Study Checklist for 50.038 Computational Data Science

> You can click on ? for relevant links on topics more in-depth outside of the lecture notes.

- Week 1: BIG DATA, Hadoop and MapReduce
  - □ The 3 V's of Big Data
  - □ CAP Theorem
  - □ Hadoop Ecosystem
  - □ What is MapReduce
- Week 2: Feature vectors, dimension reduction, evaluation
  - □ Types of Features
    - □ Ordinal, Nominal, Interval, Ratio
    - □ Discrete, Continuous
  - □ Discretization, Binarization
  - □ Curse of Dimensionality
  - □ PCA, SVD
  - □ Token Normalization
  - □ TF-IDF
- Week 3: Data Visualization
  - □ How to emphasize certain data
  - □ Reading Charts e.g. Boxplot, Scatterplot, Spider Charts, Violin Plots etc...
- Week 4: Regression algorithms – Time series
  - □ Train, validation, test sets
  - □ Types of cross-validation
    - □ leave-one-out
    - □ k-fold
  - □ Random Sampling, Stratified Sampling
  - □ How to measure the quality of a classifier
    - □ Accuracy
    - □ Precision
    - □ Recall
    - □ F-score
    - □ Receiver Operating Curve (ROC)
- Week 5: Classification algorithms
  - □ Decision Tree
    - □ Measuring Node Impurity
      - □ GINI index
      - □ Entropy
      - □ Misclassification Error
    - □ Addressing Overfit and Underfit
  - □ K-means
    - □ Elbow Method for determining optimal K
  - □ Ensemble Methods ?
    - □ Bagging

        - □ Boosting
        - □ Stacking
- Week 6: Intro to Deep Learning
    - □ Activation Functions: Sigmoid, Softmax, tanh, RELU, leaky RELU
        - □ How Softmax is an extension of Sigmoid
        - □ Why leaky RELU when there is RELU
    - □ Neural Networks
        - □ Backpropagation
        - □ Gradient Descent
        - □ Underfit, Overfit - When to stop training
- Week 9: Word Embeddings ?
    - □ One-hot vectors vs Bag of Words (BOW) vs Word Embeddings
    - □ Text Representation Models
        - □ Word2Vec: cBOW vs Skip-gram
        - □ Doc2Vec: dBOW vs dM ?
    - □ Extensions of Word Embeddings: GloVe, Elmo, BERT
- Week 10: Convolutional neural networks (CNN) ?
    - □ Image Detection/ Filter Kernel
        - □ Dimensions of output in relation to Stride size
        - □ Padding with zeros
    - □ Activation Maps
    - □ Max Pooling and Average Pooling
    - □ Flattening
    - □ Methods of Data Augmentation
- Week 11: Recurrent Neural Networks (RNN)
    - □ Types of RNN: Many-to-many, one-to-many
    - □ Problem of Vanilla RNNs: Vanishing/Exploding Gradient ?
        - □ Resolution: Long Short-term Memory (LSTM) ?
        - □ LSTM Variants: Peephole, Combined Forget/Input Gates, Gated Recurrent Units