CS 105 Study guide for Midterm 2

Topics:

1. Mathematical Models for Data Science

    - Types of models
    - Supervised, unsupervised, reinforcement learning
    - Assessing performance of a learning algorithm
    - Common splitting strategies
    - Confusion matrix
    - Bias, Variance, Bias - Variance Trade-Off
    - Loss functions
    - The curse of dimensionality
    - Distance measures

2. Supervised Learning: Regression Analysis and Classification

    - Regression
    - Linear regression
    - Line of best fit
    - KNN
    - Weighted KNN
    - Normalization and standardization
    - KNN time complexity and improvement (KD - trees)
    - Advantages and disadvantages of linear regression and KNN
    - Recommender Systems

3. Unsupervised Learning: Clustering

    - Clustering, types of clustering
    - K-means clustering
    - Initialization and choice of k in K-means clustering
    - Strengths and weaknesses of K-means clustering
    - Hierarchical clustering
    - Types of hierarchical clustering

4. Intro to Text Mining

    - Document decomposition
    - Text features
    - Bag-of-Words Model
    - One-Hot-Encoding Model
    - TF-IDF

Some tests, formulas, algorithms, techniques to perform (for examples, see problems from slides):

- Understand which algorithm/technique can be used on a given dataset to perform a given task
- Compute precision and recall from the a confusion matrix
- Calculate MSE and MAE
- Find an equation of the line of best fit
- Normalize (standardize) given data
- Run KNN on a small dataset
- Build a recommender system
- Run k-means clustering on a small dataset
- Find k for k-means clustering
- Run hierarchical clustering on a small dataset
- Vectorize a corpus using the Bag-of-Words Model
- Vectorize a corpus using the One-Hot-Encoding Model
- Calculate distance (similarity) using a given distance measure