Final Exam

Due: 11:59pm, Tuesday, December 10, 2019

Submit your plots, descriptions, and python scripts including all of your functions. You may use the Python scripts that you have written for your homework assignments from this semester or use `sklearn` directly.

1. (30 points) **PCA and linear regression**

   (1) (30 points) The data in `linear_regression_test_data.csv` contains x, y, and $y-$theoretical. Perform PCA on x and y. Plot y vs x, y-theoretical vs x, and the PC1 axis in the same plot.

   (2) (30 points) Perform linear regression on x and y with x being the independent variable and y being the dependent variable. Plot the regression line in the same plot as you obtained in (1). Compare the PC1 axis and the regression line obtained above. Are they very different or very similar?

2. (30 points) **PCA and LDA**

   In dataset `dataset_1.csv`, columns correspond to variables and there are two variables named `V1` and `V2`.

   (1) Plot `V2` vs `V1`. Do you see a clear separation of the raw data?
   (2) Apply PCA to this dataset without scaling the two variables. Project the raw data onto your first principal component axis, i.e. the PC1 axis. Do you still see a clear separation of the data in PC1, i.e. in projections of your raw data on the PC1 axis?
   (3) Add the PC1 axis to the plot you obtained in (1).
   (4) Apply LDA to this dataset and obtain W. The class information of each data point is in the `label` column.
   (5) Project your raw data onto W. Do you see a clear separation of the data in the projection onto W?
   (6) Add the W axis to your plot. At this point, your plot should contain the raw data points, the PC1 axis you obtain from the PCA analysis, and the W axis you obtain from the LDA analysis.
   (7) Compute the variance of the projections onto PC1 and PC2 axes. What is the relationship between these two variances and the eigenvalues of the covariance matrix you use for computing PC1 and PC2 axes?
   (8) Compute the variance of the projections onto the W axis.
   (9) What message can you get from the above PCA and LDA analyses?

3. (40 points) **Open-ended question**

   Apply PCA, clustering, linear regression, LDA, decision tree, logistic regression, and ANN to the heart disease dataset at the UCI Machine Learning Repository and describe what you could find about the data.

   The data and the description of the data can be found here: Heart Disease Dataset. Use the `processed.clevand.data` that you can download here: Download the Heart Disease Dataset.

   When you apply LDA, decision tree, logistic regression, and ANN, consider the problem as a binary classification problem to distinguish presence from absence of heart disease. When you apply linear regression, look for variables that are strongly correlated. When you apply clustering, do you see certain patients group together. Describe in detail what you find.