This list contains exercises of the type you will find in an exam for the course *Natuurlijke Taalmodellen en Interfaces*.

# Contents



### Points

| Question: | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Points: | 1 | 1 | 2 | 8 | 7 | 19 |

# 1 Naive Bayes Classification

1. (1 point) In Naive Bayes classification for sentiment analysis, we model the conditional probability of a class $c$ given input features $f_1^n$ (i.e. sentiment words in a movie review) by application of Bayes rule:

$$\underbrace{P_{C|F_1^n}(c|f_1^n)}_{\text{posterior}} = \frac{\overbrace{P_C(c)}^{\text{prior}} \times \overbrace{P_{F_1^n|C}(f_1^n|c)}^{\text{likelihood}}}{\underbrace{P_{F_1^n}(f_1^n)}_{\text{marginal}}}$$

where the likelihood term is further simplified by making $F_i \perp F_j | C$ for any $i \neq j$.

Assume the following likelihoods for each word being part of a positive or negative movie review, and equal prior probabilities for each class.

| $F$ | $C = +$ | $C = -$ |
|---|---|---|
| I | 0.09 | 0.16 |
| always | 0.07 | 0.06 |
| like | 0.29 | 0.06 |
| foreign | 0.04 | 0.15 |
| films | 0.08 | 0.11 |

What class will Naive Bayes assign to the sentence "I always like foreign films"?[1]

2. (1 point) Given the following short movie reviews, each labeled with a genre, either comedy or action:

| | | |
|---|---|---|
| 1. | fun, couple, love, love | comedy |
| 2. | fast, furious, shoot | action |
| 3. | couple, fly, fast, fun, fun | comedy |
| 4. | furious, shoot, shoot, fun | action |
| 5. | fly, fast, shoot, love | action |

and a new document:

*fast, couple, shoot, fly*

compute the most likely class for D. Assume a naive Bayes classifier and use Laplace (add-1) smoothing for the likelihoods (consider the vocabulary as the union of the word types in the training documents).[2] *Remark:* do not add an UNK type to the vocabulary of features.

[1] This is exercise 4.1 of https://web.stanford.edu/~jurafsky/slp3/4.pdf

[2] Exercise 4.2 in https://web.stanford.edu/~jurafsky/slp3/4.pdf

3. (2 points) Explain interpolation and Naive Bayes and discuss the main differences in how they model and estimate a conditional $P_{C|F_1^n}$ where $C$ takes on values in a set $\mathcal{C}$ of $K$ classes and $F$ takes on values in a set $\mathcal{F}$ of $v$ words.

## 2 Logistic Regression

4. Naive Bayes classification allows us to condition efficiently on high-dimensional data, but does so by making a strong independence assumption. A more general strategy involves modelling the conditional probability directly by means of an exponentiated linear model (also known as a log-linear model or logistic regression):

$$P_{Y|X}(y|x) = \frac{\exp\big(w^\top f(y,x)\big)}{\sum_{y' \in \mathcal{Y}} \exp(w^\top f(y',x))} \ .$$

Answer the questions below.

(a) (½ point) Define $f(y, x)$ and explain its role.

(b) (½ point) Define $w$ and explain its role.

(c) (½ point) What is the role of the dot product in the numerator of the definition?

(d) (½ point) What is the role of the exponential function in the numerator of the definition?

(e) (½ point) What is the role of the summation in the denominator?

(f) (½ point) What is the derivative of the numerator with respect to a parameter $w_d$?

(g) (1 point) What is the derivative of the denominator with respect to a parameter $w_d$?

(h) (1 point) Where $\mathcal{Z}(x|w)$ is the denominator, express the derivative of $\log \mathcal{Z}(x|w)$ with respect to a parameter $w_d$ in terms of the probability distribution $P_{Y|X}$ and the feature function $f$? Make sure your final answer is expressed in terms of $P_{Y|X}$ and $f$ and explain every step.

(i) (½ point) What is the log likelihood of a datapoint $(y, x)$?

(j) (½ point) What is the derivative of the log likelihood of a datapoint $(y, x)$ with respect to a parameter $w_d$?

(k) (2 points) How can we find a maximum likelihood estimate for $w_d$? Explain the global strategy including the reason for it being valid. If your answer involves updates to $w$, give a formula and explain all of its elements.

# 3 Applications

5. We have discussed improvements to HMMs where we have emission distributions condition both on the current class and the previously emitted word. This should improve the performance of the HMM as a language model, but has the downside of making the conditioning context available for emissions much sparser. We have seen 3 techniques to deal with conditionals based on high-dimensional data, namely, interpolation (used for in the notebook for the written assignment), naive Bayes, and logistic regression. Design emission distributions $P_{X|X_{\text{prev}}C}$ using these techniques. For this exercise consider that $X$ and $X_{\text{prev}}$ take on values in a vocabulary $\Sigma$ containing $v$ words and $C$ takes on values in a vocabulary $\mathcal{C}$ containing $t$ POS tags.

   (a) (1 point) First design it via interpolation, and consider $0 < \alpha < 1$ a fixed interpolation coefficient. Also provide MLE solutions (no need for smoothing).

   (b) (2 points) Now design it via naive Bayes classification. Also provide MLE solutions (no need for smoothing).

(c) (2 points) Now design it via logistic regression. There is not need for showing the MLE, but do discuss the design of a feature function for this problem: consider English as the language of interest and motivate your choices.

(d) (2 points) Discuss pros and cons of each approach and which approach would you use if you were to condition on yet more context for emissions. Also discuss why you would make your choice that way.

## Assessment

| Question | Points | Score |
|---|---|---|
| 1 | 1 | |
| 2 | 1 | |
| 3 | 2 | |
| 4 | 8 | |
| 5 | 7 | |
| Total: | 19 | |