Statistical Foundations of Learning - Concepts and Key Things to Remember

CIT4230004 (Summer Semester 2024)

Assignment 3: PAC Learning and VC Dimension

Concepts:

1. **PAC Learning:** - A class H is agnostic PAC learnable if there exists an algorithm that, for every $\epsilon > 0$ and $\delta > 0$, produces a hypothesis h such that the true error $L_D(h) \leq \inf_{h' \in H} L_D(h') + \epsilon$ with probability at least $1 - \delta$. - **Markov's Inequality:** Used to bound the probability that a non-negative random variable exceeds a certain value.

2. **VC Dimension:** - The VC dimension of a hypothesis class H is the largest number of points that can be shattered by H. - **Sauer's Lemma:** Relates the growth function to the VC dimension.

Key Things to Remember: - Understand the definition and implications of PAC learnability. - Use Markov's inequality and Sauer's lemma for proofs and bounds related to PAC learning and VC dimension. - The VC dimension provides a measure of the complexity of the hypothesis class.

Assignment 4: Validation, Rademacher Complexity, and Gaussian Kernels

Concepts:

1. **Validation:** - **Leave-One-Out Error:** An unbiased estimator of the true error. - **Generalization Error:** Expected error on new, unseen data.

2. **Rademacher Complexity:** - Measures the ability of a hypothesis class to fit random noise. - **Hoeffding's Inequality:** Used for deriving bounds on Rademacher complexity.

3. **Gaussian Kernels:** - **Universality:** A kernel is universal if it can approximate any continuous function to arbitrary accuracy.

Key Things to Remember: - Validation techniques help in estimating the true performance of a model. - Rademacher complexity provides a way to measure the capacity of a hypothesis class. - Gaussian kernels are powerful due to their universal approximation properties.

Assignment 5: SVM and One-Class SVM

Concepts:

1. **SVM (Support Vector Machine):** - **Hard SVM:** For separable data, finds the maximum margin separator. - **Soft SVM:** For non-separable data, introduces slack variables to allow some misclassifications.

2. **One-Class SVM:** - Used for anomaly detection. - **Tikhonov Regularization:** Regularization technique used to prevent overfitting.

Key Things to Remember: - Understand the difference between hard and soft SVM and when to use each. - One-Class SVM is useful for detecting outliers or anomalies. - Regularization helps in controlling the complexity of the model.

Sample Sheet 1: Probability Bounds and Bayes Risk

Concepts:

1. **Probability Bounds:** - **Markov's Inequality:** Provides an upper bound on the probability that a non-negative random variable exceeds a certain value. - **Cauchy-Schwarz Inequality:** A fundamental inequality in probability and statistics.

2. **Bayes Risk:** - The minimum possible risk achievable by the best possible classifier (Bayes classifier).

Key Things to Remember: - Use Markov's and Cauchy-Schwarz inequalities to derive probability bounds. - Bayes risk serves as a benchmark for evaluating classifiers.

Sample Sheet 2: k-Nearest Neighbours and VC Dimension

Concepts:

1. **k-Nearest Neighbours (k-NN):** - A non-parametric method used for classification and regression. - The prediction is based on the majority class among the k nearest neighbours.

2. **VC Dimension:** - Measure of the capacity of a hypothesis class. - **Growth Function:** Number of distinct labelings of a set of points.

Key Things to Remember: - k-NN relies on distance metrics and the choice of k. - VC dimension helps in understanding the learning capacity of a model.

Sample Sheet 3: Growth Function and Graph Dimension

Concepts:

1. **Growth Function:** - The maximum number of distinct labelings of a set of points by a hypothesis class.

2. **Graph Dimension:** - A generalization of VC dimension for multiclass classification.

Key Things to Remember: - The growth function provides insight into the hypothesis class's capacity. - Graph dimension is useful for analyzing multiclass classifiers.

Sample Sheet 4: Agnostic PAC Learnability and Stability

Concepts:

1. **Agnostic PAC Learnability:** - Extends PAC learnability to scenarios where the best possible hypothesis may not achieve zero error.

2. **Stability:** - Measures the sensitivity of the learning algorithm to changes in the training set. - **On-Average-Replace-One Stability:** A specific form of stability used to bound generalization error.

Key Things to Remember: - Agnostic PAC learnability accounts for cases with inherent noise. - Stability is crucial for understanding the robustness of learning algorithms.

Sample Sheet 5: Rademacher Complexity and SVM

Concepts:

1. **Rademacher Complexity:** - A measure of the hypothesis class's capacity to fit random noise.

2. **SVM:** - **Hard SVM:** For separable data. - **Soft SVM:** For non-separable data with slack variables.

Key Things to Remember: - Rademacher complexity provides a way to quantify the capacity of a hypothesis class. - Understand the conditions under which hard and soft SVM are used.

Sample Sheet 6: k-means++ and Explainable k-means Cost

Concepts:

1. **k-means++:** - An initialization algorithm for k-means clustering that improves convergence.

2. **Explainable k-means Cost:** - The cost of clustering when the clusters must be explainable, often higher than the unrestricted k-means cost.

Key Things to Remember: - k-means++ helps in achieving better clustering results. - Explainable clustering may come at a higher cost due to additional constraints.