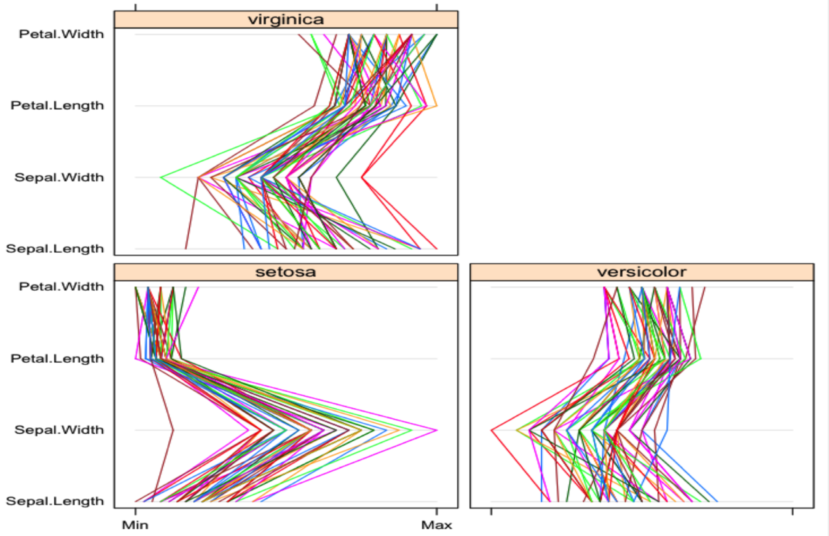
CISC 5352 Financial Data Analytics Quiz (5)

Support vector machines (SVM) Basics (20 points)

Review data visualization materials in lecture 6 and related materials in lecture 7 and 8, and solve the following problems.

1. Extract a new dataset that only contains 'setosa', (0) and 'versicolor' (1) samples and name it as Iris_binary_data.cvs (we call it iris binary data)



2. Compare SVM classification under four kernels : 'linear', 'rbf', 'poly', 'sigmoid' via a data visualization approach for iris binary data (2 dimensional data).
   (a) Your data visualization should clearly mark the decision boundaries.
   (b) You also need to draw a conclusion about the impacts of these kernels on the classification results.
3. Partition Iris binary data into training data that counts 55% of the total data and test data that counts 45% of the total data. Conduct SVM

classification under such a setting and compute the following classification measures: accuracy, sensitivity, specificity, positive predictive ratios (PPR) and negative predictive ratios. (Note: you need to use all the four kernels)

4. Read 1.4.1.1. about Multi-class classification at
    (a) http://scikit-learn.org/stable/modules/svm.html#svm-classification
    (b) Conduct Multi-class classification for the original iris data with four different kernels respectively and compare their performance.
    (c) You can choose to use "one-against-one" or 'one-vs-the-rest' approaches in multi-class classification

Revisit Credit Risk Analytics (I) (30 points)

- Credit risk analytics is key in personal loan decision making for banks. Using credit risk analytics, banks are able to analyze previous lending data, along with associated default rates, to create an effective predictive model in loan decision making.
- The file *credit_ data_ risk_ balanced.csv* has 16714 credit records with the following variables
  - 'Index': case number
  - 'Delinquency': this is a binary variable 1: means bad credit and 0: means good credit
  - 'Revolving Credit Percentage',
  - 'Capital Reserves',
  - 'Num Late 60',
  - 'Debt Ratio'
  - 'Monthly Income' ($)
  - 'Num Credit Lines' ($1000)
  - 'Num Late Past 90',
  - 'Num Real Estate',
  - 'Num Late 90',
  - 'Num Employees' (should not be more than 10 for a personal credit analytics)

1. Partition data as 80% for training and 20% for testing and use K-NN, to conduct credit risk analytics, i.e. do classification to determine good or bad credit records.
2. Compute all six classification measures and F1-measure
3. Use sklearn to compute classification measures

```
from sklearn import, metrics
metrics.classification_report( label, YOUR-classifier.predict(X_test)
```

4. Find all samples in ’TP’/’TN’/’FP’/’FN’ class
   (a) TP class: the positive samples that are correctly predicted
   (b) TN class: the negative samples that are correctly predicted
   (c) FP class: the nagtive samples that are falsely predicted
   (d) FN class: the positive samples that are falsely predicted
5. Write your own k-NN to conduct this credit risk analytics (extra credits 20 points)

Selective Learning analytics (50 points)

- dataset: NBoption.csv
- Apply selective learning for this dataset and compare its results with original machine learning methods
  - Learning machines: k-NN, SVM
  - Training/test partition: 80% training data and 20% test data
  - Train-train/train-test partition: 80%: 20%
  - Bad guys: the samples whose values $error = |Volatility - predictedVolatility|$ are at the bottom 10% of all train-test samples
  - The number of nearest neighbors: $k = 10$
  - Note: you should try selective learning both two stages: training clean and test clean
  - It's your freedom
    - to choose kernels in SVM
    - to choose distance measures in k-NN

Credit Risk Analytics (III) (30 points)

- We have the following data set for credit ranking for 12 different industry sections (it is a simulated data):
  - *credit_ data_ .csv*, where the first **1540** samples (rows) are labeled as 'good credit' (label type: '1'), i.e., whose credit rankings are 'AAA', 'AA', or 'A'
  - and the remaining **130** samples are labeled as ' bad credit', (label type: '0') whose credit ranks are 'CCC'.
- There are six variables (columns) in this data set:

```
variable 1: Working capital / Total Assets (WC_TA)
variable 2: Retained Earnings / Total Assets (RE_TA)
variable 3: Earnings Before Interests and Taxes / Total Assets (EBIT_TA)
variable 4: Market Value of Equity / Book Value of Total Debt (MVE_BVTD)
variable 5: Sales / Total Assets (S_TA)
variable 6: Industry sector labels from 1-12
```

- Complete the following problems
  - Conduct k-fold (k=10) cross validation for the data and use the following prediction to conduct classifications and compare their results
    - SVM with 'linear', 'rbf', 'poly', and 'sigmoid' respectively
    - Compare the support vectors under different kernels.
    - Compare the eigenvalues of kernel matrices under different kernels (Extra credits: 10 points)

What should you turn in?

- 1. A folder that contains
  - A report to show details of your analytics (at least 20 pages)
  - your data
  - source files
  - corresponding related output.
- 2. Please name your folder last_name1_last-name2_CISC5352_Quiz_5. For example,
- 3. Send the zipped file (.zip instead of ,rar) of your folder to Blackboard before 11:59 pm Nov 20, 2018