University of Stuttgart | ANALYTIC COMPUTING

Foundations of Machine Learning - Exercise (SS 25)

Assignment 5: Naïve Bayes

Arvindh Arunbabu
arvindh.arun@ki.uni-stuttgart.de

Nadeen Fathallah
Nadeen.Fathallah@ki.uni-stuttgart.de

Cosimo Gregucci
cosimo.gregucci@ki.uni-stuttgart.de

Akram Sadat Hosseini
Akram.Hosseini@ki.uni-stuttgart.de

Farane Jalali
farane.jalali-farahani@ki.uni-stuttgart.de

Osama Mohammed
osama.mohammed@ki.uni-stuttgart.de

Jiaxin Pan
jiaxin.pan@ki.uni-stuttgart.de

Tim Schneider
tim.schneider@ki.uni-stuttgart.de

Jingcheng Wu
jingcheng.wu@ki.uni-stuttgart.de

Daniel Frank
daniel.frank@ki.uni-stuttgart.de

This assignment consists of 6 pages and the `Gaussian Naive Bayes` notebook. with 4 tasks:

Submit your theoretical solution in ILIAS as a single PDF file.[1] Make sure to list the full names of all participants, matriculation number, study program, and B.Sc. or M.Sc on the first page. Optionally, you can *additionally* upload source files (e.g., PPTX files). Submit your programming task in ILIAS as a single Jupyter notebook. If you have any questions, feel free to ask them in the exercise forum in ILIAS.

Submission is open until Monday, 26.05.2025, 12:00 noon.

[1] Your drawing software probably allows exporting as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

Task 1: Binary Features and Bayes Independence

Let X_1, X_2, C be binary random variables, i.e. $\text{dom}(X_1) = \text{dom}(X_2) = \text{dom}(C) = \{0, 1\}$. Let furthermore be p a probability distribution over the joint domain.

Complexity

1. **Task** Assume C is a class to be determined. What is the optimal decision given features X_1, X_2?
2. **Task** Consider the extension to n binary (feature) variables $X_1, \ldots, X_n$. How many parameters are required to represent the likelihood $p(X_1, \ldots, X_n | C)$ as a table?
3. **Task** In constrast, which computations are performed by a *Naïve Bayes* classifier? Which probability distributions need to be inferred from the data? How many parameters do the tables require?

Independence

The independence of random variables A, B can be alternatively denoted as $A \perp B$, or conditional independence (given a random variable C) as $A \perp B \mid C$.

1. **Task** List all the independence assumptions of a Naïve Bayes classifier with two features X_1, X_2.
2. **Task** Does $X_1 \perp X_2 \implies X_1 \perp X_2 \mid C$ hold? Prove it or give a counter-example.
3. **Task** Does $X_1 \perp X_2 \mid C \implies X_1 \perp X_2$ hold? Prove it or give a counter-example.

Applications

For the following scenarios: What can you say about the bias error of a Naïve Bayes classifier?

1. **Task** Predicting Hospital Admission.

You are designing a model to classify whether a patient will be admitted to the hospital (class C) based on whether the patient has diabetes X_1 and gallbladder disease (X_2). Epidemiological surveys suggest that diabetes and gallbladder disease occur largely independently. However, either condition on its own can lead to hospital admission, so most records in the hospital database show at least one of the two illnesses. Within that hospital subset, doctors notice that a patient who has diabetes is actually *less* likely to have gallbladder disease, and vice versa. Conditioning on hospitalization creates a negative correlation between diabetes and gallbladder disease because if a patient is hospitalized and one condition is absent, it increases the likelihood that the other must be present to explain the admission.

2. **Task** Detecting Flu Infection.

Public health analysts aim to classify individuals as influenza-positive or not (class C), using the presence of the following two symptoms: features: fever (X_1) and headache (X_2). City-wide surveillance reports that fever and headache frequently appear together during flu season. However, physicians explain that influenza itself causes each symptom separately in many patients; one symptom does not physiologically cause the other. Imagine you focus only on patients confirmed to have the flu. Within that group, seeing a fever gives little extra information about whether a headache is also reported. Based on this narrative, decide whether X and Z are independent marginally, conditionally on Y, both, or neither.

3. **Task** Inferring Sex from Traits.

You are tasked with predicting a survey respondent's sex (class C) using the results from the ABO blood type test (X_1). On the other hand, the Big-Five personality test is a psychological model used to describe human personality. It is one of the most widely researched and accepted models in psychology. Its extraversion score(X_2) is used as a second feature. Large-scale studies have found no statistical relationship between blood type and any Big-Five trait in the overall population. Researchers then separate the data by sex and repeat the analysis within males and females. Again, no clear association is found. The ABO blood group is determined by the ABO gene, located on chromosome 9, which is not a sex chromosome.

4. **Task** Classifying Age Group from Health Metrics.

An insurance company is training a model to classify applicants into one of several age bands (class C) using two continuous features recorded during health check-ups: the systolic blood pressure (X_1) and the cholesterol concentration (X_2). When the company analyzes the entire dataset, it finds that people with higher blood pressure also tend to have higher cholesterol values. They then divide the dataset into five-year age brackets and look again. Within every age bracket, the positive correlation between blood pressure and cholesterol remains strong.

Task 2: Flu Detection with Naïve Bayes

A hospital analyses 60 emergency-room cases to train a Naïve Bayes classifier.

Table 1 Training counts for the categorical symptoms.

	Fever		Cough		Cases
Class	Yes	No	Yes	No	total
Flu	24	6	21	9	30
No Flu	8	22	12	18	30

Table 2 Class-conditional temperature (assumed Gaussian).

Class	mean μ_{Temp} (°C)	variance σ^2
Flu	38.0	0.25
No Flu	36.8	0.16

The empirical class priors are therefore $P(\text{Flu}) = P(\text{NoFlu}) = 0.5$. For all categorical likelihoods, use **Laplace smoothing with** $\lambda = 1$.

A new patient presents with **Fever = Yes, Cough = No, Temperature = 37.8 °C**.

1. **Task** Using only *Fever* and *Cough*, apply Naïve Bayes with $\lambda = 1$ to compute the posterior probability of Flu for the new patient and state the predicted class.
2. **Task** Using the Gaussian temperature parameters, evaluate the likelihoods at 37.8°C; combine these with the categorical evidence from the previous step, then report the updated posterior probability of Flu and the resulting class prediction. Briefly comment on the outcome.
3. **Task** In the previous sub-task, the classifier predicted Flu. For which new priors will the Naïve Bayes classifier change its decision to No Flu?
4. **Task** The thermometer fails, so no temperature is recorded. Explain how you would classify the patient and briefly state how omitting this feature affects the posterior probabilities.

Task 3: Height Classification with Naïve Bayes

The dataset below records the height of 14 people together with the binary class *Athlete*.

Table 3 Athlete Prediction Dataset

Inst.	Height	Athlete?
A	180	Yes
B	172	No
C	178	Yes
D	165	No
E	160	No
F	170	No
G	185	Yes
H	158	No
I	182	Yes
J	168	No
K	177	Yes
L	162	No
M	186	Yes
N	164	Yes

The training set, therefore, contains seven **Yes** and seven **No** instances.

Consider a new person of height **174** cm.

1. **Task** Using a Gaussian likelihood for the numerical feature *Height*, determine the Naïve Bayes posterior probability that the person is an athlete and state the predicted class.
2. **Task** Create *three* equal-width height bins of your choice and treat *Height* as categorical. With Laplace smoothing ($\lambda = 1$) compute the Naïve Bayes posterior for the same 174 cm person and state the predicted class.
3. **Task** Do the two models yield different classes? Quote both posteriors and, if they differ, explain why.

Task 4: Guassian Naïve Bayes with on the Boston Housing Dataset

Follow the instructions in the jupyter notebook.