Naïve Bayes Classifier

- A robust predictor that is based on historical knowledge
- Conditional independence
- The core of the predictor: Bayes law
- The naïve assumption to make live easier

Law of calculating total probability

- The total probability of A is computed by summing up the parital probilities of A at events B_i

$$P(A)=\sum_{i=1}^{m} P(A|B_i)\cdot P(B_i)$$

- Beispiel Wahrscheinlichkeit Lust auf Eiscreme zu haben
  - An heißen Tagen: 90%
  - An warmen Tagen: 50%
  - An Kalten Tagen: 10%
- Wie hoch ist die totale Wahrscheinlichkeit von "Lust auf Eiscreme" (heiße, warme & kalte Tage sind gleichhäufig)

Conditional Dependence

If A depends on B

$$P(A, B) = P(A|B) \cdot P(B)$$

If independent

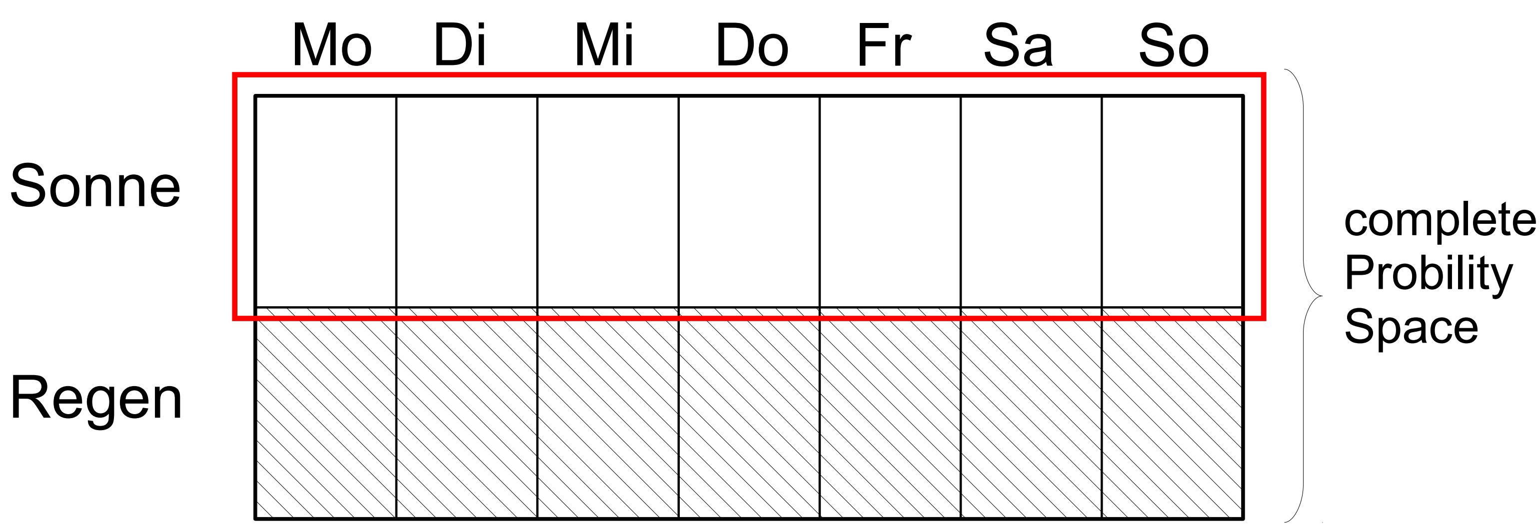
$$P(A, B) = P(A) \cdot P(B)$$

hence

$$P(A|B) = P(A)$$

Example Independence

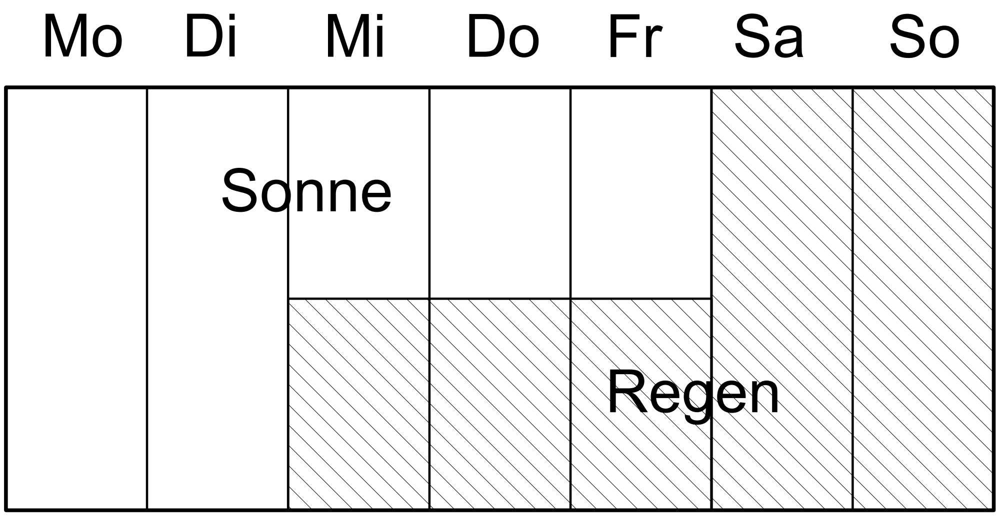
- 2 Variables with Values
  - Tag in {Mo,Di,Mi,Do,Fr,Sa,So}
  - Wetter in {Sonnig,Regen}
- A priori Probability
  - P(Tag=Mo)=~~1/7~~
  - P(Wetter=Sonnig)=1/2
- Show that Tag is independent of Wetter
  P(Mo|Sonnig)=P(Mo)



Example Dependence

2/7 1/7

Show that $P(Mo|Sonnig) \neq P(Mo)$



Naive Bayes Predition algorithm

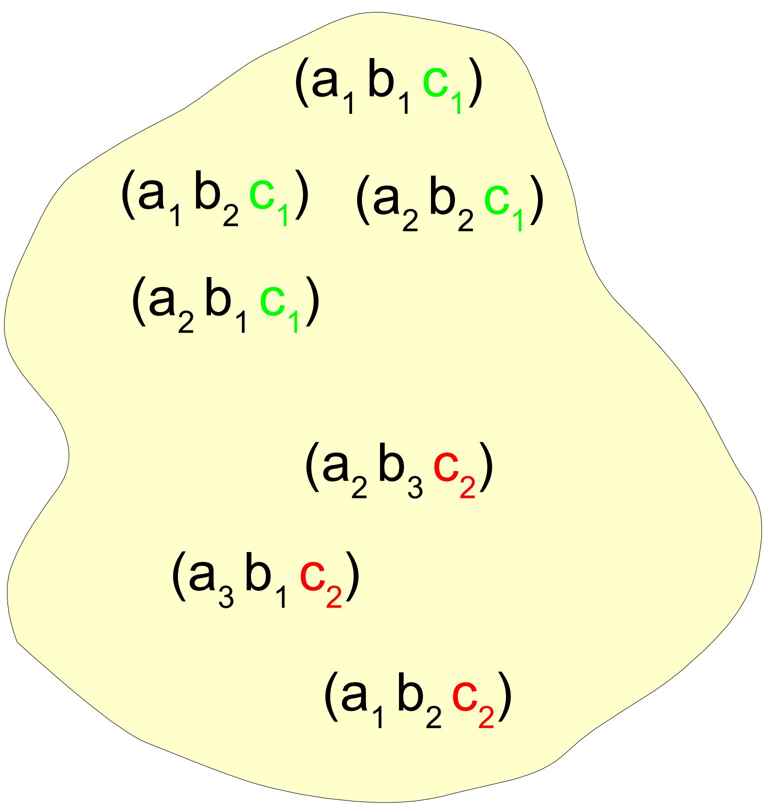
- Predict class $C=\{c_1, c_2\}$ for sample $D=[A=a_1, B=b_2]$
- Count how many times of the training samples
  - a_1 and b_2 occurred in c_1
  - a_1 and b_2 occurred in c_2
  - Take that class ĉ where a_1 and b_2 occurred more often $\hat{c} = argmax_{c \in C} P(D|c)$

Tools to explain why its working

- Bayes law $$P(c|D)=\frac{P(D|c)\cdot P(c)}{P(D)}$$
- Multiplication rule $$P(A\wedge B)=P(A|B)\cdot P(B)$$
- Conditional probability $$P(A|B)$$

Another Sample to show Conditional Independence

- Attributes $A=\{a_1,a_2\}$ $B=\{b_1,b_2\}$
- Class label $C=\{c_1,c_2\}$

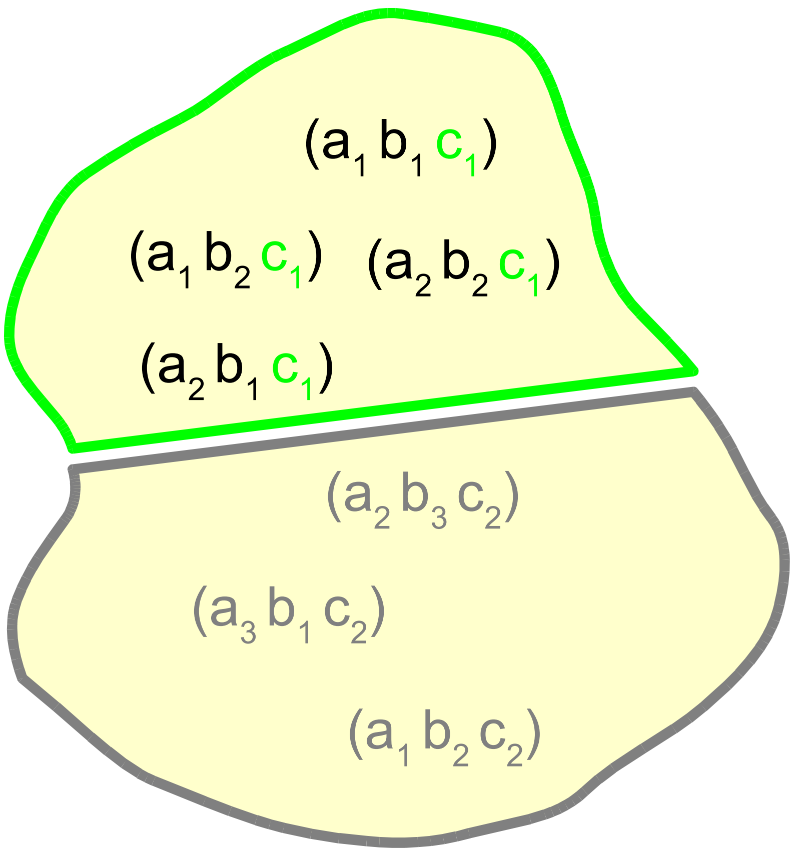


	Attribute A	Attribute B	Attribute C
Sample 1	a_1	b_1	c_1
Sample 2	a_2	b_1	c_1
Sample 3	a_1	b_2	c_1
Sample 4	a_2	b_2	c_1
…	…	…	c_2
…	…	…	c_2
…	…	…	c_2

Selection

- Select only samples with $C=c_1$

	Attribute A	Attribute B	Attribute C
Sample 1	a_1	b_1	c_1
Sample 2	a_2	b_1	c_1
Sample 3	a_1	b_2	c_1
Sample 4	a_2	b_2	c_1



a_1 & b_1 independent given c_1

Samples

	Attribute A	Attribute B	Attribute C
Sample 1	a_1	b_1	c_1
Sample 2	a_2	b_1	c_1
Sample 3	a_1	b_2	c_1
Sample 4	a_2	b_2	c_1

Statistics

	Absolute Frequency of samples with $C=c_1$	Relative Frequency of samples with $C=c_1$
$a_1 b_1$	1	1/4
$a_2 b_1$	1	1/4
$a_1 b_2$	1	1/4
$a_2 b_2$	1	1/4

	Absolute Frequency of samples with $C=c_1$	Relative Frequency of samples with $C=c_1$
a_1	2	1/2
a_2	2	1/2

	Absolute Frequency of samples with $C=c_1$	Relative Frequency of samples with $C=c_1$
b_1	2	1/2
b_2	2	1/2

Excerpt formal description

- $P(a_1 b_2 \mid c_1)=0.25$
- $P(a_1 \mid c_1)=0.5$
- $P(b_2 \mid c_1)=0.5$

Conditional Independence

- $P(a_1\, b_2 \mid c_1)=0.25$
- $P(a_1 \mid c_1)=0.5$
- $P(b_2 \mid c_1)=0.5$

- Simplification $P(A,B)$ to $P(A) \cdot P(B)$ allowed if
  - $P(a_i,b_k)=P(a_i) \cdot P(b_k)$ is true for all $a_i \in A$ and $b_k \in B$
- If Attribute A and B are conditional independent than
  - The multiplication rule $P(A \wedge B)=\boxed{P(A|B)} \cdot P(B)$
  - is simplified to $P(A \wedge B)=\boxed{P(A)} \cdot P(B)$

Prediction using naïve Bayes classifier

- Input
  - Supervised Training samples
    - Discrete independent attributes (nominal or ordinal) D ϵ [A x B]
    - Discrete class label C ϵ $\{c_1, c_2\}$
  - an unseen sample that is to be predicted (only the independent attributes D are known)
- Output
  - The most likely class $\hat{c}$ ϵ C is predicted for the unseen sample

Example

- Domain: Credit Worthiness
  - Data sample D consists of two variables
    - Income={low,middle,high}
    - furtherCredits={none, one, many}
  - Class label C
    - Creditworthy={no (not credit worthy), yes (credit worthy)}
  - e.g. Supervised data
    - [$\overbrace{\text{Income=low, furtherCredits=many,}}^{D}$ $\overbrace{\text{Creditworthy=no}}^{C}$]
    - [Income=high, furtherCredits=one, Creditworthy=yes]
  - e.g. Unseen sample
    - D= [Income=low, furtherCredits=one]

Naïve Bayes

- Find class c with highest probability given the data D

$$\hat{c} = argmax_{c \in C} P(c|D)$$

- Apply Bayes law

$$\hat{c} = argmax_{c \in C} \frac{P(D|c) \cdot P(c)}{P(D)}$$

- Skip P(D) as ineffective constant

$$\hat{c} = argmax_{c \in C} P(D|c) \cdot P(c)$$

- Skip P(c) assuming homogeneously-distributed classes

$$\hat{c} = argmax_{c \in C} P(D|c)$$

  - Maximum likelihood hypothesis

Working Prediction Example

$$\hat{c} = argmax_{c \in C} P(D|c)$$

- Supervised database
  - [Income=low, furtherCredits=many, Creditworthy=no]
  - [Income=middle, furtherCredits=many, Creditworthy=no]
  - [Income=middle, furtherCredits=one, Creditworthy=no]
  - [Income=high, furtherCredits=one, Creditworthy=yes]
  - [Income=high, furtherCredits=none, Creditworthy=yes]
- Unseen sample for which class is to be predicted
  - D= [Income=high, furtherCredits=one]
  - Compute probability

$$\hat{c} = argmax_{c \in C} P([A, B]|c)$$

    - For class “no”: P([Income=high, furtherCredits=one] | no)
    - For class “yes”: P([Income=high, furtherCredits=one] | yes)

Not working Prediction Example

- Supervised database (the same database)
  - [Income=low, furtherCredits=many, Creditworthy=no]
  - [Income=middle, furtherCredits=many, Creditworthy=no]
  - [Income=middle, furtherCredits=one, Creditworthy=no]
  - [Income=high, furtherCredits=one, Creditworthy=yes]
  - [Income=high, furtherCredits=none, Creditworthy=yes]
- Unseen sample (is now different)
  - D= [Income=low, furtherCredits=one]
  - Computing the probability is not possible
    - P([Income=low, furtherCredits=one] | no) not found
    - P([Income=low, furtherCredits=one] | yes) not found

Sparsity Problem

$$\hat{c} = argmax_{c \in C} P([A, B] | c)$$

- Most often the Database is sparse
  - The combination of attributes that you need for your unseen sample is either very rare or not existent in the database
- Solution
  - Make the naïve assumption that the independent Attributes are conditionally independent

$\hat{c} = argmax_{c \in C} P([A, B] | c)$ is replaced by

$$\underline{\hat{c} = argmax_{c \in C} \left(P(A | c) \cdot P(B | c) \right)}$$

Prediction Example

- Supervised database $\hat{c} = argmax_{c \in C} \left(P(A|c) \cdot P(B|c) \right)$
  - [Income=low, furtherCredits=many, Creditworthy=no]
  - [Income=middle, furtherCredits=many, Creditworthy=no]
  - [Income=middle, furtherCredits=one, Creditworthy=no]
  - [Income=high, furtherCredits=one, Creditworthy=yes]
  - [Income=high, furtherCredits=none, Creditworthy=yes]
- Unseen sample for which class is to be predicted
  - D= [Income=low, furtherCredits=one]
  - Compute probability
    - "no": P(Income=low | no) * P(furtherCredits=one | no)=1/3 * 1/3
    - "yes": P(Income=low | yes) * P(furtherCredits=one | yes)= 0 * 1/2

Properties of naïve Bayes

- Output is the “best” class ĉ and an evaluation value as well (likelihood of ĉ of being the correct class)
- No model required
  - With each new training sample the likelihood is updated dynamically
  - Easy and simple but very robust in production mode