Meeting Summary for the Live Professor Q&A on Supervised and Unsupervised Learning at the 2025 Astrostatistics Summer School (06/03/2025)

1 Quick recap

The meeting covered several machine learning and data analysis techniques, including t-SNE for dimensionality reduction, K-means clustering with Expectation-Maximization, and PCA for variance maximization. Ashley explained various clustering algorithms and their applications, including DBSCAN, TV scan, and the challenges of choosing appropriate kernel functions for SVMs. The discussion concluded with guidance on approaching high-dimensional data analysis, emphasizing the importance of starting with linear models like PCA before moving to more complex methods.

2 Summary

2.1 T-Sne and K-Means Explained

Ashley explained the concept of t-SNE, a dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space while preserving neighborhood relationships. She clarified why t-SNE uses a probability-based distance metric rather than direct pairwise distances, explaining that it allows for comparison of distributions using tools like KL divergence. Ashley also described the Expectation-Maximization algorithm used in K-means clustering, which alternates between assigning data points to clusters based on proximity to centroids and updating the centroids based on the assigned points.

2.2 PCA Power Iteration Method Overview

Ashley explained the power iteration method for PCA by describing how it finds the largest variance by iteratively expanding vectors in the eigenbasis, with the largest eigenvalue component growing exponentially while others diminish. She acknowledged the complexity of the explanation and agreed to share a visual map via Slack to better illustrate the concept.

2.3 Maximizing SVM Width and Clustering

Ashley discussed the challenges of maximizing the width of the screen in SVMs and the relaxation of conditions for classification when data is not easily separable. She explained the use of PCA for dimensionality reduction in light curve data, noting that while PCA can capture most variations, it may not preserve the physical structure of rare transients, potentially losing important signals. Ashley also addressed the difficulty of choosing the correct kernel function for SVMs, suggesting that standard options like Gaussian kernels are often used due to their simplicity. Jay inquired about the DBSCAN algorithm, which Ashley described as an algorithm that does not follow the simple model framework she typically teaches, but instead directly identifies clusters in data without preselecting the number of clusters.

2.4 Clustering Algorithm Demonstration

Ashley explained a clustering algorithm that starts with random points and expands clusters by checking the number of neighboring points within a defined neighborhood. She demonstrated how the algorithm identifies core points that meet a minimum threshold and labels them as cluster centers, while other points are added to existing clusters if they meet the neighborhood criteria. The process continues until all points are visited, with new clusters forming when unvisited points are encountered.

2.5 TV Scan Algorithm Overview

Ashley explained the TV scan algorithm, which automatically labels outliers without assuming cluster shapes or sizes, requiring only neighborhood size and minimum points as parameters. She advised starting with linear models, particularly PCA analysis, when working with high-dimensional data, and suggested testing whether PCA naturally forms reasonable clusters. When asked about differences between TV scan and friend-of-friend algorithms, Ashley admitted she was unfamiliar with the latter.