ECO421
University of Toronto

Marlène Koffi

November 16, 2021

The assignment is graded out of 105 points. There are 5 bonus points to be obtained.

1 Exercise 1: Build a Covid uncertainty index (45 points)

You work at the Federal Reserve in the US.[1] You are in charge of building a Covid uncertainty index and analysing the impact of Covid on the stock market. 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 (2 points).
2. Build a vocabulary of Covid-19 related words (3 points).
3. Combine the different headlines by day (1 points).
4. Use topic modelling to exhibit the key topics of the headlines (8 points).
   NB: Find the optimal number of topics, name the topics.
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).

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

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* (5 points).

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

8. 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 (5 points).

9. Select the articles that contains at least one word in the covid-related dictionary you constructed (2 points). 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) as well as Negative-Positive (5 points).
   - b) an aggregate sentiment over all the period of the database. Include the three dimensions (Negative, Neutral and Positive) as well as Negative-Positive (3 points).

10. 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).

2 Exercise 2: Bias-Variance Trade-off (28 points)

An important concept in machine learning is the bias-variance trade-off. Let us discover through this exercise what it is about. This exercise will build upon your current knowledge of machine learning, your knowledge of econometrics, and your reasoning.

Suppose we have the following model:

$$y = f(x) + \epsilon \tag{1}$$

The machine learning methods we have seen in the supervised learning aim to have an estimate of the function f, that we will call $\hat{f}$. We want to assess the error of the prediction.

We have a new observation A for which the value of x is x_0 and we predict the corresponding value y_0 (the true value) using our estimate $\hat{f}$. We want to evaluate analytically the accuracy

of our prediction using the expected test mean squared error (expected test MSE) defined as $E((y_0 - \hat{f}(x_0))^2)$.

Note: To refresh your mind, E stands for expectation and is usually estimated using the sample average. Recall that we have seen earlier in the course that to evaluate the accuracy of a machine learning regression model, we use the test mean square error.

1. Assuming that:

   - x is known (no uncertainty, x is not a random variables)
   - $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2$
   - $E[\epsilon(g(x))] = 0$ where $g(x)$ is a function of x.

   Show that equation the expected MSE $E((y_0 - \hat{f}(x_0))^2)$ is equal to:

   $$E(f(x_0) - \hat{f}(x_0))^2 + Var[\epsilon] \tag{2}$$

   (5 points)

2. Assuming that:

   - x is known (no uncertainty, x is not a random variables)
   - $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2$
   - $E[\epsilon(g(x))] = 0$ where $g(x)$ is a function of x.

   Using equation 2, show that you can further rewrite the expected MSE $E((y_0 - \hat{f}(x_0))^2)$ as:

   $$\underbrace{E[(E\hat{f}(x_0) - \hat{f}(x_0))^2]}_{Var(\hat{f}(x_0)} + \underbrace{(f(x_0) - E\hat{f}(x_0))^2}_{[Bias(\hat{f}(x_0))]^2} + \underbrace{\sigma^2}_{\text{irreducible error}} \tag{3}$$

   With $E\hat{f}(x_0) = E[\hat{f}(x_0)]$. This is the expectation of $\hat{f}(x_0)$ (bonus: 5 points).[2]

   Equation 3 gives us the bias-variance trade-off. It states that the expected test MSE of the prediction can be decomposed into a term relating to the variance, a term relating to a squared bias and an irreducible error.

   $$E((y_0 - \hat{f}(x_0))^2) = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{4}$$

3. Choose the right answers (no explanation needed, you can have multiple right answers)). What is equation (4) telling you (7 points)?

[2] If it is not clear for some of you, make the parallel with the linear regression. For example in a simple linear regression model, $\hat{f}(x_0) = \hat{\beta}x_0$ and $E\hat{f}(x_0) = x_0 E[\hat{\beta}]$. $\hat{\beta}$ is an estimator of the true value β.

- a: In order to minimize the expected test MSE, we need to select a machine learning method that simultaneously achieves low variance and low bias.
- b: A machine learning method with a bias different from 0 is a bad predictor.
- c: A machine learning method with a variance different from 0 is a bad predictor.
- d: Starting from a given situation, it is possible to achieve a lower expected test MSE by simultaneously increase the bias and reduce the variance provided that the reduction in the variance is bigger (in absolute value) than the increase in the bias.
- e: Starting from a given situation, it is impossible to achieve a lower expected test MSE by simultaneously increase the variance and reduce the bias provided that the reduction in the bias is bigger (in absolute value) than the increase in the variance.
- f: Starting from a given situation, it is always possible to achieve a lower expected test MSE by simultaneously reduce the bias and reduce the variance.
- g: The expected test MSE can never lie below $Var(\epsilon)$.

4. Now, let us interpret the different terms. In JWHT,"Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set". Bias refers to the "error that is introduced by approximating a real-life problem,..., by a much simpler model."

   - a) From this interpretation, how could you compare the Decision Tree to the Random Forest in terms of bias and variance (8 points)?
   - b) Suppose your friend has a magical learning algorithm which returns the true labelling function regardless of the training set. Consider the following statement that your friend makes to you: "My algorithm has low bias, since its predictions are always correct, but high variance, since its predictions are quite different for different datapoints." Is your friend correct about bias (4 points)? Is your friend correct about variance (4 points)? Comment on each of your answers.

3 Exercise 3: Unsupervised Learning: K-means (27 points)

3.1 Multiple choice (24 points)

You can pick more than one answer. For each question, you must explain your answer. For each question, you have a total of 3 points when you pick the correct answer and 3 points for the explanations.

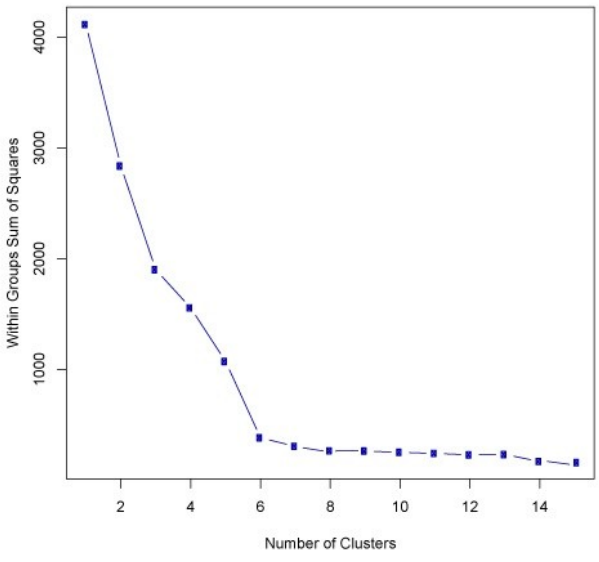
Q1: The minimum number of variables required to perform clustering is:

1. 0
2. 1
3. 2
4. 3
5. More than 5

Q2: You plan to buy a new outfit for the new year eve. You went on a shopping website. One week later, while you were working on your were on a website on ML for economist, you receive some recommendations of outfit. This recommendation is an example of:

1. Classification
2. Clustering
3. Regression

Q3: What should be the best choice for number of clusters based on the following results:



1. 5
2. 6
3. 14

4. Greater than 14

Q4: What is (are) the wrong statement(s)?

1. k-means clustering aims to partition n observations into k clusters.
2. k-nearest neighbor is same as k-means.
3. For two runs of K-mean clustering, I can have the same clustering results.
4. Assignment of observations to clusters does not change between successive iterations in K-means.

3.2 Open-ended question (3 points)

Propose one way in which clustering (Unsupervised Learning) can be used to improve the accuracy of a Linear Regression model (Supervised Learning).

References:

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.