MADS-ML – Machine Learning

Naive Bayes

Prof. Dr. Stephan Doerfel

FACHHOCHSCHULE KIEL
University of Applied Sciences

Moodle (WiSe 2024/25)

"For events A and B with $P(B) > 0$,

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A) \cdot \mathrm{P}(B|A)}{\mathrm{P}(B)}$$

holds."

Bayes' Theorem

Outline

Sample Spaces and Events

Definition 1 (Sample Space, Event)

The **sample space** Ω of an experiment is the set of all possible outcomes. The elements of Ω are called **sample outcomes**. Subsets of Ω are called **events**.

Sample Spaces and Events

Definition 1 (Sample Space, Event)

The **sample space** Ω of an experiment is the set of all possible outcomes. The elements of Ω are called **sample outcomes**. Subsets of Ω are called **events**.

Example: Role a regular die.

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Events: e.g.
  - "Role an even number" $\{2, 4, 6\}$
  - "Role a 3" $\{3\}$
  - "Role a 1 or a 4" $\{1, 4\}$

Probability Distribution

Definition 2 (Probability Distribution)

Given a sample space Ω, a function $P : 2^{\Omega} \rightarrow \mathbb{R}$ is called a **probability distribution** if the following are true:

1. for $A \subseteq \Omega : P(A) \geq 0$
2. $P(\Omega) = 1$
3. If in a set of events $\{A_i \mid i \in I\}$, the A_i are pairwise disjoint, then

$$P(\bigcup_{i \in I} A_i) = \sum_{i \in I} P(A_i).$$

- Note that the index set I can be an infinite set!
- This definition is actually too narrow. A broader version leads to the mathematical theory of measures ...

Probability Distribution – Example

Role a (fair) die as in the example before:

- $P(\Omega) = 1$
- $P(\{2, 4, 6\}) = \frac{1}{2}$
- $P(\{3\}) = \frac{1}{6}$

Toss a (fair) coin three times:

- $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
- $P(3xH) = \frac{1}{8}$
- $P(2xH) = \frac{3}{8}$
- $P(\geq 2xH) = \frac{1}{2}$

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"at least 1 heads": $P(B) = .75$

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"at least 1 heads": $P(B) = .75$
  - $P(A \wedge B) = 0.5 \neq 0.5 \cdot 0.75$ (not independent)

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"at least 1 heads": $P(B) = .75$
  - $P(A \wedge B) = 0.5 \neq 0.5 \cdot 0.75$ (not independent)
- Example 2:

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"at least 1 heads": $P(B) = .75$
  - $P(A \wedge B) = 0.5 \neq 0.5 \cdot 0.75$ (not independent)
- Example 2:
  - Event $A =$"first throw is heads": $P(A) = .5$

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"at least 1 heads": $P(B) = .75$
  - $P(A \wedge B) = 0.5 \neq 0.5 \cdot 0.75$ (not independent)
- Example 2:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"second throw is heads": $P(B) = .5$

Independence

Definition 3 (Independence)

Two events $A, B \in \Omega$ are called **independent** if $P(A \wedge B) = P(A) \cdot P(B)$. As set of events $\{A_i \mid i \in I\}$ is independent if

$$P(\bigwedge_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

Example: Throw a (fair) coin twice.

- Example 1:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"at least 1 heads": $P(B) = .75$
  - $P(A \wedge B) = 0.5 \neq 0.5 \cdot 0.75$ (not independent)
- Example 2:
  - Event $A =$"first throw is heads": $P(A) = .5$
  - Event $B =$"second throw is heads": $P(B) = .5$
  - $P(A \wedge B) = 0.25 = 0.5 \cdot 0.5$ (independent)

Conditional Probability

Definition 4 (Conditional Probability)

Given two events $A, B \in \Omega$ with $P(B) > 0$. The **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}.$$

Conditional Probability

Definition 4 (Conditional Probability)

Given two events $A, B \in \Omega$ with $P(B) > 0$. The **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}.$$

Example: Throw a (fair) coin three times.

Conditional Probability

Definition 4 (Conditional Probability)

Given two events $A, B \in \Omega$ with $P(B) > 0$. The **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}.$$

Example: Throw a (fair) coin three times.

- $A =$"exactly two heads": $P(A) = \frac{3}{8}$

Conditional Probability

Definition 4 (Conditional Probability)

Given two events $A, B \in \Omega$ with $P(B) > 0$. The **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}.$$

Example: Throw a (fair) coin three times.

- $A =$"exactly two heads": $P(A) = \frac{3}{8}$
- $B =$"first throw is heads": $P(B) = \frac{1}{2}$

Conditional Probability

Definition 4 (Conditional Probability)

Given two events $A, B \in \Omega$ with $P(B) > 0$. The **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}.$$

Example: Throw a (fair) coin three times.

- $A =$"exactly two heads": $P(A) = \frac{3}{8}$
- $B =$"first throw is heads": $P(B) = \frac{1}{2}$
- $P(A \mid B) = \frac{\frac{2}{8}}{\frac{1}{2}} = \frac{1}{2}$

Independence and Conditional Probabilities

Lemma 5

If two events $A, B \in \Omega$ are independent and $P(B) > 0$, then $P(A \mid B) = P(A)$.

Independence and Conditional Probabilities

Lemma 5

If two events $A, B \in \Omega$ are independent and $P(B) > 0$, then $P(A \mid B) = P(A)$.

Example: Throw a (fair) coin three times.

Independence and Conditional Probabilities

Lemma 5

If two events $A, B \in \Omega$ are independent and $P(B) > 0$, then $P(A \mid B) = P(A)$.

Example: Throw a (fair) coin three times.

- $A =$"second throw is heads": $P(A) = \frac{1}{2}$

Independence and Conditional Probabilities

Lemma 5

If two events $A, B \in \Omega$ are independent and $P(B) > 0$, then $P(A \mid B) = P(A)$.

Example: Throw a (fair) coin three times.

- $A =$"second throw is heads": $P(A) = \frac{1}{2}$
- $B =$"first throw is heads": $P(B) = \frac{1}{2}$

Independence and Conditional Probabilities

Lemma 5

If two events $A, B \in \Omega$ are independent and $P(B) > 0$, then $P(A \mid B) = P(A)$.

Example: Throw a (fair) coin three times.

- $A =$"second throw is heads": $P(A) = \frac{1}{2}$
- $B =$"first throw is heads": $P(B) = \frac{1}{2}$
- $P(A \mid B) = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$

Bayes' Theorem

Theorem 6 (Bayes' Theorem)

Given two events $A, B \in \Omega$ *with* $P(A), P(B) > 0$. *Then*

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}.$$

Random Variable

Definition 7 (Random Variable)

A **random variable** is a function $X : \Omega \to \mathbb{R}$.

Random Variable

Definition 7 (Random Variable)

A **random variable** is a function $X : \Omega \to \mathbb{R}$.

Example: Throw a (fair) coin three times. Count the number of heads.

The Point Mass Distribution

- There exists one element $a \in \Omega$ s.t. for $\omega \in \Omega$:

$$X(\omega) = \begin{cases} 1 \text{ if } \omega = a \\ 0 \text{ else.} \end{cases}$$

The Discrete Uniform Distribution

- Let $|X[\Omega]| = k$ (X takes k values) and $P(X = x)$ is either 0 or $\frac{1}{k}$.
- Each value of X has the same probability.

The Bernoulli Distribution

- X represents the number of heads in a single (biased) coin flip.
- $P(X = 1) = p$ and $P(X = 0) = 1 - p$.
- p is a (fix) parameter of the distribution.

The Binomial Distribution

- X represents the number of heads in n (independent) flips of a (biased) coin.
- $$P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x \in 0, 1, \dots, n \\ 0 \text{ else.} \end{cases}$$
- n and p are (fix) parameters of the distribution.
- generalization of Bernoulli distribution ($n = 1$)

The Categorical Distribution

- X represents the rolling of a k sided (biased) die.
- $$P(X = x) = \begin{cases} p_i \text{ for } i \in \{1, \dots, k\} \\ 0 \text{ else.} \end{cases}$$
- k and the p_i are (fix) parameters of the distribution with $\sum_{i=1}^{k} p_i = 1$.
- generalization of Bernoulli distribution ($k = 2$)

Outline

Naive Bayes – Basics

Which class has the maximum likelihood,
given the observed feature values?

Naive Bayes – Basics

Which class has the maximum likelihood,
given the observed feature values?

- Given an instance (feature tuple) $\boldsymbol{x}$, which class $c \in C$ has the highest likelihood to fit $\boldsymbol{x}$, i.e, for which class is it the most plausible that we would see $\boldsymbol{x}$?

Naive Bayes – Basics

Which class has the maximum likelihood,
given the observed feature values?

- Given an instance (feature tuple) $\boldsymbol{x}$, which class $c \in C$ has the highest likelihood to fit $\boldsymbol{x}$, i.e, for which class is it the most plausible that we would see $\boldsymbol{x}$?
- Compare conditional probabilities: $\mathrm{P}(c \mid \boldsymbol{x})$

Naive Bayes – Basics

Which class has the maximum likelihood,
given the observed feature values?

- Given an instance (feature tuple) $\boldsymbol{x}$, which class $c \in C$ has the highest likelihood to fit $\boldsymbol{x}$, i.e, for which class is it the most plausible that we would see $\boldsymbol{x}$?
- Compare conditional probabilities: $\mathrm{P}(c \mid \boldsymbol{x})$
- Predict the class with the highest likelihood, i.e.

$$\underset{c \in C}{argmax}\, \mathrm{P}(c \mid \boldsymbol{x})$$

Naive Bayes – Basics

Which class has the maximum likelihood,
given the observed feature values?

- Given an instance (feature tuple) $\boldsymbol{x}$, which class $c \in C$ has the highest likelihood to fit $\boldsymbol{x}$, i.e, for which class is it the most plausible that we would see $\boldsymbol{x}$?
- Compare conditional probabilities: $\mathrm{P}(c \mid \boldsymbol{x})$
- Predict the class with the highest likelihood, i.e.

$$\underset{c \in C}{argmax}\, \mathrm{P}(c \mid \boldsymbol{x})$$

How can we estimate these conditional probabilities for arbitrary $\boldsymbol{x}$?

Estimating Likelihoods

- According to Bayes Theorem:

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})}$$

Estimating Likelihoods

- According to Bayes Theorem:

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})}$$

- the likelihood of class c is computed from

Estimating Likelihoods

- According to Bayes Theorem:

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})}$$

- the likelihood of class c is computed from
  - the evidence $P(\boldsymbol{x})$,

Estimating Likelihoods

- According to Bayes Theorem:

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})}$$

- the likelihood of class c is computed from
  - the evidence $P(\boldsymbol{x})$,
  - the prior probability of c and

Estimating Likelihoods

- According to Bayes Theorem:

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})}$$

- the likelihood of class c is computed from
  - the evidence $P(\boldsymbol{x})$,
  - the prior probability of c and
  - the conditional probability of $\boldsymbol{x}$ given c, which can be interpreted as the plausibility (likelihood) of c given $\boldsymbol{x}$.

Naive Bayes – Classification

- Given an instance $\boldsymbol{x}$, $\mathrm{P}(\boldsymbol{x})$ is constant. Thus

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})} \propto \mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c).$$

Naive Bayes – Classification

- Given an instance $\boldsymbol{x}$, $\mathrm{P}(\boldsymbol{x})$ is constant. Thus

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})} \propto \mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c).$$

- Thus the decision rule becomes

$$\underset{c \in C}{argmax}\, \mathrm{P}(\boldsymbol{x} \mid c) \cdot P(c).$$

Naive Bayes – Classification

- Given an instance $\boldsymbol{x}$, $\mathrm{P}(\boldsymbol{x})$ is constant. Thus

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})} \propto \mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c).$$

- Thus the decision rule becomes

$$\underset{c \in C}{argmax}\, \mathrm{P}(\boldsymbol{x} \mid c) \cdot P(c).$$

- $P(c)$ can be estimated by counting the relative class frequency in the training data.

Naive Bayes – Classification

- Given an instance $\boldsymbol{x}$, $\mathrm{P}(\boldsymbol{x})$ is constant. Thus

$$\mathrm{P}(c \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c)}{\mathrm{P}(\boldsymbol{x})} \propto \mathrm{P}(\boldsymbol{x} \mid c) \cdot \mathrm{P}(c).$$

- Thus the decision rule becomes

$$\underset{c \in C}{argmax}\, \mathrm{P}(\boldsymbol{x} \mid c) \cdot P(c).$$

- $P(c)$ can be estimated by counting the relative class frequency in the training data.

How to determine $\mathrm{P}(\boldsymbol{x} \mid c)$?

Example 1: A Single Feature

day	wind	tennis?
1	weak	no
2	strong	no
3	weak	yes
4	weak	yes
5	weak	yes
6	strong	no

Classification task: Play tennis when wind is weak?

Example 1: A Single Feature

day	wind	tennis?
1	weak	no
2	strong	no
3	weak	yes
4	weak	yes
5	weak	yes
6	strong	no

Classification task: Play tennis when wind is weak?

- $\mathrm{P}(\mathrm{no}|\mathrm{weak}) \sim \mathrm{P}(\mathrm{weak}|\mathrm{no}) \cdot \mathrm{P}(\mathrm{no}) = \frac{1}{3} \cdot \frac{3}{6} = \frac{1}{6}$

Example 1: A Single Feature

day	wind	tennis?
1	weak	no
2	strong	no
3	weak	yes
4	weak	yes
5	weak	yes
6	strong	no

Classification task: Play tennis when wind is weak?

- $\mathsf{P}(\text{no}|\text{weak}) \sim \mathsf{P}(\text{weak}|\text{no}) \cdot \mathsf{P}(\text{no}) = \frac{1}{3} \cdot \frac{3}{6} = \frac{1}{6}$
- $\mathsf{P}(\text{yes}|\text{weak}) \sim \mathsf{P}(\text{weak}|\text{yes}) \cdot \mathit{Prob}(\text{yes}) = \frac{3}{3} \cdot \frac{3}{6} = \frac{1}{2}$

Example 1: A Single Feature

day	wind	tennis?
1	weak	no
2	strong	no
3	weak	yes
4	weak	yes
5	weak	yes
6	strong	no

Classification task: Play tennis when wind is weak?

- $\mathrm{P}(\mathrm{no}|\mathrm{weak}) \sim \mathrm{P}(\mathrm{weak}|\mathrm{no}) \cdot \mathrm{P}(\mathrm{no}) = \frac{1}{3} \cdot \frac{3}{6} = \frac{1}{6}$
- $\mathrm{P}(\mathrm{yes}|\mathrm{weak}) \sim \mathrm{P}(\mathrm{weak}|\mathrm{yes}) \cdot \mathit{Prob}(\mathrm{yes}) = \frac{3}{3} \cdot \frac{3}{6} = \frac{1}{2}$
- Decision: yes

Example 1: A Single Feature

day	wind	tennis?
1	weak	no
2	strong	no
3	weak	yes
4	weak	yes
5	weak	yes
6	strong	no

Classification task: Play tennis when wind is weak?

- $\mathrm{P}(\mathrm{no}|\mathrm{weak}) \sim \mathrm{P}(\mathrm{weak}|\mathrm{no}) \cdot \mathrm{P}(\mathrm{no}) = \frac{1}{3} \cdot \frac{3}{6} = \frac{1}{6}$
- $\mathrm{P}(\mathrm{yes}|\mathrm{weak}) \sim \mathrm{P}(\mathrm{weak}|\mathrm{yes}) \cdot \mathit{Prob}(\mathrm{yes}) = \frac{3}{3} \cdot \frac{3}{6} = \frac{1}{2}$
- Decision: yes

Probability estimate: count frequencies in the training data.

Example 2: Multiple Attributes

Attribute space is more sparsely covered (much bigger space through combinatorial explosion)

day	outlook	temperature	humidity	wind	tennis?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	clouded	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no

Classification task: Play tennis on a clouded, mild day with normal humidity and weak wind?

$$P(\text{clouded} \wedge \text{mild} \wedge \text{normal} \wedge \text{weak} \mid \text{no}) = ?$$

Computing the Probabilities 1/2

- Task: Estimate $\mathrm{P}(\boldsymbol{x} \mid c)$, for $\boldsymbol{x} = (x_1, \ldots, x_d)$

Computing the Probabilities 1/2

- Task: Estimate $\mathrm{P}(\boldsymbol{x} \mid c)$, for $\boldsymbol{x} = (x_1, \ldots, x_d)$
- **Naive Independence Assumption**: for each class c, the attribute values are independent, i.e.

$$\mathrm{P}(\boldsymbol{x} \mid c) = \prod_{i=1}^{d} \mathrm{P}(x_i \mid c)$$

Computing the Probabilities 1/2

- Task: Estimate $\mathrm{P}(\boldsymbol{x} \mid c)$, for $\boldsymbol{x} = (x_1, \ldots, x_d)$
- **Naive Independence Assumption**: for each class c, the attribute values are independent, i.e.

$$\mathrm{P}(\boldsymbol{x} \mid c) = \prod_{i=1}^{d} \mathrm{P}(x_i \mid c)$$

- Estimate $\mathrm{P}(x_i \mid c)$ for $i = 1, \ldots, d$ by counting relative frequencies.

Computing the Probabilities 1/2

- Task: Estimate $\mathrm{P}(\boldsymbol{x} \mid c)$, for $\boldsymbol{x} = (x_1, \ldots, x_d)$
- **Naive Independence Assumption**: for each class c, the attribute values are independent, i.e.

$$\mathrm{P}(\boldsymbol{x} \mid c) = \prod_{i=1}^{d} \mathrm{P}(x_i \mid c)$$

- Estimate $\mathrm{P}(x_i \mid c)$ for $i = 1, \ldots, d$ by counting relative frequencies.
- Decision rule of Naive Bayes:

$$\underset{c \in C}{argmax} \left(\mathrm{P}(c) \cdot \prod_{i=1}^{d} \mathrm{P}(x_i \mid c) \right)$$

"All models are wrong,
but some are useful."

George Box (statistician)

Computing the Probabilities 2/2

Naiveté:

Computing the Probabilities 2/2

Naiveté:

- The naive independence assumption is usually never correct.

Computing the Probabilities 2/2

Naiveté:

- The naive independence assumption is usually never correct.
- Implicit assumption: All attributes are equally important (equally contribute to the product).

Computing the Probabilities 2/2

Naiveté:

- The naive independence assumption is usually never correct.
- Implicit assumption: All attributes are equally important (equally contribute to the product).

Restriction

Works only for categorical attributes.

Computing the Probabilities 2/2

Naiveté:

- The naive independence assumption is usually never correct.
- Implicit assumption: All attributes are equally important (equally contribute to the product).

Restriction

Works only for categorical attributes.

Zero-Probabilities

Computing the Probabilities 2/2

Naiveté:

- The naive independence assumption is usually never correct.
- Implicit assumption: All attributes are equally important (equally contribute to the product).

Restriction

Works only for categorical attributes.

Zero-Probabilities

- If for one x_i/c-combination $\mathrm{P}(x_i \mid c) = 0$, then $\prod_{i=1}^{d} \mathrm{P}(x_i \mid c) = 0$.

Computing the Probabilities 2/2

Naiveté:

- The naive independence assumption is usually never correct.
- Implicit assumption: All attributes are equally important (equally contribute to the product).

Restriction

Works only for categorical attributes.

Zero-Probabilities

- If for one x_i/c-combination $\mathrm{P}(x_i \mid c) = 0$, then $\prod_{i=1}^{d} \mathrm{P}(x_i \mid c) = 0$.
- Remedy: Smoothing. Introduce hyperparameter α:

$$P(x_i \mid c\,;\alpha) = \frac{|\{t \in T \mid t_i = x_i, t \in c\}| + \alpha}{|\{t \in T \mid t \in c\}| + \alpha|\{t_i \mid t \in T\}|},$$

Naive Bayes in Python

sklearn.naive_bayes.CategoricalNB

- Parameters: Mainly α for smoothing the probabilities

Naive Bayes in Python

sklearn.naive_bayes.CategoricalNB

- Parameters: Mainly α for smoothing the probabilities
- Variables must be categorical, yet presented as consecutive integers $0, 1, \ldots, n-1$.

Naive Bayes in Python

sklearn.naive_bayes.CategoricalNB

- Parameters: Mainly α for smoothing the probabilities
- Variables must be categorical, yet presented as consecutive integers $0, 1, \ldots, n-1$.
- Variables can be transformed using the OrdinalEncoder as their order is ignored in the algorithm.

Naive Bayes in Python

sklearn.naive_bayes.CategoricalNB

- Parameters: Mainly α for smoothing the probabilities
- Variables must be categorical, yet presented as consecutive integers $0, 1, \ldots, n-1$.
- Variables can be transformed using the OrdinalEncoder as their order is ignored in the algorithm.

Notebook 08_1_bayes_tennis

Outline

Restrictions of Categorical NB

- Assumes that in each class, each feature has a categorical distribution, i.e. k values each with their own probability, summing up to 1

Restrictions of Categorical NB

- Assumes that in each class, each feature has a categorical distribution, i.e. k values each with their own probability, summing up to 1
- This makes it expensive if features have many possible categories

Restrictions of Categorical NB

- Assumes that in each class, each feature has a categorical distribution, i.e. k values each with their own probability, summing up to 1
- This makes it expensive if features have many possible categories
- This does not work with counts (all possible values for counts would have to be in the training data).

Restrictions of Categorical NB

- Assumes that in each class, each feature has a categorical distribution, i.e. k values each with their own probability, summing up to 1
- This makes it expensive if features have many possible categories
- This does not work with counts (all possible values for counts would have to be in the training data).
- This does not work with continuous values.

Multivariate Distributions

- Consider multiple random variables $X_i : \Omega \to \mathbb{R}$ for $i = 1, \ldots, n$.

Multivariate Distributions

- Consider multiple random variables $X_i : \Omega \to \mathbb{R}$ for $i = 1, \dots, n$.
- These random variables are **independent** if for all choices of intervals A_i:

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

The Multinomial Distribution

- is a multivariate distribution

The Multinomial Distribution

- is a multivariate distribution
- X_i the frequency of each side i of a k sided (biased) die in n (independent) roles.

The Multinomial Distribution

- is a multivariate distribution
- X_i the frequency of each side i of a k sided (biased) die in n (independent) roles.
- $P(X_1 = x_1, \dots, X_k = x_k)$

$$= \begin{cases} \frac{n}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \text{ for } x_i \in \{0, \dots, n\} \text{ and } \sum_{i=1}^{k} x_i = n \\ 0 \text{ else.} \end{cases}$$

The Multinomial Distribution

- is a multivariate distribution
- X_i the frequency of each side i of a k sided (biased) die in n (independent) roles.
- $P(X_1 = x_1, \ldots, X_k = x_k)$

$$= \begin{cases} \frac{n}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \text{ for } x_i \in \{0, \ldots, n\} \text{ and } \sum_{i=1}^{k} x_i = n \\ 0 \text{ else.} \end{cases}$$

- k and the p_i are (fix) parameters of the distribution with $\sum_{i=0}^{k-1} p_i = 1$.

The Multinomial Distribution

- is a multivariate distribution
- X_i the frequency of each side i of a k sided (biased) die in n (independent) roles.
- $P(X_1 = x_1, \ldots, X_k = x_k)$

$$= \begin{cases} \frac{n}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \text{ for } x_i \in \{0, \ldots, n\} \text{ and } \sum_{i=1}^{k} x_i = n \\ 0 \text{ else.} \end{cases}$$

- k and the p_i are (fix) parameters of the distribution with $\sum_{i=0}^{k-1} p_i = 1$.
- generalization of categorical distribution ($n = 1$)

The Multinomial Distribution

- is a multivariate distribution
- X_i the frequency of each side i of a k sided (biased) die in n (independent) roles.
- $P(X_1 = x_1, \dots, X_k = x_k)$

$$= \begin{cases} \frac{n}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \text{ for } x_i \in \{0, \dots, n\} \text{ and } \sum_{i=1}^{k} x_i = n \\ 0 \text{ else.} \end{cases}$$

- k and the p_i are (fix) parameters of the distribution with $\sum_{i=0}^{k-1} p_i = 1$.
- generalization of categorical distribution ($n = 1$)
- generalization of binomial distribution $k = 2$

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document
  - smoothing the probabilities is extremely important, as many document classes do not contain all words at least once

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document
  - smoothing the probabilities is extremely important, as many document classes do not contain all words at least once
  - in python: sklearn.naive_bayes.MultinomialNB

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document
  - smoothing the probabilities is extremely important, as many document classes do not contain all words at least once
  - in python: sklearn.naive_bayes.MultinomialNB
    - parameter alpha for smoothing

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document
  - smoothing the probabilities is extremely important, as many document classes do not contain all words at least once
  - in python: sklearn.naive_bayes.MultinomialNB
    - parameter alpha for smoothing
    - probabilities are computed from sums (not frequencies) of the features

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document
  - smoothing the probabilities is extremely important, as many document classes do not contain all words at least once
  - in python: sklearn.naive_bayes.MultinomialNB
    - parameter alpha for smoothing
    - probabilities are computed from sums (not frequencies) of the features
    - works with non-integer data as well (e.g. tf-idf)

Multinomial Naive Bayes

- Assumes that per class, the features are multinomial, i.e. each class has a multinomial distribution describing the (integer) values of all the features.
  - very popular in text mining, features are words with their frequency in a document as value
    - role a (biased) die with the vocabulary as sides
    - role as many times as there are words in a document
  - smoothing the probabilities is extremely important, as many document classes do not contain all words at least once
  - in python: sklearn.naive_bayes.MultinomialNB
    - parameter alpha for smoothing
    - probabilities are computed from sums (not frequencies) of the features
    - works with non-integer data as well (e.g. tf-idf)

Notebook 08_2_bayes_20_news_groups

Outline

A (very brief) Intro to Probability Theory

Basic Naive Bayes

Multinomial Bayes

Gaussian Bayes

Discussion

Continuous Distributions

We cannot always attribute probability to specific individual values.

Definition 8

A random variable X is continuous if a function f_X, called **probability density function** exists, such that

1. $f_X(x) \geq 0$ for $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. for $a \leq b$ holds

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

Continuous Distributions

We cannot always attribute probability to specific individual values.

Definition 8

A random variable X is continuous if a function f_X, called **probability density function** exists, such that

1. $f_X(x) \geq 0$ for $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. for $a \leq b$ holds

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

- Note: we can only determine the probability for intervals of real number here!

Continuous Distributions

We cannot always attribute probability to specific individual values.

Definition 8

A random variable X is continuous if a function f_X, called **probability density function** exists, such that

1. $f_X(x) \geq 0$ for $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. for $a \leq b$ holds

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

- Note: we can only determine the probability for intervals of real number here!
- The distribution is given indirectly through the density function.

Continuous Distributions

We cannot always attribute probability to specific individual values.

Definition 8

A random variable X is continuous if a function f_X, called **probability density function** exists, such that

1. $f_X(x) \geq 0$ for $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x) dx = 1$
3. for $a \leq b$ holds

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx.$$

- Note: we can only determine the probability for intervals of real number here!
- The distribution is given indirectly through the density function.
- $f(X)$ is not the same as $P(X = x)$. In fact, $P(X = x) = 0$.

The Continuous Uniform Distribution

- probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ for } x \in [a, b] \\ 0 \text{ else} \end{cases}$$

The Continuous Uniform Distribution

- probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ for } x \in [a, b] \\ 0 \text{ else} \end{cases}$$

- a, b are fix parameters of the distribution

The Continuous Uniform Distribution

- probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ for } x \in [a, b] \\ 0 \text{ else} \end{cases}$$

- a, b are fix parameters of the distribution
- the size of the area under the density function, in the interval $[a, b]$ is 1

The Continuous Uniform Distribution

- probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ for } x \in [a, b] \\ 0 \text{ else} \end{cases}$$

- a, b are fix parameters of the distribution
- the size of the area under the density function, in the interval $[a, b]$ is 1
- the probability of each individual value is 0

The Gaussian Distribution

- probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Gaussian Distribution

- probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- μ and σ are (fix) parameters of the distribution, representing mean and standard deviation of data sampled from the distribution

The Gaussian Distribution

- probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- μ and σ are (fix) parameters of the distribution, representing mean and standard deviation of data sampled from the distribution
- Gaussians are central elements of various statistical analyses

Gaussian Naive Bayes

- Allows continuous data (counting frequencies or summing them does not work with continuous data).

Gaussian Naive Bayes

- Allows continuous data (counting frequencies or summing them does not work with continuous data).
- The likelihood of c being the class of an observed instance $\boldsymbol{x}$ with feature x_i is estimated as

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right)$$

Gaussian Naive Bayes

- Allows continuous data (counting frequencies or summing them does not work with continuous data).
- The likelihood of c being the class of an observed instance $\boldsymbol{x}$ with feature x_i is estimated as

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right)$$

- one Gaussian per class and feature.

Outline

Discussion: Bayes-Classifications

Positive:

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems
- incremental learning: new training instances can easily be added

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems
- incremental learning: new training instances can easily be added
- includes knowledge about the domain (data distributions)

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems
- incremental learning: new training instances can easily be added
- includes knowledge about the domain (data distributions)

Negative:

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems
- incremental learning: new training instances can easily be added
- includes knowledge about the domain (data distributions)

Negative:

- naive assumption sometimes too naive

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems
- incremental learning: new training instances can easily be added
- includes knowledge about the domain (data distributions)

Negative:

- naive assumption sometimes too naive
- inefficiency with many attributes

Discussion: Bayes-Classifications

Positive:

- Optimality property: On average, no other classification algorithm with the same a-priori knowledge can reach a higher average classification quality.
- always a good baseline for evaluating other classifiers
- high accuracy on many classification problems
- incremental learning: new training instances can easily be added
- includes knowledge about the domain (data distributions)

Negative:

- naive assumption sometimes too naive
- inefficiency with many attributes

Exercises 1–2