FAQ

Q1. Can we use libraries (sklearn, Pandas, etc.) for dividing the dataset to Train and Test sets? (e.g. for developing the `split-data()` function?)

A1. **Yes**. (**Note** that this is in contrast to to the advice in the consultation session).

Q2. Can we use libraries (sklearn, Pandas, etc.) for Evaluating the results from our Prediction function? (e.g. for developing the `Evaluation()` function?)

A2. **Yes**. (**Note** that this is in contrast to to the advice in the consultation session).

Q3. What does stratification mean, should we use if in our split-data() function?

A3. 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. You may choose any sampling strategy you find appropriate, and be sure to explain and justify it in your response.

Q4. How strict is the word limit of 100-250 words for the answers?

A4. 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%

Q5. How should we answer the assignment questions? How should we develop the answers?

A5. Dig deep into the data set and try to find a pattern that leads to a certain behaviour of your model that you are analysing. E.g for question 1, you could try to find a group of instances that seem like the model is more successful if identifying their labels, and compare them with another possible group(s) of instances that your NB model couldn't predict their labels very well and explain which property in each of these groups may have lead to these different behaviours.

Q6. What data structure should be used for saving the NB model (the prior and conditional probabilities)?

A6. Please confer the slides of lecture 4 (Naive Bayes), which suggest different options. You will find that you need to represent your priors along 2 dimensions (1 measurement per 1 class) and your likelihoods along 3 dimensions (1 measurement per 1 feature per 1 class).

Q7. What ratio should we use for train and test division?

A7. 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!)

Q9. Do different values for epsilon (in Epsilon smoothing method) lead to different results, if so which one is better?

A9. That's an excellent question for you to explore and discuss in your solutions.

Q10. If we have tie when comparing the probabilities in the NB model what should we do?

A10. 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?

Q11. When should we use Precision and when we should use Recall?

A11. Precision tells us how many of the instances that our classifier predicted to be "interesting" are actually "interesting". Recall tells us how many of the instances that are actually "interesting" were predicted as "interesting" by our classifier. It depends on your problem whether you want to be precise (i.e., make sure to minimise the number of wrong "interesting" predictions; e.g., a spam classifier predicting "spam" only when it is really certain – at the risk to miss a few emails that should have been classified as "spam"), or to achieve a high recall (i.e., make sure to minimise the number of truly "interesting" instances that your classifier misses; e.g., a spam classifier that wants to capture every potential spam email as "spam" – at the risk of predicting "spam" for a couple of emails which weren't actually "spam"). We can only judge which measure is more appropriate in the context of the problem or application. The F-measure is a way of combining both measures into a single number.