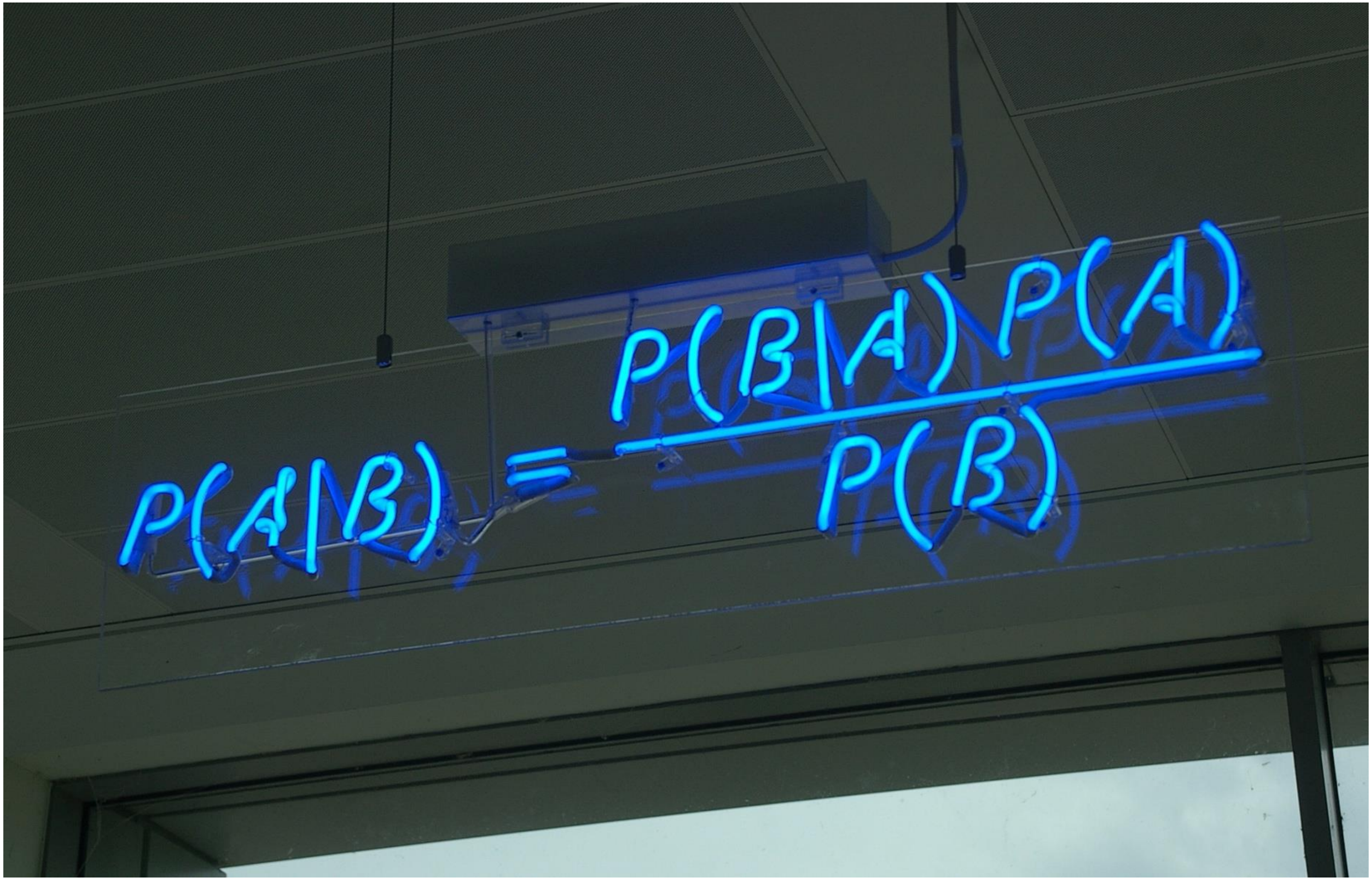
Data Mining
(Mining Knowledge from Data)

Bayes classifier

Marcel Jiřina, Pavel Kordík

Bayes formula



Bayes classification – what is it?

- Statistical classification method.
- To express the certainty with which the data had been correctly classified.
- Named after Thomas Bayes (1702-1761), who described the Bayes theorem.

Why Bayes?

- It provides a practical way of learning.
  - Example: Naive Bayes
- The prior probability and the observed data can be combined.
- Calculates explicitly the probability of hypothesis.
- Provides insight into complex learning algorithms.
- It provides the gold standard against which it is possible to compare other classifiers.
- Resistant to noise in the data.

Probability

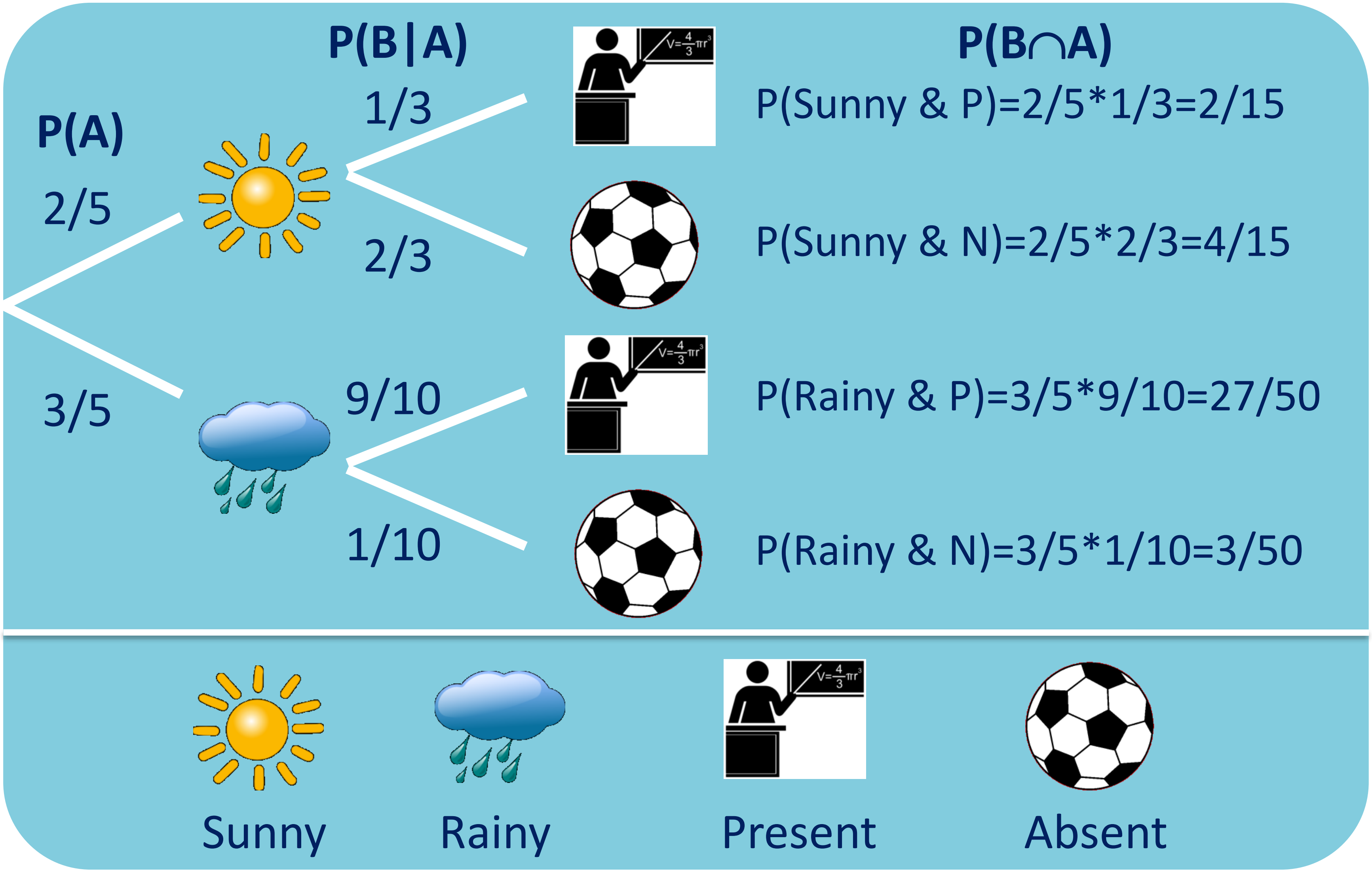
"Probability theory is nothing but common sense reduced to calculation" Pierre-Simon Laplace – 1814

Conditional probability

- P(A) is the probability of occurrence of phenomenon A.
- P(B|A) is the probability of phenomenon B, provided that phenomenon A occurred.
- P(A∩B) is the probability that both A and B phenomena occurred.

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

The probability of student presence at the lecture



Derivation of Bayes theorem

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

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

Because it holds $P(B \cap A) = P(A \cap B)$:

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

Bayes theorem/classifier

credibility
(probability of data B,
when A hypothesis it is true)

prior probability
(probability of hypothesis A
before we see data)

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

posterior probability
(probability of hypotheses A
after we saw data B)

normalizing term
(probability of data B.
Ensures that we get probability)

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Example

- **P(A/B)**: We want to find the probability that a customer buys a computer from us (A), when we know the age of B (posterior probability).
- **P(A)**: The probability that a customer bought a computer from us, regardless of his age (prior probability).
- **P(B/A)**: The probability that the customer is 35 years old when he bought a computer from us (Credibility).
- **P(B)**: The probability that the customer is 35 years (normalization term).

Maximal Posterior Probability (MPP)

- We want to find the most likely phenomenon A on the basis of training data B.

- $A_{MPP} = \max P(A_i|B)$

A_1 will buy a computer
A_2 will not buy a computer

- $A_{MPP} = \max \frac{P(B|A_i)P(A_i)}{P(B)}$

As $P(B)$ is the same for all A_i, we can ignore it

- $A_{MMP} = \max P(B|A_i)\, P(A_i)$

Maximum Likelihood

- We can suppose that $P(A_i) = P(A_j)$
- We are therefore not concerned in advance
  - In our example we assume that a half of customers will buy a computer ...
- This leads to simplification:

$$A_{MMP} = \max P(B|A_i)\, P(A_i)$$
$$A_{MV} = \max P(B|A_i)$$

Example

Customer id	Age	Income	University education	Own car	Will buy a computer?
1	35	Middle	Yes	Yes	Yes
2	30	High	No	Yes	No
3	40	Low	Yes	No	No
4	35	Middle	No	No	Yes
5	45	Low	No	No	Yes
6	35	High	No	Yes	Yes
7	35	Middle	No	Yes	No
8	25	Low	No	Yes	No
9	28	High	No	Yes	No
10	35	Middle	Yes	Yes	Yes

Example (cont.)

- P(will buy a computer = yes) = 5/10 = 0.5
- P(will buy a computer = no) = 5/10 = 0.5
- P(the customer is 35 & middle income) =
  = 4/10 = 0.4
- P(the customer is 35 & middle income | will buy a computer = yes) = 3/5 =0.6
- P(the customer is 35 & middle income | will buy a computer = no) = 1/5 = 0.2
- Will the customer buy a computer, yes or no?

Example (cont.)

- A customer will buy a computer $P(A_1|B)$
  $= P(A_1) * P(B|A_1) / P(B)$
  $= 0.5 * 0.6 / 0.4 = 0.75$
- A customer will not buy a computer $P(A_2|B)$
  $= P(A_2) * P(B|A_2) / P(B)$
  $= 0.5 * 0.2 / 0.4 = 0.25$
- Result = max $\{P(A_1|B), P(A_2|B)\}$
  = max(0.75; 0.25)

➔ The customer will buy the computer

Example (cont.)

- What if we have a customer:
  40 years old, high income?

Customer id	Age	Income	University education	Own car	Will buy a computer?
1	35	Middle	Yes	Yes	Yes
2	30	High	No	Yes	No
3	40	Low	Yes	No	No
4	35	Middle	No	No	Yes
5	45	Low	No	No	Yes
6	35	High	No	Yes	Yes
7	35	Middle	No	Yes	No
8	25	Low	No	Yes	No
9	28	High	No	Yes	No
10	35	Middle	Yes	Yes	Yes

Bayes

Basic education?

Prior probability, that someone has a basic education

High-school education?

Age above 21?

-	0.7

University education?

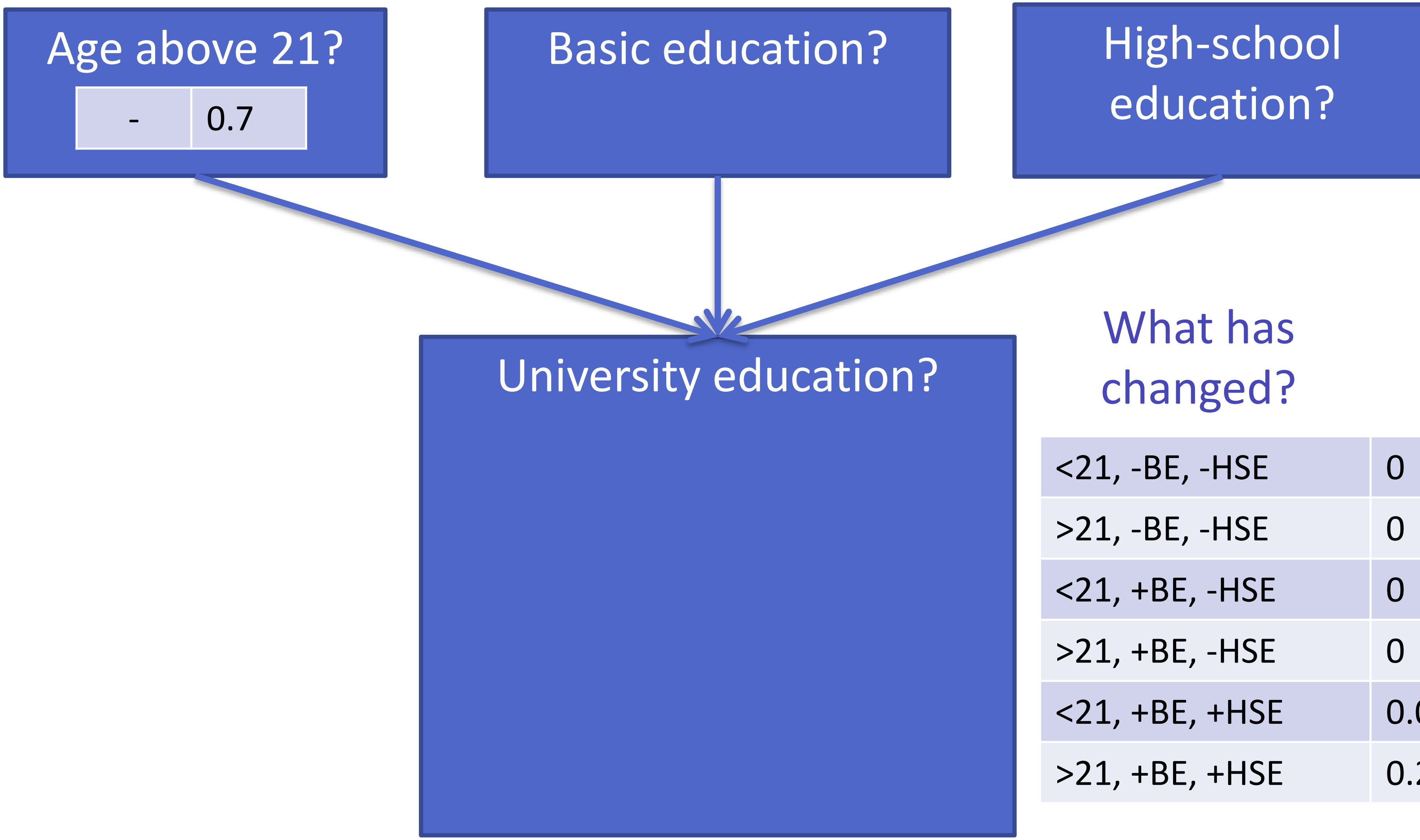
Conditional probability that someone has high-school education

<21, -HSE	0
>21, -HSE	0
<21, +HSE	0.02
>21, +HSE	0.2

Bayes

- Excellent model, but usually we do not know how much are phenomena interdependent.
- Dependencies can be estimated from the training data, but usually we do not have enough data.
- Therefore, Naive Bayes is used...

Naive Bayes



What has changed?

<21, -BE, -HSE	0
>21, -BE, -HSE	0
<21, +BE, -HSE	0
>21, +BE, -HSE	0
<21, +BE, +HSE	0.02
>21, +BE, +HSE	0.20

Naive Bayes

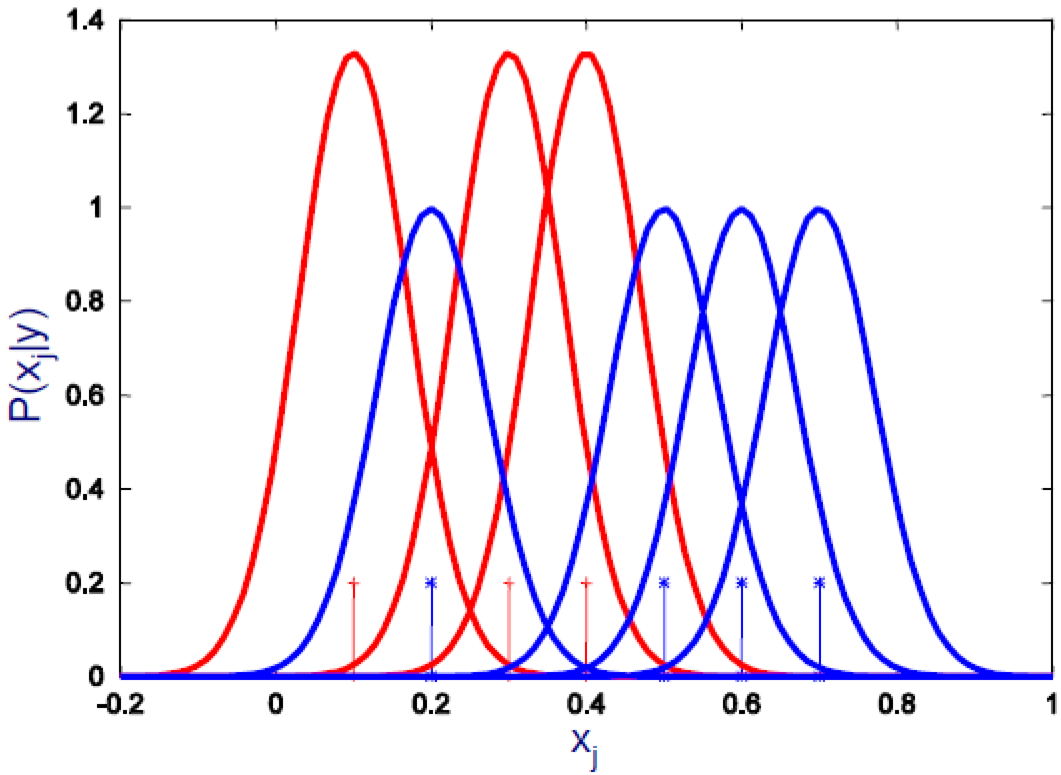
- Naive Bayes supposes

$$P(X_1,\ldots,X_N \mid C) = P(X_1 \mid C)\cdots P(X_N \mid C)$$

  thus independence of attributes.
- Each attribute X_i is independent on other attributes, once we know the value of C.

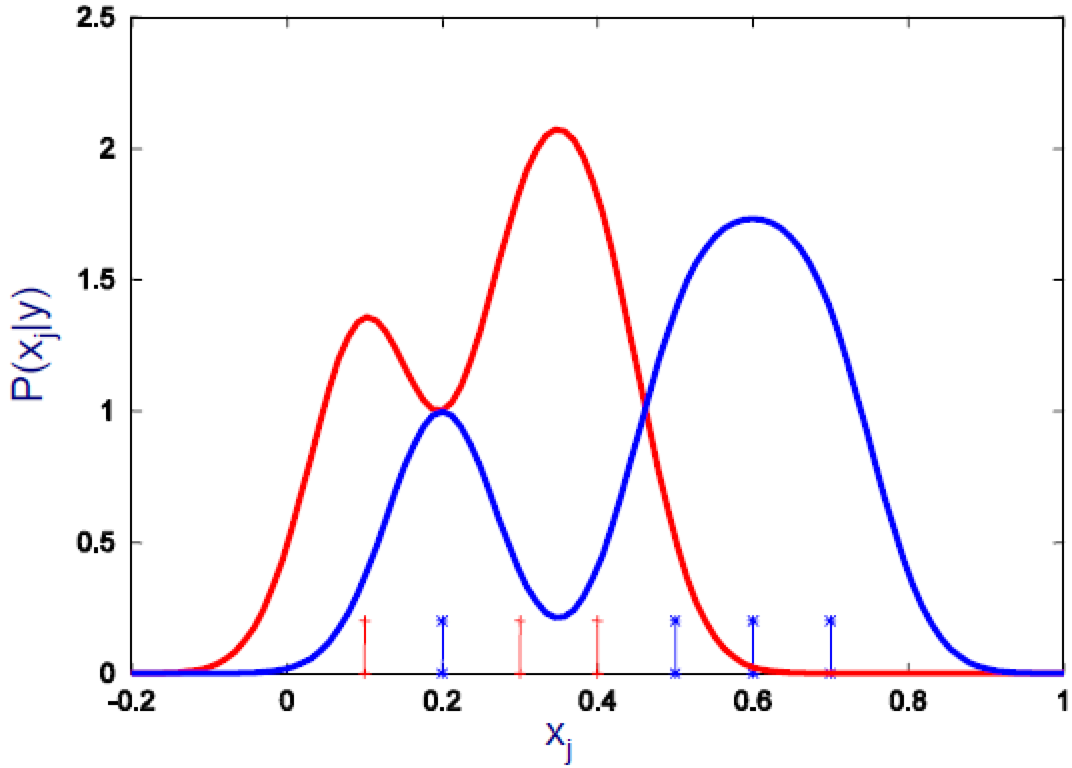
Kernel estimate

- For each sample one Gaussian function is formed, and subsequently all are summed together.

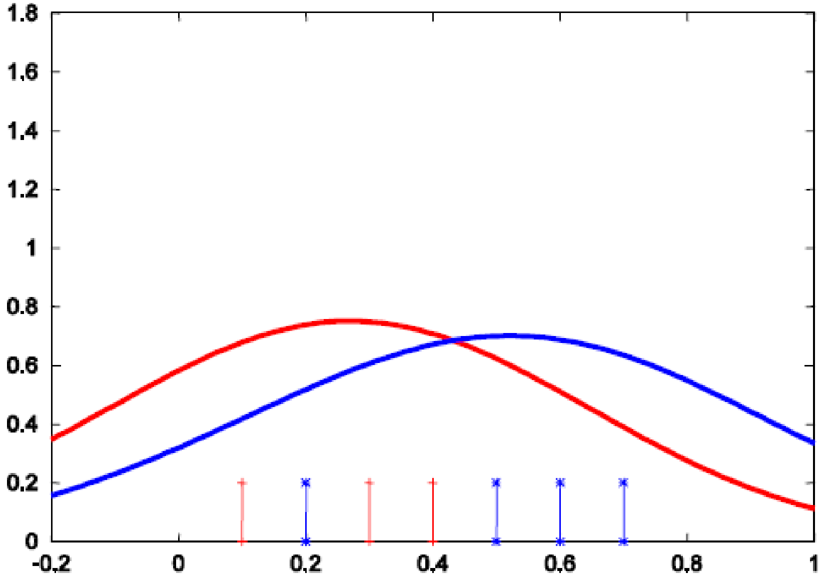


Kernel estimate

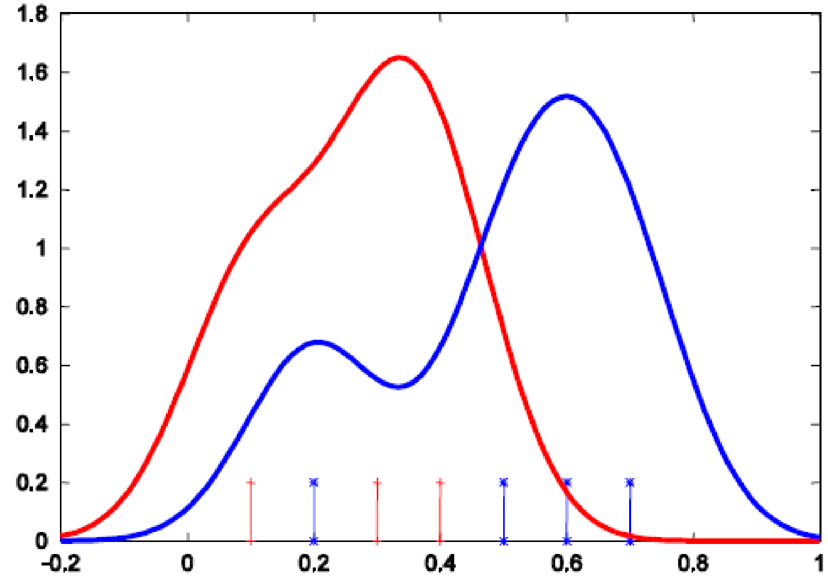
- The resulting probability density



Selection of variance of Gaussian function



σ=0.50



σ=0.15

Advantages and disadvantages of Naive Bayes

– The assumption of independence of attributes

– Assuming a normal distribution

– In the case of abundance of data other methods give usually better results

\+ Easy to implement

\+ To learn just from a few data

Comparison of Classifiers

Feature	Trees	k-NN	Naive Bayes	Neural networks
Mix of attribute types	yes	no	yes	no
Missing data	yes	some	yes	no
Outliers	yes	yes	questionable	yes
Scalability	yes	no	yes	yes
Interpretation	yes	no	yes	no
Accuracy	no	no	yes	yes

Online sources

- http://www.statsoft.com/textbook/naive-bayes-classifier/
- http://en.wikipedia.org/wiki/Bayes%27_theorem