Comprehensive Study Guide for Advanced Machine Learning

Eshaan Arora

Contents

1 Overview and Recap

1.1 Data Science Lifecycle

- Iterative steps: Problem Definition $\rightarrow$ Data Collection $\rightarrow$ Preprocessing $\rightarrow$ Exploratory Data Analysis $\rightarrow$ Model Building $\rightarrow$ Evaluation $\rightarrow$ Deployment.
- Emphasis on feedback loops for continuous improvement.
- **Maximum Likelihood (ML) Principle:**

  $$\hat{\theta} = \arg\max_{\theta} L(\theta|X) = \arg\max_{\theta} \prod_{i=1}^{N} P(x_i|\theta),$$

  where $L(\theta|X)$ is the likelihood function for parameter θ given data X.
- **No Free Lunch Theorem:** No universally best model exists; trade-offs must be understood.

1.2 Predictive Analytics

- **Descriptive Analytics:** Summarizes past data to find patterns.
- **Predictive Analytics:** Uses models like regression and classification to predict unknown or future outcomes.
- **Prescriptive Analytics:** Suggests actions, often involving reinforcement learning.

2 Advanced Multivariate Regression

2.1 Basis Function Expansion

- Allows non-linear relationships by transforming input features, e.g.,

$$\phi_1(x) = x, \quad \phi_2(x) = x^2, \quad \phi_3(x) = \sin(x).$$

2.2 Regularization Techniques

- **Ridge Regression:**

$$\min_w \left\{ \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2 \right\}.$$

- **Lasso Regression:**

$$\min_w \left\{ \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \|w\|_1 \right\}.$$

- **Elastic Net Regularization:** Balances Ridge and Lasso penalties for feature selection and coefficient shrinkage.

2.3 Bias-Variance Tradeoff

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}.$$

- High bias indicates underfitting; high variance indicates overfitting.
- Learning curves help identify bias-variance issues.

2.4 Polynomial Regression

- Fits non-linear relationships using polynomial terms like x, x^2, x^3.

3 Neural Networks and Optimization

3.1 Activation Functions

- **ReLU:** Efficient but may suffer from the dying ReLU problem.
- **Leaky ReLU:** Allows a small slope for negative inputs to address dying ReLUs.
- **Sigmoid and Tanh:** Useful for smooth outputs but suffer from vanishing gradients.

3.2 Stochastic Gradient Descent (SGD)

- Iteratively minimizes loss functions.
- Weight update rule:

$$w_{t+1} = w_t - \eta \frac{\partial L(w_t)}{\partial w},$$

  where η is the learning rate.
- Momentum-based SGD accelerates convergence by incorporating past updates.
- **Applications of SGD:**
  - Large datasets where batch gradient descent is computationally expensive.
  - Streaming data or non-stationary problems.

3.3 Backpropagation

- Efficiently computes gradients for multi-layer networks using the chain rule.
- Steps:
  1. Forward pass to compute predictions.
  2. Compute the error at the output.
  3. Propagate the error backward to update weights.

3.4 Multi-Layer Perceptrons (MLPs)

- Universal approximators capable of modeling continuous functions.
- Use non-linear activations in hidden layers for expressiveness.
- Softmax function ensures output probabilities sum to 1 in classification tasks.

3.5 Transfer and Multi-task Learning

- **Transfer Learning:** Sequentially reuses pre-trained models for related tasks.
- **Multi-task Learning:** Simultaneously learns related tasks by sharing features.

4 Loss Functions and Regularization

4.1 Huber Loss

Combines MSE for small residuals and MAE for large residuals:

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & \text{if } |r| \leq \delta, \\ \delta(|r| - \frac{\delta}{2}), & \text{if } |r| > \delta. \end{cases}$$

4.2 Regularization in Ridge Regression

- Increasing regularization (λ) reduces variance but increases bias.
- At $\lambda \to \infty$, weights shrink to zero except the intercept, leading to MSE = Var(y).

5 Dimensionality Reduction and PCA

5.1 Principal Component Analysis (PCA)

- Identifies principal components that capture maximum variance.
- **Reconstruction:** Original data can be reconstructed as:

$$\mathbf{X} \approx \mathbf{Z}\mathbf{V}^{\top},$$

  where $\mathbf{Z}$ are PCA scores and $\mathbf{V}$ are eigenvectors.

5.2 t-SNE and Visualization

- t-SNE maps high-dimensional data into 2D or 3D spaces, preserving local structure for visualization.

6 Probabilistic Models and Bayesian Networks

- **Bayesian Belief Networks:** Models conditional dependencies.
- Probabilities are calculated using conditional probability tables (CPTs).
- Example:

$$P(C, S, \neg R | \neg W) = \frac{P(C)P(S|C)P(\neg R|C)P(\neg W|S, \neg R)}{P(\neg W)}.$$

7 Classifier Calibration and Decision Theory

- **Calibration Curves:** Evaluate how predicted probabilities align with observed outcomes.
- **Decision Boundaries:** Logistic regression defines boundaries as hyperplanes, e.g.,

$$\mathbf{w}^\top \mathbf{x} = \log\left(\frac{P(C_1)}{1 - P(C_1)}\right).$$

8 Ensemble Methods and Transformers

8.1 Bagging and Boosting

- **Bagging:** Combines models to reduce variance.
- **Boosting:** Sequentially focuses on errors to improve bias.

8.2 Transformers

- Includes attention mechanisms for sequential data:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

9 Data Pre-Processing and Metrics

9.1 Data Transformations

- Scaling, normalization, and logarithmic transformations improve model performance.

9.2 Imputation and Handling Outliers

- Imputation strategies include mean, median, mode, and k-NN.
- Outliers are handled via thresholds, distance-based methods, or statistical rules.

9.3 Performance Metrics

- Metrics include Accuracy, Precision, Recall, F1-Score, ROC-AUC.