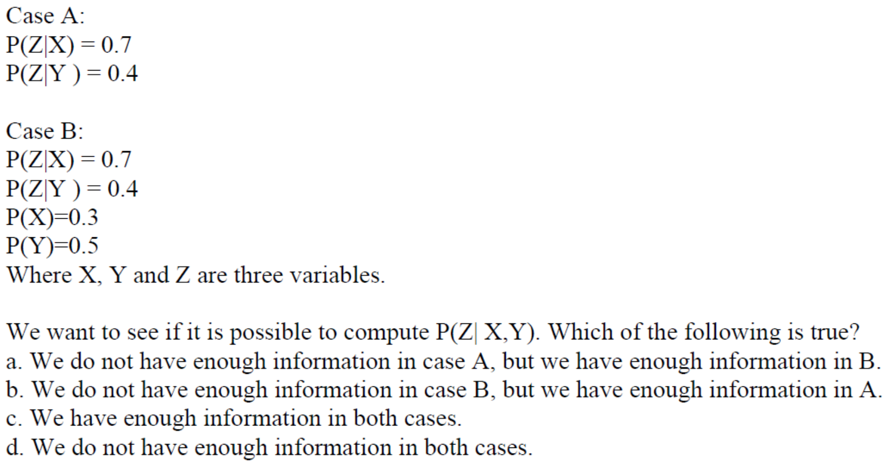
Assignment 2
ECO481
University of Toronto

Marlène Koffi

March 9, 2023

1 Exercise 1: Multiple-choice question (please provide an explanation for your choice (12 points)

1. Which statement is more likely to be true?
   - a. A classifier trained on less training data is less likely to overfit.
   - b. When the feature space is larger, overfitting is less likely.
   - c. As the number of training examples goes to infinity, your model trained on that data will have a lower variance.
   - d. As the number of training examples goes to infinity, your model trained on that data will have a lower bias.
2. Suppose I give you the following information:



Case A:
$P(Z|X) = 0.7$
$P(Z|Y) = 0.4$

Case B:
$P(Z|X) = 0.7$
$P(Z|Y) = 0.4$
$P(X)=0.3$
$P(Y)=0.5$
Where X, Y and Z are three variables.

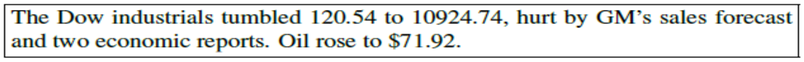
We want to see if it is possible to compute $P(Z| X,Y)$. Which of the following is true?
a. We do not have enough information in case A, but we have enough information in B.
b. We do not have enough information in case B, but we have enough information in A.
c. We have enough information in both cases.
d. We do not have enough information in both cases.

3. Which of the following statement is true for both Naïve Bayes classifier and decision trees?

- a. In both classifiers, a pair of features are assumed to be independent.
- b. In both classifiers, a pair of features are assumed to be dependent.
- c. In both classifiers, a pair of features are assumed to be independent given the class label.
- d. In both classifiers, a pair of features are assumed to be dependent given the class label

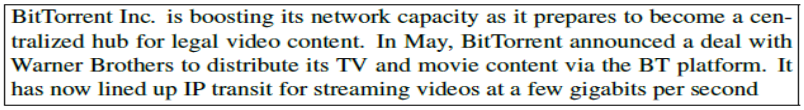
2 Exercise 2: Text classification using Naive Bayes (25 points)

We want to classify some texts using Naive Bayes Classifier.



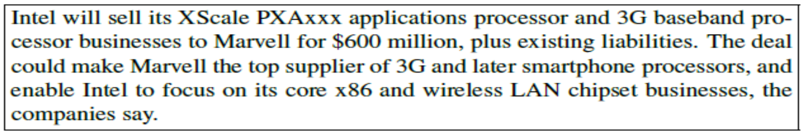
The Dow industrials tumbled 120.54 to 10924.74, hurt by GM's sales forecast and two economic reports. Oil rose to $71.92.

from `www.wsj.com/`



BitTorrent Inc. is boosting its network capacity as it prepares to become a centralized hub for legal video content. In May, BitTorrent announced a deal with Warner Brothers to distribute its TV and movie content via the BT platform. It has now lined up IP transit for streaming videos at a few gigabits per second

from `slashdot.org/`

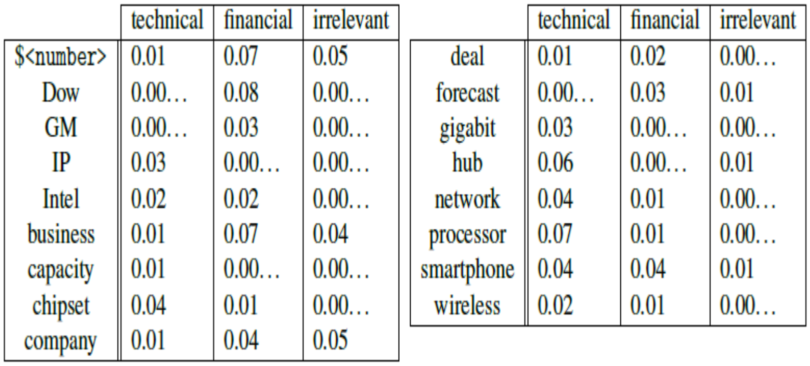


Intel will sell its XScale PXAxxx applications processor and 3G baseband processor businesses to Marvell for $600 million, plus existing liabilities. The deal could make Marvell the top supplier of 3G and later smartphone processors, and enable Intel to focus on its core x86 and wireless LAN chipset businesses, the companies say.

from `www.linuxdevices.com/`

The potential labels are: technical, financial and irrelevant.

In addition, we know the word frequencies:



	technical	financial	irrelevant
$<number>	0.01	0.07	0.05
Dow	0.00…	0.08	0.00…
GM	0.00…	0.03	0.00…
IP	0.03	0.00…	0.00…
Intel	0.02	0.02	0.00…
business	0.01	0.07	0.04
capacity	0.01	0.00…	0.00…
chipset	0.04	0.01	0.00…
company	0.01	0.04	0.05

	technical	financial	irrelevant
deal	0.01	0.02	0.00…
forecast	0.00…	0.03	0.01
gigabit	0.03	0.00…	0.00…
hub	0.06	0.00…	0.01
network	0.04	0.01	0.00…
processor	0.07	0.01	0.00…
smartphone	0.04	0.04	0.01
wireless	0.02	0.01	0.00…

Assume that ”0.00...” $= \epsilon$ is a very small value.

Moreover, the prior distribution is given by: 50% for technical, 40% for financial, and 10% for no interest. Last, ignore all the other words not in the previous table.

1. Apply the Naive Bayes Classifier to those texts. Provide a full explanation of all the steps and computations that lead to your results. (15 points)

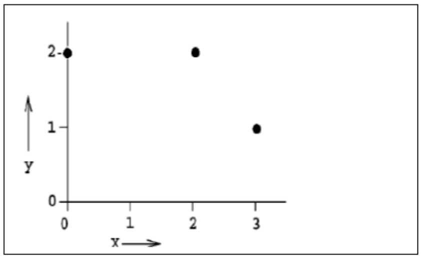
2. In the next step, we would like to focus on expressions like ”network capacity”.

    - a. How does a Naive Bayes classifier handle the expressions? (Hints: you should discuss two cases). (5 points)

- b. What Python package and module will you use to build the expressions? Be specific and clearly explain the different steps you will take. (5 points)

3 Exercise 3: MSE (20 points)

You are given the following data points.



X	Y
0	2
2	2
3	1

1. You are doing 3-fold cross-validation. Each time the model is learned from the non-left-out data points. Assume you use a trivial algorithm that predicts a constant y = c. What is the mean square error from the 3-fold cross-validation? (10 points)

2. You are doing a 3-fold cross-validation. Each time the model is learned from the non-left-out data points. What is the mean square error from the 3-fold cross-validation assuming you fit a linear regression ($Y = \beta_0 + \beta_1 X + \epsilon$). (10 points)

4 Exercise 4: Build a Covid uncertainty index (43 points)

You work at the Federal Reserve in the US.[1] You are in charge of analysing the impact of Covid on the stock market for the last 30 days. You also want to assess the general feeling using the news for this period.

Your boss gives you the database of headlines from the New York Times that I constructed by webscrapping the information on the archives page of The New York Times. The data is contained in the csv file called NYT_headline.csv. There are two columns. The first one is related to the headline. The second one is related to the date (the date of publication of the article with the corresponding headline). I restrict the collection on articles about the US. The period covered is: February 1, 2021 to March 12, 2021.

1. Read the file NYT_headline.csv on python and drop the duplicates (1 point).

[1] Unfortunately, it was difficult to find historical news on Canada.

2. Build a vocabulary of Covid-19 related words (3 points).

3. Combine the different headlines by day (1 point).

4. Use topic modelling to exhibit the key topics of the headlines (10 points).
   NB: Find the optimal number of topics, name the topics and display the topics using wordclouds.

5. Using the vocabulary constructed, build a daily covid related index (that we will call the covid uncertainty index) by estimating the relative fraction of articles related to covid to the total number of articles per day (5 points).

6. Use the following words "uncertainty", "uncertain", "economic", "economy", "Congress", "deficit", "Federal Reserve", "legislation", "regulation", or "White House", "uncertainties", "regulatory", or "the Fed" to construct a daily economic policy uncertainty index. In the same manner as for the covid uncertainty index, build the current index by estimating the relative fraction of articles that use any of those words. We will call it a *coarse economic policy uncertainty index* (3 points).

7. Can you argue why this is not a very good way of assessing economic policy uncertainty (this is why it is "coarse") (3 points)?

8. Use the variable "Adj Close" to compute the return on S&P500 (ĜSPC) (3 points).

9. Using a plot and simple correlations, exhibit the link between the Covid uncertainty index, the coarse economic policy index and the returns. Comment on your findings (3 points).

10. Select the articles that contains at least one word in the covid-related dictionary you constructed. For those articles, use the Vader sentiment lexicon and construct:

    - a) a daily sentiment index and plot. (Just consider the dates with a covid-related word. The dates without any covid-related word are considered as missing values). Include the three dimensions: Negative, Neutral and Positive (5 points).

    - b) an aggregate sentiment over all the period of the database. Include the three dimensions: Negative, Neutral and Positive (3 points).

11. Your boss asks you to write a short paragraph highlighting your key findings on this study. What will this paragraph look like? (No more than 5 lines) (3 points).

References:

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. Measuring economic policy uncertainty. The quarterly journal of economics 131.4 (2016): 1593-1636.

Baker, S. R., Bloom, N., Davis, S. J., Kost, K. J., Sammon, M. C., & Viratyosin, T. (2020). The unprecedented stock market impact of COVID-19 (No. w26945). National Bureau of Economic Research.

Baker, S. R., Bloom, N., Davis, S. J., & Terry, S. J. (2020). Covid-induced economic uncertainty (No. w26983). National Bureau of Economic Research.