High Dimensional Data

An Investigation into General EDA when working with multi-dimensional datasets

* If you are looking at this at a later date, comments are in the speaker notes below

How many dimensions count as high dimensional data?

What this presentation is about?

1. General approach to tackling high dimensional data
2. How to select a subset of features
3. How to make pretty visualizations of clusters using dimensionality reduction algorithms

1. Questions to Ask When Doing EDA

Questions:

- What constitutes an outlier?
- Do I really need all of my features?
- Are there latent variables at play?
- Does missing data really matter (ignore vs. impute)?
- Are my classes unbalanced which distorts the distribution of covariates?

Less obvious questions:

- What plots really tell the story of the data?
- Are there any biases in data sampling?

1. Quickly Finding Dependencies And Problems

- Heatmaps?
- Lineplots?
- Pandas Profiler: https://github.com/pandas-profiling/pandas-profiling

2. Feature Selection Algorithms

Given P features, rank them by importance then select K of them.

Popular Strategies:

- statistical tests
- forward/recursive feature elimination
- model based importance

2. Feature Selection Algorithms

Questions

- Does feature selection introduce bias?

- Does your selection strategy deal with multicollinearity?
- Are there any underlying assumptions in your selection strategy?

Ideas

- Perfected correlated variables are truly redundant.
- Two variables that are useless by themselves can be useful together.

Source: http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf

2. Questions to Ask During Feature Selection

More obvious questions:

- What constitutes an outlier?
- Do I really need all of my features?
- Are there latent variables at play?
- Does missing data really matter (ignore vs. impute)?

Less obvious questions:

- What plots really tell the story of the data?
- Are there any biases in data sampling?

3. Dimensionality Reduction Methods

Find a mapping, $f(x) \rightarrow y$, where x is dim N and y is dim M, N >> M.

Given some distance metric, D, these functions aim to have $D(x_i, x_j) \sim D(y_i, y_j)$.

Popular Examples:

- t-distributed stochastic neighbor embedding (T-SNE)
- principal components analysis (PCA)
- spectral embedding

Additional Resources

Feature Selection Paper: http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf

Dimensionality Reduction Algorithms on MNIST: https://colah.github.io/posts/2014-10-Visualizing-MNIST/

Google Tensorflow High Dimensional Visualization: https://experiments.withgoogle.com/visualizing-high-dimensional-space