Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Tutorial sheet 4: Naïve Bayes

Prof. Dr. Martin Bichler
Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise T4.1 *Cookie factory*

You are the manager of a company which produces cookies and you want to introduce a new product. Your R&D department has proposed and developed the following two alternatives:

1. Unicorn cookies (UC)
2. Vanilla-chip cookies (VC).

As part of your market research, you are interested in predicting whether certain customers are likely to buy one of the new products. For that, you have already collected data from a large number of test persons. In particular, you asked them to fill out a query with the following questions:

1. *What do you think is the most fascinating: Rainbows, Black holes or Cats?* (variable $preferences$)
2. *How much money do you spend on cookies per month?* (variable $money$)
3. *Which of our cookies would you buy?* (variable $product$)
   Note: The variable $product$ can also take on the value "No product" (NP).

You can find the data in *cookie-factory.csv*. We recommend using the provided notebook template to solve sub-tasks b) - e).

a) For each of the questions 1-3, decide
   (i) whether the answers are continuous or discrete outcomes,
   (ii) which range the outcomes could have

To infer which products new customers are likely to buy, you set up a probabilistic model. You assume that the answers to questions 1 - 2 are conditionally independent (Naive Bayes) given $product$ and model the dependencies as follows:

$$f(preferences, money, product) = \mathbb{P}(preferences \mid product) \cdot f_{money}(money \mid product) \cdot \mathbb{P}(product)$$

b) Estimate the parameters of your categorical prior $\mathbb{P}(product)$ by using maximum likelihood:

$$\mathbb{P}(product = UC) = p_{UC} \qquad \mathbb{P}(product = VC) = p_{VC} \qquad \mathbb{P}(product = NP) = p_{NP}$$

Hint: The maximum likelihood estimate of the parameters for categorically distributed variables is simply the fraction of samples from a category.

Based on your observations in a), you decide to model the likelihoods as follows:

1. $preferences$ follows a Categorical distribution where the parameters depend on the product the customers would buy:

$$\begin{aligned}
\mathbb{P}(preferences = \text{"Rainbows"} \mid prod. = \text{UC}) &= \pi_R^{UC} \\
\mathbb{P}(preferences = \text{"Black holes"} \mid prod. = \text{UC}) &= \pi_B^{UC} \\
\mathbb{P}(preferences = \text{"Cats"} \mid prod. = \text{UC}) &= \pi_C^{UC} \\
\\
\mathbb{P}(preferences = \text{"Rainbows"} \mid prod. = \text{VC}) &= \pi_R^{VC} \\
\mathbb{P}(preferences = \text{"Black holes"} \mid prod. = \text{VC}) &= \pi_B^{VC} \\
\mathbb{P}(preferences = \text{"Cats"} \mid prod. = \text{VC}) &= \pi_C^{VC} \\
\\
\mathbb{P}(preferences = \text{"Rainbows"} \mid prod. = \text{NP}) &= \pi_R^{NP} \\
\mathbb{P}(preferences = \text{"Black holes"} \mid prod. = \text{NP}) &= \pi_B^{NP} \\
\mathbb{P}(preferences = \text{"Cats"} \mid prod. = \text{NP}) &= \pi_C^{NP}
\end{aligned}$$

2. $money$ follows an exponential distribution where the parameter $\lambda_{product}$ depends on the product the customers would buy ($\eta_{product} = \eta_{UC}$, $\eta_{product} = \eta_{VC}$ or $\eta_{product} = \eta_{NP}$):

$$f_{money}(m|product) = \begin{cases} \eta_{product} \cdot e^{-\eta_{product} \cdot m} & m \geq 0 \\ 0 & \text{else} \end{cases}$$

Intuitively, your model describes the profile ($preferences$, $money$) of a customer if you already know which product they would buy ($product$).

c) Using the data, derive maximum likelihood estimates for all parameters.

Hint: The maximum likelihood estimate of the parameters for exponentially distributed variables is the inverse of their sample mean: $\bar{x}^{-1}$.

You now have access to a joint density over your data:

$$f(preferences, money, product) = \mathbb{P}(preferences \mid product) \cdot f_{money}(money \mid product) \cdot \mathbb{P}(product)$$

d) With the fitted model, predict the (posterior) probability

$$\mathbb{P}(product \mid preferences, money)$$

that the customers below buy a unicorn cookie, a vanilla-chip cookie or no cookie at all:

Customer	$preferences$	$money$
Anna	Cats	53.10 €
Ben	Rainbows	2.30 €
Caroline	Black holes	10.25 €

e) From a fourth customer, you only know that they like rainbows. Predict the probability that they buy unicorn cookies.

f) *[Bonus]* You may have noticed that the data also contains information about age. What would you need to do to include this information as well?

Exercise T4.2 *Sushi or Pizza*

Anna and Ben want to have dinner together, but they cannot decide if they prefer pizza or sushi. For this reason, they design the following mechanism to make a decision.

1. They both roll a dice (fair, six-sided), obtaining events A (Anna's dice) and B (Ben's dice).
2. They determine the sum of the outcomes and note down if it is even or odd. (Event $C \in \{"e", "o"\}$)
3. They throw a biased coin which shows head with 80% probability and tails with 20%. If $C = "e"$, they opt for pizza ($"P"$) if head is shown and sushi ($"S"$) if tails is shown. If $C = "o"$, they opt for sushi if head is shown and pizza if tails is shown. (Event $D \in \{"P", "S"\}$)

a) Visualize the process as a Bayesian network.

b) Next to the nodes of the Bayesian network, note the likelihood tables of the conditional probabilities.

c) Write down the joint probability distribution in terms of A, B, C, and D.

d) Compute the probability that they decide on pizza, given that Anna's dice shows 2 and Ben's dice shows 3.

e) Are A and B independent? Are A and B *conditionally* independent given C?