Classification Labs

Lab: KNN

This lab on K-Nearest Neighbors is a python adaptation of p. 163-167 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Originally adapted by Jordi Warmenhoven (github.com/JWarmenhoven/ISLR-python), modified by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Spring 2016).

In [ ]:
```
import pandas as pd
import numpy as np
```

4.6.5: K-Nearest Neighbors

In this lab, we will perform KNN on the $Smarket$ dataset from $ISLR$. This data set consists of percentage returns for the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005. For each date, we have recorded the percentage returns for each of the five previous trading days, $Lag1$ through $Lag5$. We have also recorded $Volume$ (the number of shares traded on the previous day, in billions), $Today$ (the percentage return on the date in question) and $Direction$ (whether the market was Up or $Down$ on this date). We can use the $head(...)$ function to look at the first few rows:

In [ ]:
```
df = pd.read_csv('Smarket.csv', usecols=range(1,10), index_col=0, parse_dates=True)
df.head()
```

https://github.com/jcrouser/islr-python

https://github.com/jcrouser/islr-python/blob/master/Lab%203%20-%20K-Nearest%20Neighbors%20in%20Python.ipynb

http://faculty.marshall.usc.edu/gareth-james/ISL/

Lab: Logistic Regression

This lab on Logistic Regression is a Python adaptation from p. 154-161 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Adapted by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Spring 2016).

In [1]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
```

4.6.2 Logistic Regression

Let's return to the $Smarket$ data from $ISLR$.

In [2]:
```
df = pd.read_csv('Smarket.csv', usecols=range(1,10), index_col=0, parse_dates=True)
df.describe()
```

https://github.com/jcrouser/islr-python

https://github.com/jcrouser/islr-python/blob/master/Lab%204%20-%20Logistic%20Regression%20in%20Python.ipynb

http://faculty.marshall.usc.edu/gareth-james/ISL/

Lab: LDA & QDA

This lab on Logistic Regression is a Python adaptation of p. 161-163 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Adapted by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Spring 2016).

In [ ]:
```
import pandas as pd
import numpy as np

from sklearn.lda import LDA
from sklearn.qda import QDA
from sklearn.metrics import confusion_matrix, classification_report, precision_score

%matplotlib inline
```

4.6.3 Linear Discriminant Analysis

Let's return to the $Smarket$ data from $ISLR$.

In [ ]:
```
df = pd.read_csv('Smarket.csv', usecols=range(1,10), index_col=0, parse_dates=True)
df.head()
```

Now we will perform LDA on the $Smarket$ data from the $ISLR$ package. In $Python$, we can fit a LDA model using the $LDA()$ function, which is part of the lda module of the $sklearn$ library. As we did with logistic regression and KNN, we'll fit the model using only the observations before 2005, and then test the model on the data from 2005.

https://github.com/jcrouser/islr-python

https://github.com/jcrouser/islr-python/blob/master/Lab%205%20-%20LDA%20and%20QDA%20in%20Python.ipynb

http://faculty.marshall.usc.edu/gareth-james/ISL/