Hao Wu

MET CS 677

Assignment 2

Question 1

the default probability p ∗ that the next day is a ”up” day is 0.5298

the probability of seeing ”−, −, −, +” as opposed to seeing ”−, −, −, −”. Compute this for k = 1 is 0.5092, k = 2 is 0.4865 and k =3 is 0.4893.

the probability of seeing ”+, +, +, +” as opposed to seeing ”+, +, +, −”? Compute this for k = 1 is 0.4957, k = 2 is 0.4256 and k = 3 is 0.4124

Question 2

percentage of true labels (both positive and negative) have you predicted correctly for the last two years is 0.7396.

when w=2 is most accuracy

Question 3

The accuracy rate is still 0.7396.

Imporved my accuracy rate on predict “-“ label.

Question 4

TP = 263

FP = 131

TN = 109

FN = 0

TPR = TP/(TP+FN) = 1

TNR = TN/(TN+FP) = 0.4541

Maybe my model wrong, the FN shouldn’t be zero.

Code：

import pandas as pd

import numpy as np

df1 = pd.read\_csv("/Users/haowu/Desktop/Boston University Graduate Study/CS 677/assignment 2/HA.csv")

df1.head()

df1.assign(True\_Label = "")

true\_label = []

for value in df1["Return"]:

if value >= 0:

true\_label.append("+")

if value < 0:

true\_label.append("-")

df1["True\_Label"] = true\_label

df1.head()

df2 = df1.groupby(['True\_Label','Year'])['Day'].count()

prob = 400/755

prob

np.random.seed(100)

df3 = pd.Series(np.random.choice(['+','-'], 750))

ia = df3.groupby((df3 == '+').cumsum()).cumcount()

prob1 = (ia.diff() <= -3).sum() / (ia >= 3).sum()

prob2 = (ia.diff() <= -2).sum() / (ia >= 2).sum()

prob3 = (ia.diff() <= -1).sum() / (ia >= 1).sum()

print("prob1:",prob1,"prob2:",prob2,"prob3:",prob3)

df4 = df1.loc[df1['Year'] > 2017]

df4.reset\_index(drop=True, inplace=True)

df4

df5 = df4.assign(Predict = "")

ib = df5.groupby((df5["True\_Label"] == '+').cumsum()).cumcount()

ib

df6=ib.to\_frame()

df6.columns = ['A']

predict = []

for value in df6['A']:

if value > 1:

predict.append("-")

if value <= 1:

predict.append("+")

df5["Predict"] = predict

df5.head()

comparison\_column = np.where(df5["True\_Label"] == df5["Predict"], True, False)

count = sum(comparison\_column)

rate = count / len(comparison\_column)

rate

df7 = df5.loc[df5['Predict'] == "+"]

df7.reset\_index(drop=True, inplace=True)

df7

df8 = df5.loc[df5['Predict'] == "-"]

df8.reset\_index(drop=True, inplace=True)

df8

comparison\_column2 = np.where(df7["True\_Label"] == df7["Predict"], True, False)

comparison\_column3 = np.where(df8["True\_Label"] == df8["Predict"], True, False)

TP = sum(comparison\_column2)

FP = len(comparison\_column2) - TP

TN = sum(comparison\_column3)

FN = len(comparison\_column3) – TN

文本

描述已自动生成电脑萤幕的截图

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成