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. The following sections explain the concepts, the approach taken to solve each problem, and key points to remember.

Problem 1: Bayes Risk I

Concepts: - **Bayes Risk:** The minimum possible risk (error) that can be achieved by any classifier for a given distribution of data. - **Bayes Classifier:** A classifier that assigns each data point to the class with the highest posterior probability.

Approach: To compute the Bayes classifier and the Bayes risk: 1. **Calculate Posterior Probabilities:** Use the given distributions and conditional probabilities to calculate the posterior probabilities $P(Y = y|X = x)$. 2. **Determine Bayes Classifier:** Assign each x to the class y that maximizes $P(Y = y|X = x)$. 3. **Compute Bayes Risk:** Sum the minimum conditional probabilities across all x to find the expected error rate.

Key Points to Remember:

- The Bayes classifier minimizes the conditional risk for each observation. - Bayes risk is the theoretical lower bound on the error rate for any classifier. - Posterior probabilities are derived from the given distributions and conditional probabilities.

Problem 2: Bayes Risk II

Concepts: - **Conditional Probability:** The probability of an event occurring given that another event has already occurred. - **Class Probability Function $\eta(x)$:** Represents the probability of a particular class given an input x.

Approach: To compute the Bayes classifier and the Bayes risk for a given conditional probability function: 1. **Identify Class Probability Function:** Use the given function $\eta(x)$ to determine the class probabilities. 2. **Determine Bayes Classifier:** Assign each x to the class y that has the higher probability according to $\eta(x)$. 3. **Compute Bayes Risk:** Calculate the expected error by integrating the minimum class probabilities over the distribution of x.

Key Points to Remember:

- The Bayes classifier uses the class probability function to assign labels. - Bayes risk involves integrating the error probabilities over the input distribution. - Understanding the distribution of x and the conditional class probabilities is crucial.

Problem 3: Universal Consistency of ϵ-neighbourhood classifiers

Concepts: - **ϵ-neighbourhood Classifier:** A non-parametric classifier that makes predictions based on the majority class within an ϵ-radius neighborhood of each input. - **Weighted Average Estimator:** An estimator that computes the weighted average of observed values within a neighborhood.

Approach: To analyze the universal consistency of ϵ-neighbourhood classifiers: 1. **Define the Estimator:** Express the ϵ-neighbourhood classifier as a plug-in classifier using a weighted average estimator. 2. **Simplify for Specific Cases:** Simplify the estimator for cases where $X = \{0, 1\}$ and show convergence to the true class probability function $\eta(x)$. 3. **Prove Consistency:** Show that the risk of the ϵ-neighbourhood classifier converges to the Bayes risk as the sample size increases, demonstrating universal consistency.

Key Points to Remember:

- The ϵ-neighbourhood classifier is non-parametric and relies on local information. - Consistency is shown by proving that the estimator converges to the true class probability function. - Universal consistency means that the classifier performs well for any distribution as the sample size grows.

Problem 4: VC Dimension

Concepts: - **VC Dimension:** A measure of the capacity of a hypothesis class, defined as the largest number of points that can be shattered (correctly classified) by the class. - **Affine Functions:** Functions defined by linear combinations of input variables plus a bias term.

Approach: To determine the VC dimension of a given hypothesis class: 1. **Analyze the Hypothesis Class:** Understand the structure of the hypothesis class and how it defines decision boundaries. 2. **Upper Bound on VC Dimension:** Show that the hypothesis class cannot shatter more than a certain number of points by analyzing the geometric properties of the decision boundaries. 3. **Exact VC Dimension:** Prove conditions under which the hypothesis class can shatter a specific number of points, thereby determining the exact VC dimension.

Key Points to Remember:

- The VC dimension provides a measure of the complexity and capacity of a hypothesis class. - Understanding the geometric properties of the hypothesis class helps in determining the VC dimension. - The exact VC dimension is found by analyzing the conditions under which the hypothesis class can shatter a set of points.