Data Mining 2025
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) We want to calculate

$$\begin{aligned}
&\hat{P}(\text{Funk} \mid \text{s5}) \\
&= \frac{\hat{P}(\text{Funk}, \text{s5})}{\hat{P}(\text{s5})} && \text{(definition of conditional probability)} \\
&= \frac{\hat{P}(\text{Funk})\hat{P}(\text{s5} \mid \text{Funk})}{\hat{P}(\text{s5})} && (\hat{P}(A, B) = \hat{P}(B)\hat{P}(B|A)) \\
&= \frac{\text{prior} \times \text{likelihood}}{\text{(marginal) probability of evidence}} \\
&\propto \hat{P}(\text{Funk})\hat{P}(\text{s5} \mid \text{Funk}) && \text{(We don't need to compute } \hat{P}(\text{s5}) \text{ (yet).)} \\
&= \hat{P}(\text{Funk}) \prod_{w \in \text{s5}} \hat{P}(w \mid \text{Funk}) && \text{(Naive Bayes assumption)} \\
&= \hat{P}(\text{Funk}) \prod_{w_i \in \text{s5} \cap V} \frac{\text{count}(w_i, \text{Funk}) + 1}{\sum_{w_j \in V} \text{count}(w_j, \text{Funk}) + |V|} && \text{(smoothed probabilities)} \\
&= \hat{P}(\text{Funk}) \prod_{w_i \in \text{s5} \cap V} \frac{\text{count of } w_i \text{ in class Funk} + 1}{\text{total count of all words in class Funk} + |V|} \\
&= \hat{P}(\text{Funk})\hat{P}(\texttt{hell} \mid \text{Funk})\hat{P}(\texttt{ya} \mid \text{Funk}) && (\texttt{burn} \notin V\text{, so ignore it)} \\
&= \frac{1}{2} \times \frac{0+1}{9+9} \times \frac{3+1}{9+9} = \frac{1}{162} \\
&= \hat{P}(\text{Funk}, \text{s5}) && \text{(Reminder: not done yet! This isn't } \hat{P}(\text{Funk} \mid \text{s5}).)
\end{aligned}$$

The proportionality above holds because $\hat{P}(\text{s5})$ does not depend on the class. In general it is the case that a conditional probability is proportional to the corresponding joint probability: $P(A \mid B) = P(A, B) \times \frac{1}{P(B)} \propto P(A, B)$.

Now we know the joint probability of class "Funk" and document s5. Let's calculate the same for class "Metal":

$$\begin{aligned}
&\hat{P}(\text{Metal}, \text{s5}) \\
&= \hat{P}(\text{Metal})\hat{P}(\texttt{hell} \mid \text{Metal})\hat{P}(\texttt{ya} \mid \text{Metal}) && \text{(again ignore } \texttt{burn}\text{)} \\
&= \frac{1}{2} \times \frac{3+1}{7+9} \times \frac{0+1}{7+9} = \frac{1}{128}.
\end{aligned}$$

Now we only need to calculate the marginal probability of the evidence (of observing s5): $\hat{P}(\text{s5})$. We get this by marginalising over all (two) of the classes:

$$\hat{P}(\text{s5}) = \hat{P}(\text{Funk}, \text{s5}) + \hat{P}(\text{Metal}, \text{s5}) = \frac{1}{162} + \frac{1}{128} = \frac{145}{10368}.$$

Then, we can calculate our final answer:

$$\begin{aligned}
\hat{P}(\text{Funk} \mid \text{s5}) &= \frac{\hat{P}(\text{Funk}, \text{s5})}{\hat{P}(\text{s5})} \\
&= \frac{1}{162} \div \frac{145}{10368} \\
&= \frac{1}{162} \times \frac{10368}{145} \\
&= \frac{64}{145} \approx 0.44.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\hat{P}(\text{Metal} \mid \text{s5}) &= \frac{\hat{P}(\text{Metal}, \text{s5})}{\hat{P}(\text{s5})} \\
&= \frac{1}{128} \div \frac{145}{10368} \\
&= \frac{1}{128} \times \frac{10368}{145} \\
&= \frac{81}{145} \approx 0.56,
\end{aligned}$$

though of course we could have calculated this also as

$$\begin{aligned}\hat{P}(\text{Metal} \mid \text{s5}) &= 1 - \hat{P}(\text{Funk} \mid \text{s5}) \\ &= 1 - \frac{64}{145} \\ &= \frac{81}{145} \approx 0.56.\end{aligned}$$

If all we care about is prediction of the class with maximum probability, we don't actually even need to calculate $\hat{P}(\text{s5})$, since, as mentioned before, the joint probabilities are already proportional to the conditional ones (i.e. the class with highest joint probability is also the class with the highest conditional probability).

(b) $\hat{P}(\texttt{ya} \mid \text{Metal}) = \frac{0}{7} = 0$ and $\hat{P}(\texttt{hell} \mid \text{Funk}) = \frac{0}{9} = 0$. Hence both $\hat{P}(\texttt{hell} \mid \text{Metal})\hat{P}(\texttt{ya} \mid \text{Metal})$ and $\hat{P}(\texttt{hell} \mid \text{Funk})\hat{P}(\texttt{ya} \mid \text{Funk})$ are zero, which leads to division by zero when we try to compute $\hat{P}(\text{Funk} \mid \text{s5})$ and $\hat{P}(\text{Metal} \mid \text{s5})$.

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

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

then player a wins, otherwise player b wins.