Stat 435
Arnav Palkhiwala

1. Summary:

Over the last 2 weeks, we got a good understanding of unsupervised learning and deep learning, both areas I would love to dive more deeply into over the next year. In unsupervised learning, we learned about methods such as PCA and clustering in order to understand and essentially sort through unlabeled data. We then went on to learn different methods in clustering and how they differ and work. Lastly, we discussed deep learning and the basics of neural networks and its bias variance tradeoff.

2. Concepts:

Goal of Unsupervised Learning:

- Discovering important "dimensions" of data
- 2 Methods
    - PCA - performing dimension reductions
    - Clustering - grouping observations and creating subgroups of data
- Challenges - there is no clear goal or way to assess results
- Benefits - it is much easier to obtain because it is unlabeled

Principal Component Analysis PCA:

- Lower dimension representation before using trees
- Reducing number of features
- Single "new" feature that summarized the data
    - This is a linear combination of features that lead to the highest variance
- Goal: Minimize Euclidean distance between original points and their projection
- You want to maximize variance through this because it best represents the variance of the original data
- Find lower dimensional representation of raw data
- Basic Equation: $Z = XP$
    - Z is nxk and p is dxk
    - Each col is principal component
    - K is much smaller than d
- How to choose p:
    - No correlation among z features (little redundancy)
    - Rank ordered features by variance
- Choosing k:
    - Pick top 2, 3 dimensions to plot

Clustering:

- Partition unlabeled data into similar groups
- Evaluation of clustering quality: calculate the within cluster variation
- K means clustering uses means
- Hierarchical cluster builds cluster in bottom up fashion
    - Ecah pt own cluster
    - Identify 2 closest clusters, merge
    - Repeat
    - End when all points in a cluster
- Practical issues
    - Scaling
    - Dissimilarity measure
    - How many clusters to choose
- Can do unsupervised learning as preprocessing for supervised learning

Neural Networks:

- XOR problem not solvable by logistic regression (same sign one class rest other)
- NN's can approximate complicated functions
- Common choices for activation functions: RELU, Sigmoid
- Double descent graphically

3. Uncertainties:

My main uncertainties are around using unsupervised learning as a preprocessing method which we did in class. If we do PCA on the data to reduce the number of dimensions of the data, but then doesn't the response variable change? Then how do we maintain the original basis of the problem and the main concept? I am also curious about Neural Networks and convulsion NNs. I want to look into it a lot more and get a better understanding of them, so I may take additional statistics courses in this realm if they are available. I am also curious about when to use Neural Networks. I heard that it's better when we have more data but in general should we test both traditional methods and nns or is there some sign that tells sus we should just only use nns and not waste time/compute on other methods>