Applied Statistics for Machine Learning Exam

Statistics and Machine Learning (9 – 12%)

Relevance of Statistics in Big Data and Machine Learning

- Describe ways of obtaining data, the different types of data, and how each type of data is analyzed
- Define big data and identify smart applications produced by it
- Compare and contrast of machine learning and classical statistics
- Explain the importance of statistics in machine learning

Terminology and Vocabulary

- Relate statistical terminology with machine learning
- Compare variable types and level of measurements
- Explore common modeling vocabulary

Fundamental Statistical Concepts (17 – 21%)

Basics of Statistical Analysis

- Distinguish between populations and samples
- Describe the process of statistical analysis
- Compare and contrast inferential and descriptive statistics
- Explain different methods of sampling data including event-based sampling

Descriptive Statistics

- Define measures of central tendency, position, and dispersion
- Explain how to visualize a distribution with different graphics
- Describe usefulness of the normal distribution in machine learning
- Explain measures of distribution shape

Inferential Statistics

- Explain sampling distributions and how to make inferences from data
- Explain confidence intervals and hypothesis tests
- Define a one-sample t Test
- Describe usefulness of p-values in machine learning

Explanatory Modeling Using Linear Regression (18 – 24%)

Correlation and Simple Linear Regression

- Define explanatory modeling
- Explore bivariate relationships using scatterplots
- Compare and contrast correlation and covariance
- Identify irrelevant and redundant predictors using correlation
- Explain simple linear regression and OLS estimation
- Test regression hypothesis and assess model fit

Multiple Regression and Model Selection

- Define multiple linear regression
- Use categorical predictors
- Define ANOVA and relate it with regression
- Explain interaction effects
- Compare regression models using R-square, Adjusted R-square, and Information Criteria
- Describe sequential model selection methods

Model Diagnostics

- Define the assumptions of linear regression
- Verify assumptions with Residual Plots
- Diagnose and remedy collinearity
- Explain problems with outliers, leverage points, and influential observations
- Diagnose influential and outlier cases

Predictive Modeling Using Logistic Regression (25 – 31%)

Introduction to Predictive Modeling

- Compare and contrast explanatory and predictive modeling
- Describe predictive modeling concepts
- Define honest assessment including data partitioning
- Explain how to incorporate different time frames for predictive modeling
- Explain how to optimize model complexity for prediction
- Explain the bias-variance tradeoff

Categorical Associations

- Explain association between categorical predictors
- Define and use Cramer's V statistic
- Explain and interpret odds ratios

Logistic Regression Model

- Define logistic regression
- Define odds and log odds
- Describe maximum likelihood estimation
- Interpret logistic regression coefficients
- Assess logistic regression model fit
- Use categorical predictors
- Explain interaction effects
- Compare logistic regression models using concordant/discordant pairs, c-statistic, and Information Criteria
- Describe sequential model selection methods

Model Deployment

- Explain how to deploy a logistic regression model
- Describe scoring a logistic regression model
- Explain the classification cutoff for scoring

Statistical Foundations of Machine Learning (18 – 24%)

Overview of Machine Learning

- Define machine learning
- Define supervised, unsupervised, semi-supervised, and reinforcement learning
- Explain neural networks
- Name common algorithms in machine learning
- Distinguish between data preparation and data preprocessing

Data Pre-processing for Machine Learning Models

- Describe common difficulties with modeling data for machine learning
- Describe challenges in visualizing big data
- Diagnose and correct problems with errors, missing values, and outliers
- Explain why transform input variables and discuss some simple transformations
- Diagnose problems with high dimensional data and feature engineering remedy
- Discuss feature scaling

Model Evaluation, Estimation, and Post-training Tasks

- Explain signal-noise dynamics
- Define cross-validation and bootstrap aggregation
- Explain coefficient shrinkage and why it can be useful
- Define L1, L2 and L12 regularizations
- Explain learning process and estimation criteria in machine learning
- Differentiate between parameters and hyperparameters
- Explain model interpretability

Note: All 15 main objectives will be tested on every exam. The expanded objectives are provided for additional explanation and define the entire domain that could be tested.