W3WI_DS304.1 Applied Machine Learning Fundamentals

Exercise Sheet # 2 - Decision Theory and Probability Density Estimation

Question 1 EX 2021, 2022 (Bayes' theorem)

Write down Bayes' theorem and name its components!

Question 2 (Probability theory, Bayes' theorem)

Consider the following three statements:

❶ A person with a cold has back pain 25 % of the time.

❷ 4 % of the world population has a cold.

❸ 15 % of those who do not have a cold, still have back pain.

Please answer the following questions:

1. Identify random variables from the statements above and define a unique symbol for each of them. Also, define the domain of each random variable.

2. Represent the three statements above using random variables you have defined.

3. If you suffer from back pain, what are the chances that you suffer from a cold? *(Please show all the intermediate steps of your calculations.)*

Question 3 EX 2022 (Multinomial naïve Bayes algorithm)

You work at a mail-order company. Your task is to train a classifier which decides whether or not a given customer is allowed to purchase on account. You decide to use a naïve Bayes classifier. Table 1 contains historical data which can be used for training. Each customer is described by the four features

- **customer type (T)**,
- **payment speed (PS)**,
- **purchase frequency (PF)**, and
- **origin (O)**.

The last column called **invoice** indicates whether or not a purchase on account was granted for that specific customer.

Customer type	Payment speed	Purchase frequency	Origin	Invoice
new customer	high	low	national territory	**yes**
new customer	low	low	abroad	**no**
new customer	low	high	national territory	**no**
existing customer	low	high	national territory	**yes**
new customer	high	low	national territory	**no**
existing customer	high	high	abroad	**yes**
existing customer	high	high	national territory	**yes**

Table 1: Training dataset containing historical customer data.

Please answer the following questions:

1. This morning a customer with the attributes

$$\boldsymbol{x}_q = \big[\text{new customer, high, low, abroad}\big]$$

has placed an order. Can you grant a purchase on account? Please calculate all probabilities necessary and compute the final prediction. Why is the sum of the probabilities you have computed not equal to one?

2. Another customer orders your products in the afternoon. This customer is described by the following attributes

$$\widetilde{\boldsymbol{x}}_q = \big[\text{regular customer, low, high, abroad}\big].$$

What problem do you face when feeding the data into your model? How can you mitigate this issue? *(You don't have to compute the prediction.)*

Question 4 EX 2021 (Bayes optimal classifiers)

Briefly explain the concept of **Bayes optimality**. Is every classifier Bayes optimal?

Question 5 EX 2021

What does the abbreviation **i.i.d.** stand for? What does it mean?

Question 6 (Maximum likelihood estimation for the Gaussian distribution)

This exercise is not easy and is therefore optional! In the lecture we have seen that given n data points $x^{(1)}, \dots, x^{(n)}$, the maximum likelihood estimators for the parameters of the Gaussian distribution μ and σ^2 are computed according to

$$\mu_{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \qquad \sigma^2_{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n}\big(x^{(i)} - \mu_{\mathrm{ML}}\big)^2.$$

You can find the proof for μ_{ML} in the lecture slides. Please show that the formula for σ^2_{ML} is indeed the maximum likelihood solution.

Hint: Use the log-likelihood function

$$\log p\big(\boldsymbol{X}|\mu, \sigma^2\big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(x^{(i)} - \mu\big)^2$$

and compute the partial derivative with respect to σ^2. Set the derivative to zero and subsequently solve for σ^2. Don't forget to verify that the solution you have found is indeed a local maximum of the likelihood function!

Question 7 EX 2021 (Gaussian naïve Bayes algorithm)

You are given a training dataset $\mathcal{D}$ consisting of the two classes O (orange) and G (gray):

$$\begin{aligned}\mathcal{D} = \big\{&(0.0, O), (1.7, O), (2.0, O), (2.3, O),\\ &(3.1, G), (3.5, G), (4.0, G), (4.3, G), (5.0, G), (5.3, G)\big\}\end{aligned}$$

Let $\boldsymbol{X}_O := \{0.0, 1.7, 2.0, 2.3\}$ and $\boldsymbol{X}_G := \{3.1, 3.5, 4.0, 4.3, 5.0, 5.3\}$ denote the class partitions of the dataset. The data is depicted in the following figure 1:

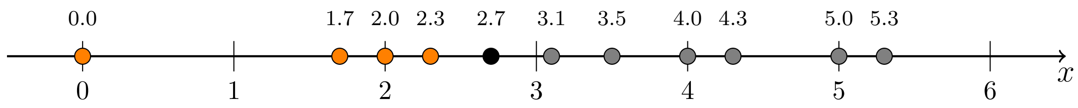

Figure 1: Plot of the training dataset consisting of the two classes O and G.

Compute the most probable class label y_q of the data point $x_q = 2.7$ (black dot in figure 1) using the Gaussian naïve Bayes algorithm.

Hint: Estimate μ_O, σ^2_O, μ_G, and σ^2_G. Use the results to compute the respective class conditional probabilities and plug them into Bayes' theorem to get the class posterior probabilities!

Question 8 (Maximum likelihood estimation for the exponential distribution)

The probability density function of the **exponential distribution** $f(x|\lambda)$ which is governed by the single parameter $0 < \lambda \in \mathbb{R}$ is given by

$$f(x|\lambda) := \begin{cases}\lambda e^{-\lambda x} & x \geq 0\\ 0 & x < 0.\end{cases}$$

Please answer the following questions:

1. Verify that $f(x|\lambda)$ is indeed a valid probability density function, i. e. show that
   - $f(x|\lambda) \geq 0 \ \forall x \in \mathbb{R}$ and
   - $\int_{-\infty}^{\infty} f(x|\lambda)\ \mathrm{d}x = 1$.

2. Suppose you are given a set of n data points

$$\boldsymbol{X} = \left\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\right\}.$$

Find the maximum likelihood solution λ_{ML} for the parameter λ!

Question 9 (Implement the multivariate Gaussian naïve Bayes algorithm)

Generate a training dataset consisting of $K = 2$ classes using the following Python snippet:

```python
from sklearn.datasets import make_classification

# DATA CREATION
# ---------------------------------------------------------------
# number of classes
n_classes = 2

X, y = make_classification(
    n_samples=200,
    n_features=2,
    n_redundant=0,
    n_informative=2,
    n_clusters_per_class=1,
    n_classes=n_classes,
    class_sep=3.25,
    random_state=42
)
```

Figure 2 depicts the training dataset generated by the above code snippet:

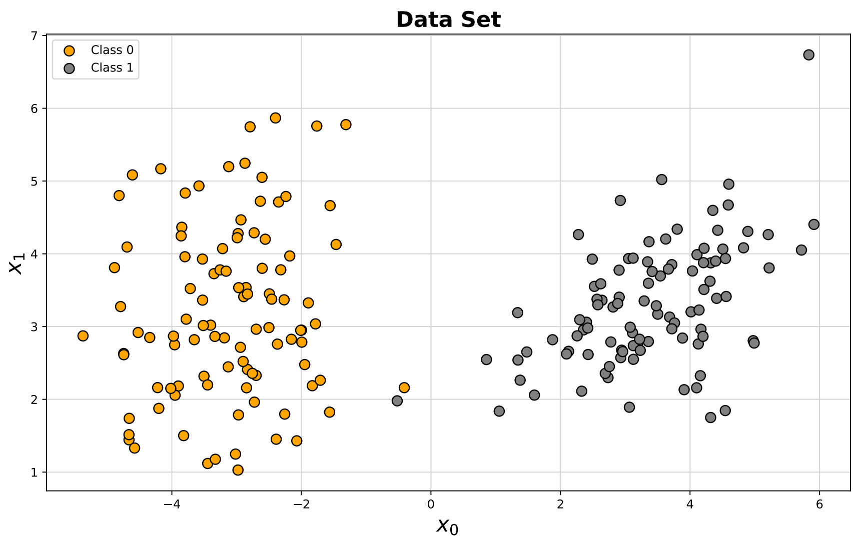

Figure 2: Training dataset generated by the code snippet.

Estimate the class prior probabilities as well as the class conditional probabilities, i. e. the mean vectors $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and the covariance matrices $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ of the (multivariate) Gaussian distributions for both classes 0 and 1. Plot the posterior distributions and the decision boundary similiary to the plot shown in the lecture slides.

Hints:

- The maximum likelihood solution for the parameters of the multivariate Gaussian distribution $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by:

$$\boldsymbol{\mu}_{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)} \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{\mathrm{ML}}\big)\big(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{\mathrm{ML}}\big)^{\intercal}$$

- To evaluate the multivariate normal distribution you may want to use the `scipy` package: `from scipy.stats import multivariate_normal`.