MODULE 5 UNIT 2
Activity submission

Learning outcomes:

LO4: Articulate an understanding of the practical application of generative models in relation to classification problems.

LO5: Evaluate the prediction of a generative model in R.

Plagiarism declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. This assignment is my own work.

3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work.

4. I acknowledge that copying someone else's assignment (or part of it) is wrong and declare that my assignments are my own work.

Name: Vasileios Tsoumpris

1. Instructions and guidelines (Read carefully)

Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M5 U2 Activity Submission** – **e.g. Lilly Smith M5 U2 Activity Submission. NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, "Start writing here" under each question. Please type your answer there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

PLEASE NOTE: Plagiarism cases will be investigated in line with the Terms and Conditions for Students.

IMPORTANT NOTICE: Please ensure that you have checked your course calendar for the due date for this assignment.

Guidelines

1. There are seven pages and two questions in this assignment.

2. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully but concisely and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

3. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

4. At the end of your assignment, please provide feedback on areas where you require further assistance or would like the Assessor to expand on.

2. Mark allocation

Each question receives a mark allocation. However, you will only receive a final percentage mark and will not be given individual marks for each question. The mark allocation is there to show you the weighting and length of each question.

Question 1	18
Question 2	18
TOTAL	**36**

3. Questions

After completing the different steps in the IDE notebook, you have applied the naive Bayes classifier to the reduced newsgroups data set in R.

With specific reference to what you’ve learnt in the assessment IDE activity, answer the following questions.

Question 1

Now that you have generated an output in the final code cell in R, you have a better understanding of how accurately your model distinguishes between different topics in text.

Refer to the output generated in the final code cell in the assessment IDE activity, analyse the confusion matrix and statistics, and answer the following questions:

1.1 Consider the number of incorrect classifications made by the model. Why do you think the model incorrectly classified rec.auto classes as rec.motorcycles, and rec.motorcycles classes as rec.auto?

1.2 What is the accuracy rate of the model expressed as a percentage? Elaborate on whether this accuracy rate is good or not, and substantiate your answer.

(Max. 300 words)

Start writing here:

The prediction accuracy of the model is demonstrated by its confusion matrix. The confusion matrix depicts that the Naive Bayes model accurately predicted 156 "rec.autos" and 151 "rec.motorcycles" cases. On the other hand, the incorrect predictions were 32 for the "rec.autos" and 19 for the "rec.motorcycles". Most of the time, these inaccuracies derive from the data preprocessing. In our case, we work with text data which generally are harder to handle. For example, identifying the most frequent terms in documents (freqwords R variable) plays a significant role in how the model will predict. Apart from just counting the frequency of words, which is generally a simple approach, other techniques could improve the performance of the model, enabling it to reach its full potential and maybe to same accuracy level of more advanced methods. These techniques include grouping different inflections of similar words (lemmatising words), counting sequences of words (n-grams) or penalising words that appear frequently in most of the text. Lastly, the Naive Bayes model implicitly assumes that all the attributes are mutually independent. In reality, it's almost impossible to work with a set of predictors that are entirely independent of one another.

The model accurately predicts 85.75% of the test data. Generally speaking, the accuracy rate is quite good; however, whether the accuracy rate is considered acceptable depends heavily on the context in which the model will be applied. Given that the Sensitivity (82,9%) and Specificity (88,8%) are both high, the model acceptably answers the question of what percentage of cases are correctly identified as "rec.auto" (positive class) and "rec.motorcycle" (negative class). What is more, the p-value (=2e-16) is smaller than 0.001, meaning that the results are replicable and of substantial significance.

Question 2

Write a brief business pitch in which you argue that a specific area within your context could benefit from applying the naive Bayes classifier in solving a specific problem. When substantiating your point, focus on the following aspects:

- The problem statement describing the problem that should be solved
- The reason why you would use the naive Bayes classifier instead of another classification method to solve this specific problem

(Max. 350 words)

Start writing here:

The improvement of the manufacturing pipeline of an industrial company and the industrial automation are essential in order to enter the Industry 4.0 ecosystem. NLP is one approach that could help the manufacturing company to analyse an exorbitant quantity of shipment documents and provide a better insight into what areas of their supply chain lag behind. In this way, the manufacturing company will better elaborate its operational processes and make logistical changes to increase efficiency.

The NLP algorithm should classify whether or not one document shows evidence that a specific supply chain area needs optimisation by categorising the shipping documents into two distinctive categories "Late Delivery" and "On time Delivery". The model pipeline will

involve data pre-processing (removal of short structured words, punctuation, stop words, conversion of alphabet cases, grouping sentences based on similarity etc.), the creation of an industry-specific lexical dictionary and the application of the ML model that makes a binary classification.

For this specific case, the probabilistic Naive Bayes classifier could be used. The calculation of probabilities is the primary reason this algorithm performs much better than other algorithms like Logistic Regression in text classification. Moreover, a Naive Bayes model does not require a lot of training data to give an acceptable accuracy. In addition, if the conditional independence assumption holds true, this model could give excellent results. Furthermore, it is quite a fast model compared to neural networks, which also have an advantage in text classification problems. So if the training speed is an aspect the company considers, Naive Bayes could prove an advantageous model. Lastly, the shipping documents have a specific structure and dictionary, so it is pretty rare for a categorical variable to have a class in the test data that was not observed in the training data. As a result, it would be easier to handle the major disadvantage of Naive Bayes, the zero probability problem. Therefore, for the reasons mentioned above, the Naive Bayes model could be an optimal solution for this NLP business problem.

4. Rubric

The following rubric will be used to grade your submission for this activity submission.

	Unsatisfactory	Limited	Accomplished	Exceptional
Question 1: Understanding and accurate use of module content *The answer demonstrates that the student engaged with the content.* *The answer demonstrates that the student has an informed grasp of how to interpret the output generated from fitting a naive Bayes classifier onto a data set, and has the ability to use the output to analyse the accuracy of the model.*	The answer demonstrates that the student did not engage with the content. OR The answer fails to demonstrate a basic understanding of the content. (0)	The answer demonstrates that the student engaged with the content. The answer demonstrates an inadequate understanding of the content. (2)	The answer demonstrates that the student engaged with the content and understands most of it. (4)	The answer demonstrates that the student engaged with the content and has an excellent understanding of the module's content. (6)
Question 1: Adherence to the brief *The answer provides a brief analysis of the output generated in the IDE notebook, and refers to the results generated therein.*	No submission. OR Answer fails to adhere to any of the elements contained in the brief. (0)	Answer adheres to some elements contained in the brief, but some key elements are missing. Answer does not fall within the prescribed word count (50 words over the word count). (2)	Answer adheres to most, but not all, elements of the brief. Almost all information is provided and relevant. Answer falls within the prescribed word count. (4)	Answer adheres to all the elements of the brief. All information provided is comprehensive and relevant. Answer falls within the prescribed word count. (6)

Question 1: Coherence and clarity *The answer is clearly structured and written in a way that is comprehensible.*	Answer is incoherent or lacks clarity. Answer is not logically structured, or is incomprehensible. (0)	Answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. (2)	Answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. (4)	Answer is extremely well structured and written with exceptional clarity and coherence. (6)
Question 2: Understanding and accurate use of module content *The answer demonstrates that the student engaged with the content.* *The answer demonstrates that the student has an informed grasp of the application of the naive Bayes classifier and the specific types of problems it can solve.*	The answer demonstrates that the student did not engage with the content. OR The answer fails to demonstrate a basic understanding of the content. (0)	The answer demonstrates that the student engaged with the content. The answer demonstrates an inadequate understanding of the content. (2)	The answer demonstrates that the student engaged with the content and understands most of it. (4)	The answer demonstrates that the student engaged with the content and has an excellent understanding of the module's content. (6)
Question 2: Adherence to the brief *The answer provides a concise pitch of the specific problem that the naive Bayes classifier could be used to solve. The answer includes a description of each guideline in a concise and structured way.*	No submission. OR Answer fails to adhere to any of the elements contained in the brief. (0)	Answer adheres to some elements contained in the brief, but some key elements are missing. Answer does not fall within the prescribed word count (50 words over the word count). (2)	Answer adheres to most, but not all, elements of the brief. Almost all information is provided and relevant. Answer falls within the prescribed word count. (4)	Answer adheres to all the elements of the brief. All information provided is comprehensive and relevant. Answer falls within the prescribed word count. (6)

Question 2: Coherence and clarity *The answer is clearly structured and written in a way that is comprehensible.*	Answer is incoherent or lacks clarity. Answer is not logically structured, or is incomprehensible. (0)	Answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. (2)	Answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. (4)	Answer is extremely well structured and written with exceptional clarity and coherence. (6)

Total: 36 marks

Feedback:

Start writing here: