Data Mining 2020
Exercises Naive Bayes and Logistic Regression

Exercise 1: Multinomial Naive Bayes for Text Classification

You are given the following collection of recipes with corresponding kitchen:

recipeID	words in recipe	Kitchen
r1	`spaghetti tomato minced meat`	Italian
r2	`spaghetti gorgonzola eggplant zucchini`	Italian
r3	`spaghetti olive oil garlic`	Italian
r4	`feta tomato olive oil`	Greek
r5	`yogurt cucumber garlic`	Greek

(a) Estimate $P(\texttt{spaghetti} \mid \text{Italian})$, $P(\texttt{spaghetti} \mid \text{Greek})$, $P(\texttt{yogurt} \mid \text{Italian})$, and $P(\texttt{yogurt} \mid \text{Greek})$ according to the multinomial naive Bayes model for text classification. *Use Laplace smoothing.*

(b) Estimate the class prior probabilities $P(\text{Greek})$ and $P(\text{Italian})$.

(c) Use the probabilities estimated at (a) and (b) to compute $P(\text{Italian} \mid \text{r6})$ according to the multinomial naive Bayes model, with r6: `spaghetti yogurt`.

Exercise 2: Naive Bayes for Text Classification

You are given the following collection of song lyrics with corresponding music genre:

songID	words in lyrics	Genre
s1	`shake ya fanky fanky ya ya`	Funk
s2	`shake baby shake`	Funk
s3	`fire hell thunder hell`	Metal
s4	`blood hell venom`	Metal
s5	`hell ya burn`	?

Here s1-s4 are the training examples, and s5 is a test example with unknown class label.

(a) Use s1-s4 to estimate $P(\texttt{ya} \mid \text{Funk})$ and $P(\texttt{ya} \mid \text{Metal})$ according to the multinomial Naive Bayes model. Use Laplace smoothing.

(b) Compute $P(\text{Funk} \mid \text{s5})$ en $P(\text{Metal} \mid \text{s5})$ according to the multinomial Naive Bayes model. Use Laplace smoothing.

Exercise 3: Logistic Regression

We analyse data of professional darts games in order to predict the winner. Only games of type "best of x legs" have been taken into account. In the model, the probability that the player who begins, player a, wins against player b depends on the difference in average and checkout percentage between the two players, and a constant β_0:

$$P(a \text{ wins against } b) = \frac{\exp(\beta_0 + \beta_1(\text{Av}_a - \text{Av}_b) + \beta_2(\text{Check}_a - \text{Check}_b))}{1 + \exp(\beta_0 + \beta_1(\text{Av}_a - \text{Av}_b) + \beta_2(\text{Check}_a - \text{Check}_b))}$$

Here Av_x denotes the average of player x, Check_x the checkout percentage of player x, and a denotes the player who starts throwing in the first leg. Estimation by maximum likelihood yields the following results:

Coefficient	Estimate
β_0 (Intercept)	0.120
β_1	0.135
β_2	0.025

Answer the following questions:

(a) How big is the advantage of the right to start the game according to the fitted model?

(b) Do the signs of the coefficient estimates make sense? Explain.

(c) Michael van Gerwen has average respectively checkout percentage of 102.7 and 46.2%. Vincent van de Voort has average respectively checkout percentage of 92.6 and 40.4%. According to the model, what is the probability that Michael van Gerwen wins against Vincent van de Voort if van Gerwen starts?

(d) What if van de Voort starts?

(e) Give the linear classification rule corresponding to this model that predicts the player most likely to win.