Okay, here's a comprehensive summary of the key concepts from the PDFs, organized into a clear and complete study guide. I've included both English and Chinese explanations for key terms and concepts, as they appear in the original document.

Study Guide: Principles of Machine Learning - Supervised Learning: Classification

I. Introduction (绪论)

- **Machine Learning Taxonomy (机器学习分类):**
  - Supervised Learning (监督学习): Classification (分类), Regression (回归)
  - Unsupervised Learning (无监督学习): Density Estimation (密度估计), Structure Analysis (结构分析)

II. Classification Problem Formulation (分类问题)

- **Definition:** Classification, also known as decision or detection, involves building a model that produces a discrete label (标签) when given a set of predictors (预测因子).
  - *Example:* Sentiment analysis (情感分析) - determining if a movie review is positive or negative.
- **Key Components:**
  - **Predictors (预测因子):** Input features (输入特征) used to make predictions.
  - **Label (标签):** The discrete output value, also known as class (类别).
  - **Dataset (数据集):** A collection of samples (样本), each consisting of a predictor-label pair (预测因子-标签对).
  - **Model Quality (模型质量):** A measure of how well the model performs.

III. Representing Data and Classifiers (数据和分类器的表示)

- **Attribute Space (属性空间):** A space where each dimension represents a predictor variable. Labels can be represented on a vertical axis, but remember that ordering and distance might not apply to categorical variables.
- **Predictor Space (预测空间):** A more convenient representation where different symbols represent different labels within the space defined by the predictors.
- **Decision Regions (决策区域):** Regions in the predictor space associated with the same label.
- **Decision Boundaries (决策边界):** The boundaries separating decision regions.
- **Classifier Representation (分类器表示):** A partition of the predictor space into decision regions separated by decision boundaries.

IV. Linear Classifiers (线性分类器)

- **Definition:** Classifiers that use linear boundaries (线性边界) between decision regions.
- **Linear Equation (线性方程):** `w^T x = 0` defines a linear boundary, where `w` is the coefficients vector (系数向量) and `x` is the extended predictor vector (扩展预测向量) `[1, x1, x2...]`.
- **Classification Rule (分类规则):**
  - `w^T x > 0`: Sample belongs to one class.
  - `w^T x < 0`: Sample belongs to the other class.
  - `w^T x = 0`: Sample lies on the boundary.
- **Separable vs. Non-separable Cases (可分和不可分情况):**
  - **Linearly Separable (线性可分):** A linear classifier can achieve perfect accuracy (A=1, E=0).
  - **Non-linearly Separable (线性不可分):** Perfect accuracy is not possible (A<1, E>0).

V. Quality Metrics (质量指标)

- **Basic Metrics (基本指标):**
  - **Accuracy (准确率) (A):** `A = #correctly classified samples / #samples` (正确分类的样本数 / 总样本数)
  - **Error Rate (错误率) (E):** `E = #incorrectly classified samples / #samples = 1 - A` (错误分类的样本数 / 总样本数)
- **Limitations:** These metrics don't consider the cost of misclassifying different classes and treat all classes equally.

VI. Logistic Regression (逻辑回归)

- **Concept:** A method to train linear classifiers by modeling the classifier's certainty (确定性) using the logistic function (逻辑函数).
- **Logistic Function (逻辑函数):**
  `p(d) = ed / (1 + ed) = 1 / (1 + e-d)`
  - `p(0) = 0.5`
  - As `d → ∞, p(d) → 1`
  - As `d → -∞, p(d) → 0`
- **Distance to Boundary (到边界的距离):** `d = w^T x` represents the distance from a sample to the boundary.

- **Certainty (确定性):**
  - `p(xi)`: Classifier's certainty that `yi = o` is true.
  - `1 - p(xi)`: Classifier's certainty that `yi = ●` is true.
- **Likelihood Function (似然函数):**
  `L = ∏ (1 - p(xi)) ∏ p(xi)` for `yi=o` and `yi=●` respectively.
- **Log-likelihood (对数似然):**
  `l = Σ log[1 - p(xi)] + Σ log[p(xi)]` for `yi=o` and `yi=●` respectively.
- **Logistic Regression Classifier (逻辑回归分类器):** The linear classifier that maximizes `L` or `l`. It can be found using gradient descent (梯度下降).

VII. Nearest Neighbors (最近邻)

- **Concept:** A non-parametric (非参数) approach where new samples are assigned the label of the closest (most similar) training sample.
- **Instance-based Method (基于实例的方法):** The entire training dataset needs to be memorized.
- **Boundaries (边界):** Not defined explicitly but exist.
- **k-Nearest Neighbors (kNN) (k最近邻):**
  - Find the `K` closest samples (neighbors).
  - Assign the label of the most popular class among the neighbors.
  - As `K` increases, boundaries become less complex, moving from overfitting (过拟合) (small `K`) to underfitting (欠拟合) (large `K`).
  - `K` is usually odd in binary problems.

VIII. Beyond Accuracy: Class-Sensitive Metrics (超越准确率：类别敏感指标)

- **Misclassification Costs (误分类成本):** Different costs associated with misclassifying samples from different classes.
- **Bayesian Extension (贝叶斯扩展):**
  - Modify the decision rule to minimize expected cost rather than maximizing accuracy.
  - `Co × P(y = ● |x) ≤ 1` or `P(y = ● |x) / P(y = ○ |x) ≤ Co / C●`
  - `Co` and `C●` represent the cost of misclassifying a sample from class ● and class ○, respectively.
- **Confusion Matrix (混淆矩阵):** Shows the number or proportion of samples from each class that are assigned to each class.
- **Detection Problems (检测问题):**
  - **Positive (阳性):** Presence of a property.
  - **Negative (阴性):** Absence of a property.
  - **True Positive (TP) (真阳性):** Correctly classified positive.
  - **False Positive (FP) (假阳性):** Incorrectly classified as positive.
  - **True Negative (TN) (真阴性):** Correctly classified negative.
  - **False Negative (FN) (假阴性):** Incorrectly classified as negative.
- **Class-Sensitive Rates (类别敏感比率):**
  - **Sensitivity (灵敏度) / Recall (召回率) / True Positive Rate (TPR) (真阳性率):** `TP / (TP + FN)`
  - **Specificity (特异度) / True Negative Rate (TNR) (真阴性率):** `TN / (TN + FP)`
  - **Precision (精确度):** `TP / (TP + FP)`
  - **F1-score (F1分数):** `2 × (precision × recall) / (precision + recall)`
- **ROC Plane (ROC平面):**
  - **Receiver Operating Characteristic (ROC) curve (受试者工作特征曲线):** Plots sensitivity against 1-specificity.
  - **Calibration (校准):** Adjusting the threshold `T` to achieve a desired balance between sensitivity and specificity.
  - **Area Under the Curve (AUC) (曲线下面积):** A measure of goodness for a classifier that can be calibrated.

IX. Important Considerations (重要考虑因素)

- **Know your data (了解你的数据)!** Ensure data is representative (有代表性) of the population.
- **Flexibility and Complexity (灵活性和复杂性):** Choose a model with appropriate complexity to avoid overfitting and underfitting.
- **Generalization (泛化):** The ability of a model to perform well on unseen data.
- **The Best Metric (最佳指标):** Carefully choose the metric that captures your notion of quality, considering class-sensitive scenarios.

This study guide provides a solid foundation for understanding the principles of supervised classification in machine learning. Remember to review the examples and practice applying these concepts to different scenarios. Let me know if you have any other questions.