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, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise T4.1 *States of a car*

Consider the following Boolean random variables related to the state of a given car.

- Battery (B):	is the battery charged?
- Fuel (F):	is the fuel tank filled?
- Ignition (I):	does the ignition system work?
- Moves (M):	does the car move?
- Radio (R):	is the radio working?
- Starts (S):	does the engine work?

a) Represent the joint probability density function using a Bayesian Network.

b) Express the joint probability $\mathbb{P}(B, F, I, M, R, S)$ as a product of conditional probabilities using the chain rule.

Exercise T4.2 *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. How old are you? (variable age)

2. What do you think is the most fascinating: Rainbows, Black holes or Cats? (variable $preferences$)

3. How much money do you spend on cookies per month? (variable $money$)

4. 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 1-4, decide

   (i) whether the answers are continuous or discrete outcomes,

(ii) which range the outcomes could have,

(iii) to which scale of measurement (nominal, ordinal, interval, ratio) the outcomes belong to.

b) To infer which products new customers are likely to buy, you set up a probabilistic model. You assume that the answers to questions 1 - 3 are conditionally independent (Naive Bayes) given $product$ and model the dependencies as follows:

$$\begin{aligned} &f(age, preferences, money, product) = \\ &\qquad \mathbb{P}(age \mid product) \cdot \mathbb{P}(preferences \mid product) \cdot f_{money}(money \mid product) \cdot \mathbb{P}(product) \end{aligned}$$

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.

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

1. age follows a Poisson distribution where the parameter $\lambda_{product}$ depends on the product the customers would buy ($\lambda_{product} = \lambda_{UC}$, $\lambda_{product} = \lambda_{VC}$, or $\lambda_{product} = \lambda_{NP}$):

$$\mathbb{P}(age = k|product) = \frac{\lambda_{product}^{k}}{k!} e^{-\lambda_{product}}$$

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

$$\begin{aligned} \mathbb{P}(preferences = \text{"Rainbows"} \quad &| \; prod. = \text{UC}) = \pi_{UC}^{R} \\ \mathbb{P}(preferences = \text{"Black holes"} \; &| \; prod. = \text{UC}) = \pi_{UC}^{B} \\ \mathbb{P}(preferences = \text{"Cats"} \qquad &| \; prod. = \text{UC}) = \pi_{UC}^{C} \end{aligned}$$

$$\begin{aligned} \mathbb{P}(preferences = \text{"Rainbows"} \quad &| \; prod. = \text{VC}) = \pi_{VC}^{R} \\ \mathbb{P}(preferences = \text{"Black holes"} \; &| \; prod. = \text{VC}) = \pi_{VC}^{B} \\ \mathbb{P}(preferences = \text{"Cats"} \qquad &| \; prod. = \text{VC}) = \pi_{VC}^{C} \end{aligned}$$

$$\begin{aligned} \mathbb{P}(preferences = \text{"Rainbows"} \quad &| \; prod. = \text{NP}) = \pi_{NP}^{R} \\ \mathbb{P}(preferences = \text{"Black holes"} \; &| \; prod. = \text{NP}) = \pi_{NP}^{B} \\ \mathbb{P}(preferences = \text{"Cats"} \qquad &| \; prod. = \text{NP}) = \pi_{NP}^{C} \end{aligned}$$

3. $money$ 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 (age, $preferences$, $money$) of a customer if you already know which product they would buy ($product$).

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

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

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

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

With the fitted model, predict the (posterior) probability

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

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

Customer	$preferences$	age	$money$
Anna	Cats	81	53.10 €
Ben	Rainbows	15	2.30 €
Caroline	Black holes	42	10.25 €

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