Frequently Asked Questions – New questions will be added as needed

Q1. Can we use the permitted libraries (*sklearn, pandas, numpy, math, mathplotlib*) for dividing the dataset to Train and Test sets? (e.g. for developing the split-data() function?)

A1. Yes.

Q2. Can we use permitted libraries (*sklearn, pandas, numpy, math, mathplotlib*) for Evaluating the results from our Prediction function? (e.g. for developing the Evaluation() function?)

A2. Yes.

Q3. Can we use any other libraries (except the permitted one)?

A3. No, you cannot use any other library or package other that the permitted ones.

Q4. What does stratification mean, should we use in our split-data() function?

A4. NO, you are asked to use the hold-out strategy. Stratification (in the context of sampling) means to select a subset of the data such that different groups (= strata) of the data are represented with the same distribution in the subset as in your overall data. In the context of classification, these “groups” of interest may for example be class labels.

Q5. What ratio should we use for train and test division?

A5. You should make an informed decision based on what you have learned in the lectures, and explain that decision in your answer to the question (the most important aspect here is that you show that you thought about the problem and can provide reasons for your decision – there’s no single correct answer!).

Hint: Using different ratios and comparing the results can lead to interesting results! ;)

Q6. Is the word restriction for each question or each sub-question?

A6. We expect 100 – 250 words per each subsection (e.g. Q1.a). You can also use codes to produce diagrams and plots to clarify your answers (Do not attach image(s) to your file).

Q7. How strict is the word limit of 100-250 words for the answers?

A6. The limit indicates what we expect: meaningful explanations and analyses, but no full essays. Try to stick to it as best you can, we allow for a margin of +/- 10%

Q8. How should we answer the assignment questions? How should we develop the answers?

A8. Dig deep into the data set and the results from your implementation and try to find some information (as evidence) that leads to a certain behaviour of your model that you are analysing. *E.g. I observed A in the results of my implementation and that could be explained by property B in Naïve Bayes; Or due to property X in Naïve Bayes model, and the observation Y in the dataset, the results of the implementation should be Z which conforms with the results from my implementation.*

Q9. How should we answer Q1 part a?

A9. Q1.a has 3 sections:

- Can you see any interesting characteristic in features, classes or categories?

  You can use diagrams and plots to analyse the features, their relation with each other and with the class and discover a pattern. Here, we are not looking for a certain correct answer. There are many things that you can find interesting. You can also use your prior knowledge in the “stroke prediction” domain.

- What is the main issue with the data?

  Here we are looking for a certain response. There is only one correct answer.

- Considering the issue, how would the Naive Bayes classifier work on this data?

If you find the answer to the previous question correctly, using the characteristics of the Naïve Bayes classifier, you would be able to infer a logical answer. Again here, we are looking for a certain correct answer.

Q10. What data structure should be used for saving the NB model (the prior and conditional probabilities)?

A10. Please confer the slides of lecture 4 (Naive Bayes), which suggest different options.

Q11. Do different values for epsilon (in Epsilon smoothing method) lead to different results, if so which one is better?

A11. That's an excellent question for you to explore and discuss in your solutions.

Q12. If we have tie when comparing the probabilities in the NB model what should we do?

A12. Again, that's a great issue for analysis and discussion in your solutions. You may want to check for the reasons of the tie: e.g., is it caused by missing features?

Q13. Where should we answer the questions?

A13. All the code and the answer to the questions should be in one .ipynb (Jupiter notebook) file.

Q14. How many files should we submit?

A14. You should submit exactly ONE .ipynb file that includes your code and your answers to the questions 1 to 3.

Q15. Should we use a certain distance metric in the KNN?

A15. No. You can use any distance metric you find appropriate for the given dataset. Remember as always you need to justify your decision. It's part of the challenge.

Q16. In Q2.b, how can we combine both Gaussian and Categorical Naive Bayes?

A16. It's up to you as a part of the challenge. One possible approach can be that you choose different methods and compare the results and explain your observations based on the theoretical characteristics of the Naïve Bayes classifier.

Q17. Are there any requirements regarding the level of performance they need to achieve?

A17. No, we are not marking your answers based on the level of the accuracy you have achieved. It means if one student has achieves 90% accuracy, and another 70%; it does not necessarily mean that the student with higher accuracy will achieve a higher mark. What we are looking for is the evidence that you can intuitively explain the why behind your observations (or decisions) and make enough connection with the empirical results. In other words, you need your results from the code to identify the special characteristics of the given dataset and be able to explain the behaviour of NB (and KNN) classifiers when being applied on the dataset, making connections between observed results and the theoretical features of the method.

Q18. How can I convert numeric feature(s) to categorical (nominal) features?

A18. There are many ways of converting continuous(numeric) data to discrete data, usually covered under the title of *"discretisation methods"* in Machine Learning content. One of them is "equal-width" discretisation, another is "equal-frequency" discretisation. There are many other methods as well, including clustering methods (e.g. k-means), supervised methods, and more. We'll cover some of these methods briefly in week 5 workshops. You are allowed and encouraged to use any method you find appropriate. Keep in mind, for any method you choose (for any part of the assignment) you have to include a valid justification.

Q19. How many features we should use for training and evaluating the model?

A19. You need to use all 10 features in the given dataset for training and evaluating your model(s).