M462/562 Final Project

You may complete the project as an individual or with a partner.

Your project will include three deliverables:

1. A project proposal.
2. A presentation of your project to the class during the final exams week.
3. A final project report in the format of Jupyter Notebook.

Project Proposal (due Friday, Feb. 19): Your proposal should include the following information

- Project title (i.e., the algorithm you have chosen from the list of project ideas).
- Group members

Project Presentation (during the scheduled final exams week: April 26—30).

Your project presentation should be 15 minutes long with 5 minutes for questions.

Project Report (due Friday, April 23)

Your final project report must be in the format of a Jupyter Notebook. Your report should sufficiently describe your project, including:

- An introduction, describing the problem you are solving.
- The algorithm implementation.
- Examples

Project Ideas:

Passive-Aggressive algorithms: The passive-aggressive algorithms are a family of online classification algorithms for massive datasets.

In online machine learning algorithms, the input data comes in sequential order, and the model is updated step-by-step. This is very useful in situations where there is a huge amount of data and it is computationally infeasible to use the entire dataset.

Decision Trees: Decision Trees are versatile Machine Learning algorithms that can perform both classification and regression tasks. They are powerful algorithms, capable of fitting

complex dataset. They are also the fundamental components of Random Forests, which are among the most powerful Machine Learning algorithms available today.

Naïve Bayes classifiers and spam detection: Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes theorem. These classifiers are very popular for text classification tasks.

Graphs/Networks and graph centrality: Graphs are mathematical structures used to study pairwise relationships between objects. They can be used to model, for example, communication and transportation networks, social networks, the web, supply chain networks, etc.

Graph centrality identifies the most important vertices within a graph. Applications include identifying the most influential person in a social network, key infrastructure nodes in the internet or urban networks, super-spreaders of a disease, etc.

Anomaly detection and the local outlier factor (LOF) algorithm: Anomaly detection (aka outlier detection) is the identification of rare observations that raise suspicions by differing significantly from the majority of the data. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text.

The local outlier factor is an algorithm proposed in 2000 for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbors.

Clustering algorithms: Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (cluster).

The goal is to implement one of the following clustering algorithms:

1. **Gaussian mixture models** (a distribution-based clustering algorithm).
2. **The DBSCAN algorithm** (a density-based clustering algorithm).
3. **Spectral clustering** (a similarity-based algorithm)

Dimensionality reduction algorithms: Dimensionality reduction is the transformation of data from high-dimensional space into a low-dimensional space so that the low-dimensional

representation retains some meaningful properties of the original data. Dimensionality reduction is extremely useful for data visualization. Reducing the number of dimensions down to two (or three) makes it possible to plot a view of a high dimensional set.

The goal is to implement one of the following dimensionality reduction algorithms:

1. **Principal component analysis (PCA):** This algorithm first identifies the hyperplane that lies closest to the data, and then it projects the data onto it.
2. **Linear discriminant analysis (LDA):** This algorithm finds the linear combination of features that characterizes or separates two or more classes of objects.

More advanced algorithms:

3. **T-distributed stochastic neighbor embedding (t-SNE)**
4. **Uniform manifold approximation and projection (UMAP)**

Both t-SNE and UMAP reduce dimensionality while trying to keep similar instances close and dissimilar instances apart. These algorithms are extremely useful for visualizing high-dimensional data.