Data Mining 2020
Naive Bayes and Logistic Regression

Exercise 1: Multinomial Naive Bayes for Text Classification

(a) The vocabulary consists of:

1. `spaghetti`
2. `tomato`
3. `minced`
4. `meat`
5. `gorgonzola`
6. `eggplant`
7. `zucchini`
8. `olive`
9. `oil`
10. `garlic`
11. `feta`
12. `yogurt`
13. `cucumber`

Hence, $|V| = 13$. The total number of words in Italian recipes is 12, and in Greek recipes 7. Hence, we get:

$$\hat{P}(\texttt{spaghetti} \mid \text{Italian}) = \frac{3+1}{12+13} = \frac{4}{25}$$
$$\hat{P}(\texttt{spaghetti} \mid \text{Greek}) = \frac{0+1}{7+13} = \frac{1}{20}$$
$$\hat{P}(\texttt{yogurt} \mid \text{Italian}) = \frac{0+1}{12+13} = \frac{1}{25}$$
$$\hat{P}(\texttt{yogurt} \mid \text{Greek}) = \frac{1+1}{7+13} = \frac{2}{20}$$

(b)

$$\hat{P}(\text{Greek}) = \frac{2}{5} \qquad \hat{P}(\text{Italian}) = \frac{3}{5}$$

(c)

$$\begin{aligned}\hat{P}(\text{Italian} \mid \texttt{spaghetti yogurt}) &\propto \hat{P}(\texttt{spaghetti} \mid \text{Italian})\hat{P}(\texttt{yogurt} \mid \text{Italian})\hat{P}(\text{Italian}) \\ &= \left(\frac{4}{25}\right)\left(\frac{1}{25}\right)\left(\frac{3}{5}\right) = \frac{12}{3125}\end{aligned}$$

$$\begin{aligned}\hat{P}(\text{Greek} \mid \texttt{spaghetti yogurt}) &\propto \hat{P}(\texttt{spaghetti} \mid \text{Greek})\hat{P}(\texttt{yogurt} \mid \text{Greek})\hat{P}(\text{Greek}) \\ &= \left(\frac{1}{20}\right)\left(\frac{2}{20}\right)\left(\frac{2}{5}\right) = \frac{4}{2000}\end{aligned}$$

$$\hat{P}(\text{Italian} \mid \texttt{spaghetti yogurt}) = \frac{12/3125}{12/3125 + 4/2000} \approx 0.66$$

Exercise 2: Naive Bayes for Text Classification

(a) $P(\texttt{ya} \mid \text{Funk}) = \frac{2}{9}$ and $P(\texttt{ya} \mid \text{Metal}) = \frac{1}{16}$.

(b) $P(\text{Funk} \mid \text{s5}) = \frac{64}{145} \approx 0.44$ and $P(\text{Metal} \mid \text{s5}) \approx 0.56$.

Exercise 3: Logistic Regression

(a) If the players have the same average and checkout percentage, then the probability that the player who begins wins is:

$$\frac{e^{0.12}}{1 + e^{0.12}} = 0.53.$$

Hence the advantage is 6 percentage points (53% against 47%).

(b) Yes. For example, β_1 is positive which means that the bigger the difference in average in a's favor, the more likely it is that a will win the game. This is in accordance with common sense.

(c) The difference in average is $102.7 - 92.6 = 10.1$, and the difference in checkout percentage is $46.2 - 40.4 = 5.8$. Hence the probability that van Gerwen wins is:

$$\frac{\exp(0.12 + 0.135 \times 10.1 + 0.025 \times 5.8)}{1 + \exp(0.12 + 0.135 \times 10.1 + 0.025 \times 5.8)} \approx 0.84$$

So approximately 84%.

(d) The probability that van de Voort wins is:

$$\frac{\exp(0.12 + 0.135 \times -10.1 + 0.025 \times -5.8)}{1 + \exp(0.12 + 0.135 \times -10.1 + 0.025 \times -5.8)} \approx 0.20$$

So the probability that van Gerwen wins is approximately 80%.

(e) If

$$0.135 \times (\mathrm{Av}_a - \mathrm{Av}_b) + 0.025 \times (\mathrm{Check}_a - \mathrm{Check}_b)) > -0.12$$

then player a wins, otherwise player b wins.