Honor Code

Each student should solve problems in this assignment independently. Specifically, *code sharing in any form is strictly prohibited*. However, you are permitted to discuss aspects of code syntax/theory/algorithms/implementation with your peers. You can list names of your colleagues (below) with whom you collaborated in solving this assignment.
Collaborators

1.
2.
3.
4.

Students found plagiarising will lose all credit for this assignment.

Submission Instructions

Submit the code for the assignment in .zip file under the name 'firstname_lastname.zip'. The zip file must contain the code for all the problems along with all the necessary files to run the code. Also, include in the .zip file, a .pdf file containing all results, comments and explanations requested in the assignment.

1 Dynamic Mode Decomposition

Consider a simple pendulum with length $l = 1\ m$, $g = 9.81\ m/s^2$ and state $\mathbf{x} = \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix}$.

1.1 Linear dynamics

Consider linear dynamics of simple pendulum and initial state $\theta(0) = 0.1$ rad , $\dot{\theta}(0) = 0$ rad/s.

- Generate the state vectors $\mathbf{x}$ with a sampling time of 0.2s over the interval $t \in$ [0,20] using ode45.
- Construct the snapshot matrices $\mathbf{X_1} = \left[\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_{n-1}\right]$, $\mathbf{X_2} = \left[\mathbf{x}_2, \mathbf{x}_3 ... \mathbf{x}_n\right]$.
- Compute the linear evolution operator $\mathbf{A} = \mathbf{X_2}\mathbf{X_1}^\dagger$ characterizing the system's dynamics. How does this matrix $\mathbf{A}$ compare to the original linear dynamics matrix?
- Plot and compare the solution the using DMD with original dynamics for the interval $t \in$ [0,25].

1.2 Non linear dynamics

Consider the full dynamics of the pendulum and initial amplitude $\theta(0) = 1$ rad, $\dot{\theta}(0) = 0$ rad/s.

- Generate the state vectors $\mathbf{x}$ with a a sampling time of 0.2s over the interval $t \in$ [0,20] using ode45.
- Construct the snapshot matrices $\mathbf{X_1} = \left[\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_{n-1}\right]$, $\mathbf{X_2} = \left[\mathbf{x}_2, \mathbf{x}_3 ... \mathbf{x}_n\right]$.
- Compute the linear evolution operator $\mathbf{A} = \mathbf{X_2}\mathbf{X_1}^\dagger$ characterizing the system's dynamics.
- Plot and compare the solution the using DMD with original dynamics for the interval $t \in$ [0,25] and compute RMSE.
- Comment on the solution using DMD approximation for the non-linear dynamics with linearized model approximation for same initial conditions.

2 Regression

2.1 Polynomial regression

- Consider the dataset contained in $regression1.mat$. Split the data into training and test sets, with an 80-20 distribution.
- Perform 2nd degree polynomial regression and determine the coefficients of fit.
- Using the resulting expression, predict the values for the test dataset and calculate mean squared errors(MSE for train and test datasets).
- Comment on the fit quality.(Overfitting/Underfitting etc.,)

2.2 Regularization

2.2.1 Lasso and Ridge regularization

- Consider the dataset contained in $regression2.mat$. Split the data into training and test sets, with an 80-20 division.
- Perform a 10th-degree polynomial regression without regularization on the training data. Predict the test data values, compute the MSEs, and assess the fit for indications of overfitting or underfitting. Reason.
- Perform Lasso and Ridge regularization to the above model using regularization coefficient $\alpha = 0.1$ and recompute the training MSE and test MSE for each. Comment on the effects of regularization by explaining your observations.
- Plot the fits of all three models (unregularized, Lasso-regularized, and Ridge-regularized).

2.2.2 Pareto frontier for best α

This exercise (2.2.2) will not be graded. You can still submit your solutions if you are interested.

- Plot the Pareto frontier (training MSE vs test MSE) for Lasso and Ridge for different α and identify the optimal α value for both the methods. (*Hint:* Selecting α initially in log space and narrow it down to get the best value.)
- Plot the fits of all three models (unregularized, Lasso-regularized, and Ridge-regularized for the best α).

3 Clustering Algorithms

Consider the dataset provided in the $clustering\_data.mat$. Plot a dendrogram for the data using an appropriate linkage (average, centroid etc) and a distance metric (euclidean, sqeuclidean etc). Select a cutoff criterion to decide on the number of clusters in the data.

- Plot the dendrogram with the applied cutoff.
- Justify the choice of linkage, metric and cutoff.

For the selected number of clusters, use the K-Means(Lloyd's), agglomerate clustering and gaussian mixture models(GMM) to cluster the data.

- By varying the number of clusters, plot average silhouette score vs number of clusters and identify the best possible number of clusters n.
- Provide the clustered scatter plot and the silhouette plot (with average silhouette score mentioned) for each clustering algorithm with n clusters.

4 Classification

Consider the UCI wine Dataset. It is available both in python and MATLAB environments. Use only Alocohol and Malic Acid as input features. Divide the dataset into training and testing datasets containing 75% and 25% of the data, respectively. While creating the training and testing datasets, ensure that the created datasets have significant contribution from all the three classes of wines so that there is no class imbalance in data.

Using the training datasets, train a naive bayes , discriminant anlysis and K- Nearest Neighbour classifier. Plot the following

- Categorical pie charts of training and test data to ensure that they have more or less an equal contribution from all the three species.
- A visualization of the decision surfaces of all the three classifier. (A decision surface is a plot that shows how a fit machine learning algorithm divides the input feature space by class label.)
- Confusion Matrix plots for all the three algorithms with the test data to compare their classification accuracy.