Final Exam Review

Week	THURS. Week 10
Assignment Due	HW5 Due — (Clustering)
Assignment Done	☐
Due Date	@March 17, 2023 11:59 PM
Notes Done	☑

Final Exam Notes

This course covered:

- Curse of dimensionality
  - More dimensions = less coverage of feature space
  - Looked at techniques that would reduce dimensionality
- Supervised datasets
  - SMAPE
  - RMSE
  - Accuracy
  - Confusion matrix
  - Precision
  - Recall
  - F-measure
  - Trade-off between precision/recall
- Linear discriminate analysis
  - Projects data to K dimensions, where K is the number of classes, to do classification

    - Similar to PCA, but instead of maximizing variance of data after projection, we want to maximize distance of the means of class data
- KNN
    - Straight-forward, expensive algorithm b/c you need to carry data around with you
- Statistical approaches
    - Joint distributions (if available)
    - Bayes' rule
    - Naïve Bayes rule
- Decision trees
    - Intuitive in how to build one, starting w/ node that gives best split of data so class entropy is lowest
- Logistic regression
    - Being able to do binary classification via a probability with a weighted sum of the features and processing it through a logistic activation function
        - Gives value between 0 and 1, which we interpret as probability of class 1
        - To get weights, we looked at gradient based learning
- SVMs
    - Linear approach where we're trying to do binary classification
    - Want hyperparameters that separates data by a hyperplane
- Midterm
- Linear regression
    - Computes target value as weighted sum of features plus a bias, finding them through a gradient based approach or a direct solution
- Ensembles
    - Idea of there "no free lunch theorem" → no *one* algorithm is the best
        - Maybe we want several algorithm to work together

    - Talked about voting schemes, bagging
        - Bagging → for each system we grab random set of training observations with replacement
        - Boosting → idea that after training a sytem with some bag of observations, we can see in our training data what we got wrong/right and focus new training to focus on what we got wrong
        - Random forest → trees are build s.t. any time we have to choose a feature to split on, we choose the best of *random options, adding a bit of randomness*
- Clustering
    - K-means clustering
    - Mixture models
        - Talked about how mixture models are a generalization of k-means
        - Require you to know clusters k ahead of time
    - Agglomerative clustering tree
        - Possibly can do
- ANN
    - Extension of linear/logistic regression
        - Have idea of a weight matrix and bias offset as with linear/logistic regression
            - Put this into an activation function
            - Could take output optimally and have multiple outputs and weight this with a second weight matrix, add more bias, and take this new weighted sum output into another activation function until we get to a place that we'd like

Format

Similar to midterm, but longer

1. Short responses (a word or sentence or a plot/drawing)

2. Computations
3. Derivations (given an objective function, computing a gradient and/or closed form solution