L2. PR & Assessment

Pattern Recognition and Classification:
- Explanation of pattern recognition and its application in machine learning.
- Key focus: teaching machines to recognize and classify patterns.

Basic Concepts in Pattern Recognition:
- Discussion of fundamental ideas in pattern recognition, including observations, judgments, measurements, classes, features, and decisions.
- Emphasis on the translation from human observations to machine classifications.

Terms and Concepts in Machine Learning:
- Sample: Explanation of what constitutes a sample in statistics and machine learning.
- Features: Quantitative attributes of a sample relevant to the learning task.
- Class: A subset of samples with shared properties.
- Classifier/Discriminant: A function that classifies samples into classes.
- Training/Learning: The process of teaching a machine using known samples.
- Supervised vs Unsupervised Learning: Different approaches based on the types of samples used.

Machine Learning vs. Pattern Recognition:
- Analysis of the relationship between machine learning and pattern recognition, emphasizing that PR is a type of task for ML.

Assessment of Classifiers:
- Discussion on how to evaluate the effectiveness of classifiers, focusing on error rates and precision.
- Introduction of concepts like Type-I and Type-II errors, sensitivity, specificity, precision, accuracy, F1 score, and Matthews correlation coefficient.

ROC Curves and Performance Assessment:
- Explanation of ROC curves and their use in assessing machine performance.
- Trade-offs in adjusting thresholds and the importance of the area under the curve (AUC).

Multi-Class Classifier Evaluation:
- Methods for assessing performance in multi-class classification scenarios, including accuracy, kappa score, and different averaging techniques (macro, micro, weighted).

Experiment Design for Assessing Classifiers:
- Techniques for estimating error rates, including training error, test error, and true error rate.
- Discussion of cross-validation methods like n-fold CV and LOOCV (Leave-One-Out Cross Validation).

Error Assessment & Decision Making in Real-World Situations:
- Case studies demonstrating the importance of considering various factors like prevalence, sensitivity, and specificity in real-world scenarios.

Inferring Relations based on Classification Performances:
- Utilization of pattern recognition for scientific discovery and inference, including the use of permutation tests.

L3. Linear Machines

Description: Linear regression is a basic scenario of machine learning where a linear relationship is modeled between a dependent variable (e.g., student's final score) and one or more independent variables (e.g., study hours per week).
Example: The document provides a toy data example illustrating this concept. It shows student IDs along with their final scores and the number of study hours per week, demonstrating how linear regression can be used to predict a student's score based on their study hours.

L4. Nonlinear & KNN

The Need for Nonlinear Classifiers:
- In cases where data is linearly non-separable, either stick to linear classifiers with minimized errors or turn to nonlinear classifiers.
- Nonlinearity does not always mean nonlinearly separable.

Quadratic Discriminant Analysis (QDA):
- Utilizes quadratic functions, which might be the simplest nonlinear functions.
- The decision for QDA is based on the largest value from the quadratic discriminant function.
- Suitable for large training datasets with roughly normal distribution and differing covariance matrices between classes.

Piecewise Linear Classifiers:
- Piecewise linearity is used to approximate any nonlinearity.
- Involves classification of multiple subclasses and decisions based on minimizing distances.

Nearest Neighbour (NN) Methods:
- NN method is a lazy learning approach where each sample can be considered a subclass.
- Decision making is based on the nearest neighbour or discriminant function.
- Various distance measures are used, including Minkowski, Euclidean, City-Block, and others.

k-Nearest Neighbours (k-NN) Method:
- A generalization of the 1-NN method, considering the k nearest neighbours.
- The decision is based on the maximum discriminant among the k nearest neighbours.

Asymptotic Errors of Nearest Neighbour Methods:
- Discusses the error rate of NN methods in comparison to the best possible error rates.
- The asymptotic error of k-NN is also considered, especially in relation to the Bayesian error.

Challenges and Solutions in Nearest Neighbour Methods:
- Addresses memory and computation costs, and the influence of noisy data.
- Introduces fast k-NN algorithms, editing nearest neighbour method, and the condensed nearest neighbour (CNN) method to address these challenges.

L5. ANN

History of Perceptrons and AI Schools: Discusses the early development of perceptrons, a type of artificial neuron, and the controversy in the AI community during the 1950s and 1960s.
Linear and Nonlinear Classifiers: Explains the limitations of linear classifiers and the need for nonlinear functions in certain scenarios.
Early Forms of Multiple-Layer Perceptrons: Describes the development of neural networks with multiple layers of perceptrons and the challenges in designing and training them.
MLP with BP Algorithm:
- MLP is the most popular type of artificial neural network, often using a sigmoid function for activation.
- The BP algorithm is central to training MLPs, involving the calculation of gradients to minimize error.

Kolmogorov Theorem and MLP's Representative Power: States that MLPs can approximate any continuous function, given sufficient hidden units, proper activation function, and weights.
Designing MLPs: Covers aspects like network structure, choice of activation function (e.g., sigmoid), and weight training algorithm (e.g., Gradient Descent).

MLP Applications:
- Used for classification, regression, and multi-dimensional mapping.
- Examples include splicing site prediction in genomes, protein structure prediction, and cancer classification with gene expression data.

Practical Techniques for Improving BP:
- Discusses various methods like choosing the right activation function, scaling input, selecting target values, initializing weights, and using augmented pseudo-samples.
- Covers the importance of the number of hidden units/layers, learning rates, momentum, weight decay, and stopping criteria to avoid overfitting.

General Structure of ANNs and Types: Details the structure of ANNs, including aspects like connection structure, activation functions, and learning algorithms. It also mentions major types of ANNs like Feedforward, Feedback, and Competitive Learning NNs.

L6. SVM

Large Margin and Optimal Hyperplane:
- Discusses the concept of an optimal hyperplane in SVMs, emphasizing the importance of maximizing the margin between classes for better classification.

Perceptron Recall:
- Reviews the perceptron as a precursor to understanding SVMs, focusing on its basic structure and learning algorithm.

Linearly Non-separable Cases:
- Addresses scenarios where data cannot be separated linearly, underlining the necessity for more complex models like SVMs.

Optimal Hyperplane Definition and Separation:
- Defines the optimal hyperplane in SVMs as the one that maximizes the margin between two classes without error, using linear decision functions.

Normalization and Support Vectors:
- Explains the canonical form of the separating hyperplane and introduces the concept of support vectors as key elements in SVMs.

Lagrangian and Saddle Point in SVMs:
- Discusses the saddle point of the Lagrangian function in SVM optimization, crucial for finding the optimal hyperplane.

Generalized Optimal Hyperplane for Non-separable Cases:
- Presents solutions for cases where data is not linearly separable, including the use of slack variables and the soft-margin optimal hyperplane.

Dual Problem of SVM and the Decision Function:
- Covers the dual formulation of the SVM problem and its importance in determining the decision function for classification.

Nonlinear Classification via Kernel Trick:
- Introduces the kernel trick for transforming data into higher-dimensional spaces, enabling SVMs to handle nonlinear classification.

Practical Aspects of SVMs:
- Discusses the challenges, algorithms, and optimization techniques related to SVMs, such as computational issues and convex optimization.

Applications and Comparisons of SVMs:
- Highlights the effectiveness of SVMs in various applications, comparing them with other methods like decision trees and MLPs.

Multi-Class Classification and SVMs:
- Explores the extension of SVMs to multi-class classification, including strategies like one-vs-all and multicategory SVMs.

SVM for Regression (Support Vector Regression):
- Describes how SVMs can be adapted for regression tasks, detailing the primal and dual problems in support vector regression.

L7. SLT

Generalization in Learning Machines:
- Addresses the issue of under-fitting and over-fitting.
- Focuses on the significance of generalization in learning processes.

Two Schools in Learning Process Analysis:
- Applied Analysis: Emphasizes the need to minimize training errors and the inductive principles for constructing learning models.
- Theoretical Analysis: Challenges the principles of applied analysis and seeks better generalization methods.

Learning Problem Setting:
- Outlines the basic model of supervised learning involving a generator, supervisor, and learning machine.
- The learning objective is to choose a function that best approximates the supervisor's response.

Risk Minimization:
- Defines the loss function and risk functional in learning.
- Goal: To find a function that minimizes the risk functional.

ERM (Empirical Risk Minimization) Inductive Principle:
- Explains the approach to approximate the function that minimizes expected risk by one that minimizes empirical risk.

Statistical Learning Theory Parts:
- Conditions for ERM process consistency.
- Rate of convergence and control of the generalization ability.
- Constructing algorithms for generalization control.

Consistency of ERM Learning Processes:
- Discusses the convergence of empirical risk to expected risk and the conditions necessary for ERM learning consistency.

VC Dimension and Generalization Bounds:
- Explores the VC dimension concept and its impact on the generalization ability of learning machines.
- Examines bounds on the generalization ability in relation to the VC dimension.

Controlling Generalization in High-Dimensional Spaces:
- Theoretical analysis of managing generalization in complex feature spaces.

Structural Risk Minimization (SRM):
- Combines ERM with regularization methods to balance empirical risk and model complexity.

Regularization Methods:
- Overview of various regularization techniques, including L0, L1 (Lasso), L2 (Tikhonov), and Elastic Net regularization.

L8. Features

Feature Engineering:

- Importance of feature acquisition (real-value and nominal features), feature selection, and feature extraction or transformation. Reasons for feature selection/extraction include dealing with noisy original features, computational considerations, and identifying key features for downstream tasks.

Criteria and Algorithms for Feature Selection/Extraction:

- Metrics to evaluate the "goodness" of features for classification.
- Different feature selection and extraction/transformation algorithms.

Metrics for Class-Separability of Features:

- Error rate, distance-based metrics, overlapping of distributions, entropy and mutual information, statistical tests on feature differences between classes.
- Average distance between classes, calculations using scatter matrices, Fisher's criterion.
- Other types of distance-based criteria (e.g., Bhattacharyya distance, Chernoff distance, divergence).

Searching for the Best Features:

- Challenges in finding the best combination of features from many candidates.
- Strategies for searching, including enumerating (optimal) and heuristic (sub-optimal) approaches.
- Branch-and-Bound Algorithm for organized searching.

Sub-optimal Searching:

- Sequential Forward Selection (SFS), Generalized SFS, Sequential Backward Selection (SBS), Generalized SBS, and L-R selection.

Genetic Algorithm (GA) for Feature Selection:

- A stochastic searching algorithm using the principles of evolution: inheritance, variation (mutation), and selection.
- Encoding the problem as a chromosome and finding the most fit "individual" from the "population".

Filtering, Wrapper, and Embedded Methods for Feature Selection:

- Filtering methods use stand-alone criteria or metrics for feature selection, followed by classification.
- Wrapper methods involve classification with all features, feature selection based on classification performance, and reclassification.
- Embedded methods integrate feature selection into the classification algorithm (e.g., Lasso).

Feature Extraction and K-L Transform:

- Fisher's Linear Discriminant (FLD) for feature extraction.
- Karhunen-Loève Transform (KLT) for function/vector expansion with orthogonal or orthonormal bases, leading to compressive representation of original data and minimal representation entropy.

Assessment of Feature-engineered Classifiers:

- Importance of cross-validation (CV) for reliable error estimation.
- Different schemes of CV and their impact on the evaluation of classifier performance.

L9. DT & Ensemble

Decision Tree (DT):

- DTs are used in various decision-making processes like credit risk assessment, medical diagnosis, and policy making.
- The process involves making tree-like decisions stepwise.
- A basic algorithm for building DTs includes finding the best splitting feature and creating child nodes accordingly.

ID3 Algorithm:

- Developed by Quinlan in 1979, it uses information entropy as a measure of impurity for each feature (node).
- The algorithm chooses nodes based on information gain, which is a decrease in impurity when expanding a node.
- It considers factors like entropy impurity, Gini impurity, and classification error.

Overfitting and Tree Pruning:

- Overfitting is a risk in DTs, particularly if the tree grows too deep.
- Methods to avoid overfitting include pruning (stopping growth in time or growing a full tree then pruning).
- Ockham's Razor principle is applied in machine learning to prefer simpler models over complex ones when the sample size is limited.

Random Forest (RF):

- RF involves decision-making with many decision trees to reduce the risk of a single decision-maker error.
- It uses bootstrapping (random sampling with replacement) to generate multiple trees.
- Each node of a tree randomly chooses a subset of features to base the decision on, and the RF classifier outputs the class that is the mode of the classes output by individual trees.

Ensemble Learning:

- Focuses on how to make classification machines better using different methods, including extracting better features, using more data, and boosting existing methods.
- Combining multiple classifiers to build a strong classifier from several weak classifiers.
- AdaBoost and other improved boosting methods like Gradient Boosting, GBDT (Gradient Boosting Decision Tree), XGBoost (eXtreme Gradient Boosting), and LightGBM (Light Gradient Boosting Machine) are discussed.

Pros and Cons of Boosting Methods:

- Good generalization properties and flexibility, but AdaBoost may not work well with too complex weak learners and is sensitive to noise.
- Gradient Boosting methods are introduced as improvements to address some of these challenges.

L10. Bayesian

Probabilistic View of Classification Task:

- Introduces classification tasks from a probabilistic perspective.
- Discusses decision-making under uncertainty, using prior and posterior probabilities for decision-making.

Bayesian Decision for Minimal Error:

- Explains how Bayesian decision-making aims to minimize error.
- Introduces the concept of minimizing error rate by considering prior and posterior probabilities.

Bayesian Classifiers for Normal Distributions:

- Demonstrates how Bayesian classifiers work with data that follows a normal (Gaussian) distribution.
- Covers discriminant analysis and decision boundaries in the context of Gaussian distributions.

Bayesian Decision with Gaussian Distribution:

- Discusses the general and specific cases of Bayesian decision-making with Gaussian distributions.
- Highlights the use of quadratic discriminants and minimal distance classifiers.

Bayesian Classifiers for Discrete Sequence Data:

- Applies Bayesian classifiers to genomic sequence analysis, particularly CpG islands in the human genome.
- Uses Markov Chains to model the probability of sequences in DNA.

Naïve Bayes Classifier:

- Introduces the Naïve Bayes Classifier and its application in high-dimensional feature spaces.

- Discusses the naïve assumption of conditional independence of elements.

Minimal Risk Bayes Classifiers:
- Focuses on minimizing risk in Bayesian decision-making.
- Discusses various aspects like sensitivity, specificity, and predictive value in the context of diagnostics.

Decision Problem and Loss Minimization:
- Presents the framework for decision-making in Bayesian classifiers, considering the loss associated with different decisions.
- Introduces the concept of minimizing risk and expected loss in decision-making.

Minimax Criterion and Unknown Priors:
- Deals with scenarios where prior probabilities are unknown.
- Discusses the minimax criterion as a strategy to minimize the maximum possible risk.

Two-step Bayesian Decision:
- Describes a two-step approach for Bayesian decision-making, especially when dealing with unknown priors and conditional densities.

Practical Applications and Exercises:
- Provides examples of Bayesian classifiers in action, such as predicting the survival of patients in ICUs using Naïve Bayes classifiers.
- Highlights the importance of understanding and applying Bayesian methods in real-world scenarios.

L11. Density Estimation

Introduction to Density Estimation:
- Emphasizes the importance of density estimation in Bayesian decision-making, highlighting that knowing the density is crucial but often only training data is available.

Parametric Estimation:
- Discusses estimating density functions from data using parametric methods.
- Focuses on samples that are independently and identically distributed (i.i.d.) from a density of a specific form, where only the parameters are unknown.
- Two main approaches: Maximum Likelihood Estimation (MLE) and Bayesian Estimation.

Maximum Likelihood Estimation (MLE):
- Involves estimating parameters from given samples.
- Uses the principle of maximizing the likelihood function, which is a product of the probabilities of observed samples.
- The solution is found by setting the derivative of the likelihood function to zero.
- For Gaussian distributions, MLE can estimate mean and variance.

Bayesian Estimation:
- Focuses on minimizing risk in parameter estimation.
- Involves using a loss function and expected risk.
- The Bayesian estimate is the conditional expectation of the parameter given the sample.
- This method calculates the a posteriori probabilities and then the expectation.

Non-parametric Estimation:
- Used when the form of the density is unknown.
- The aim is to estimate the density directly from samples.

Histograms:
- A simple non-parametric method for estimating density.
- Involves dividing data into bins and estimating the probability density based on the frequency of samples in each bin.

k-Nearest Neighbor Method:
- Another non-parametric method.
- Density estimation is based on the number of samples within a certain range of a point.

Parzen-Window Method (Kernel Density Estimation):
- Involves placing a "window" or kernel around each data point and averaging these to get the density estimate.
- Different types of kernels (like Gaussian or hypercube) can be used.

Bayesian Learning and Inference:
- Applies Bayesian methods to infer hypotheses from data.
- This involves estimating posterior probabilities of hypotheses based on given data and can be linked to maximum likelihood estimation.

L12. HMM & GM

PDF Estimation (concept):
- This would involve describing the estimation of probability density functions from given data.

Parametric Estimation (concept):
- This concept includes methods of estimating parameters within a certain assumed distribution from sample data.

Non-parametric Estimation (concept):
- Methods that do not assume a fixed distribution and estimate the underlying distribution directly from the data.

Maximum Likelihood Estimation (MLE) (method):
- A method of estimating the parameters of a statistical model, where the estimated parameters maximize the likelihood that the process described by the model produced the actual observed data.

Parzen-Window Estimation (method):
- A non-parametric technique for estimating the probability density function of a random variable.

Relations among Methods and Concepts in Machine Learning