mathbb{R} \right\}$$

Upper Bound on the VC Dimension

1. **Linear Combination and Affine Function:** The hypothesis in $\mathcal{H}$ is a sign of an affine function defined as:

$$f(x) = \sum_{i=1}^{n} \alpha_i \langle v_i, x \rangle + b$$

This is a linear combination of n inner products plus a bias term b.

2. **Points in $\mathbb{R}^d$:** Given that $n < d$, we have fewer vectors v_i than the dimensionality of the space. This means our linear combination is restricted to n degrees of freedom.

3. **Affine Hyperplanes:** Each hypothesis in $\mathcal{H}$ corresponds to a decision boundary (hyperplane) in $\mathbb{R}^d$. The position and orientation of this hyperplane are determined by the coefficients α_i and the bias b.

4. **VC Dimension and Hyperplanes:** The VC dimension of the class of affine hyperplanes in $\mathbb{R}^d$ is $d + 1$. However, since our affine hyperplanes are determined by only n vectors, we are limited to n dimensions of freedom.

5. **Shattering $n+1$ Points:** To shatter $n+1$ points, we need to classify all possible 2^{n+1} labelings of these points. However, because our affine functions are constrained by only n degrees of freedom, we cannot create 2^{n+1} distinct classifications. Thus, the maximum number of points we can shatter is $n + 1$.

Conclusion

We have established that:

$$\text{VCdim}(\mathcal{H}) \leq n + 1$$

Thus, the VC dimension of the given hypothesis class $\mathcal{H}$ is at most $n + 1$.

Approach to Solving These Questions

1. **Bayes Risk and Bayes Classifier:** - Identify the given distributions and conditional probabilities. - Use the definitions of the Bayes classifier and Bayes risk to derive the classifier and compute the risk. - Calculate the posterior probabilities and select the label with the highest probability for the classifier. - Sum the minimum posterior probabilities to find the Bayes risk.

2. **VC Dimension:** - Understand the definition of VC dimension and how it relates to the capacity of a hypothesis class. - Identify the hypothesis class and how it can shatter a set of points. - Use the properties of affine functions and linear combinations to determine the maximum number of points that can be shattered. - Conclude by establishing the upper bound of the VC dimension.

3. **Universal Consistency:** - Define the ϵ-neighbourhood classifier and the weighted average estimator. - Show convergence using the Law of Large Numbers and properties of binomial distributions. - Prove consistency by comparing the risk of the classifier to the Bayes risk and demonstrating convergence as the sample size increases.