Lecture 8 (march 25, 2019)

Quiz #3

April 29 - final project presentations

April 29 – May 6 (Mon – Mon) Final

kNN

Naïve Bayesian classifier

Linear Models and regression

Clustering (k-means), logistic regression, decision trees and Support Vector machines

Bayes Formula:

P(X) = P(X|Y)P(Y) = P(Y|X)P(X)

P(X | Y) = P(Y|X) P(X) / P(Y)

Given a particular x\*, what is

Label given x\*

P(label = ‘green’ | x\*)

P(label = ‘red’ | x\*)

To compute P(‘red’ | x)

Suppose I know P(x | ‘red’)

P(X | Y) = P(Y|X) P(X) / P(Y)

P(‘red’ | x\*) = P(x\* | ‘red’) P(‘red’) / P(Y)

P(X\* | ‘red’) = P(X1 | ‘red’) …. P(Xn | ‘red’)

Homework:

Part 1: Implement Naïve Bayesian for your labeled data set

Take 2017 set as training set: use Gaussian NB

Apply naïve Bayesian to 2018 and compute its accuracy

Make the following Table:

Method Accuracy:

kNN – manhattan

kNN – p=1.5

kNN – Euclidean

Naïve bayesian

You can use sklearn for both

Linear Models and Regressions

Part 2:

Take W=10

P1, P2, ….., P10:

Use linear regression to estimate P11

Let P\* be the estimated price

If P\* > P10, then tomorrow return > 0

If P\* < P10 then tomorrow return < 0

In this way, we generate estimates of daily returns for 2018

Compare with real daily returns and compute the accuracy

We would like to investigate the accuracy of using linear regression for different W

W = 10, 20, 30

For this assignment, you can use polyfit (degree 1) or LinearRegression

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

import numpy as np

x = np.array([1,2,3,4,5,6,7,8,9,10])

y = np.array([])

degree = 1

weights = polyfit(x,y,1)

model = np.polyfit(weights)

predicted = model(11)

if predicted > y[-1]:

rate = 1

else:

rate = -1

Prepare the following dataset:

Take 2017 and 2018 data (x, y, label)

104 points

Draw a line to separate points

Remove red labels below the line and green above the line

Task: create a “linearly separable” points

Questions: