INFSCI 0419: Final Exam, Fall 2019

Study Guide

1. Measurement scales: nominal, ordinal, interval, ratio
2. Measures of central tendency, how they are affected by different distributions and by outliers
3. Missing data
    a. Data missing completely at random (MCAR)
    b. Data missing at random (MAR)
    c. Data missing not at random (MNAR)
4. Approaches to dealing with missing values
    a. Converting the missing values into a new value
    b. Creating new categories (i.e. “MISSING”)
    c. Discarding rows / columns with missing values
    d. Discarding columns with missing values
    e. Replace missing values with mean, median, mode, or predicted / estimated value
5. Outliers, what they are, how they occur, how they affect distributions, biases, and ML algorithm selection
6. Scaling, centering, and transforming data
7. Simple linear regression
    a. How OLS algorithm works
    b. How to interpret the model (i.e. what do the coefficients indicate)
    c. How to evaluate the model (R-squared vs. adjusted R-squared vs. RMSE)
    d. Residuals
    e. Statistical assumptions
8. Supervised vs. unsupervised machine learning
9. Bias / variance trade-off in ML
10. Probabilities
11. Naive Bayes, SVM, kNN, Decision Trees, Random Forest
    a. How each algorithm works
    b. What are the key parameters
    c. When to use
    d. Weaknesses
    e. Strengths
12. Evaluating classifiers
    a. Confusion matrix
        i. True positive, false positive, true negative, false negative
    b. Accuracy score
    c. Cross-validated accuracy score

    d. ROC / AUC
    e. Sensitivity vs. Specificity
13. Model validation
    a. Cross-validation
    b. Holdout method
    c. Leave-one-out
    d. Stratified cross-validation
14. Dimensionality reduction
    a. Collinearity
    b. Features with low variance
    c. Univariate feature selection
    d. Backward feature elimination
    e. Forward feature selection
    f. Random forest feature scores
15. k-Means Clustering
    a. Algorithm
    b. Stop conditions
    c. Evaluating using the elbow method
16. ML Algorithm Selection
    a. Bias/variance trade-off
    b. Size of training set
    c. The accuracy of the model
    d. The interpretability of the model
    e. The complexity of the model
    f. The scalability of the model
    g. How long does it take to build, train, and test the model?
    h. How long does it take to make predictions using the model?
    i. Linearity
    j. Number of parameters
    k. Number of features