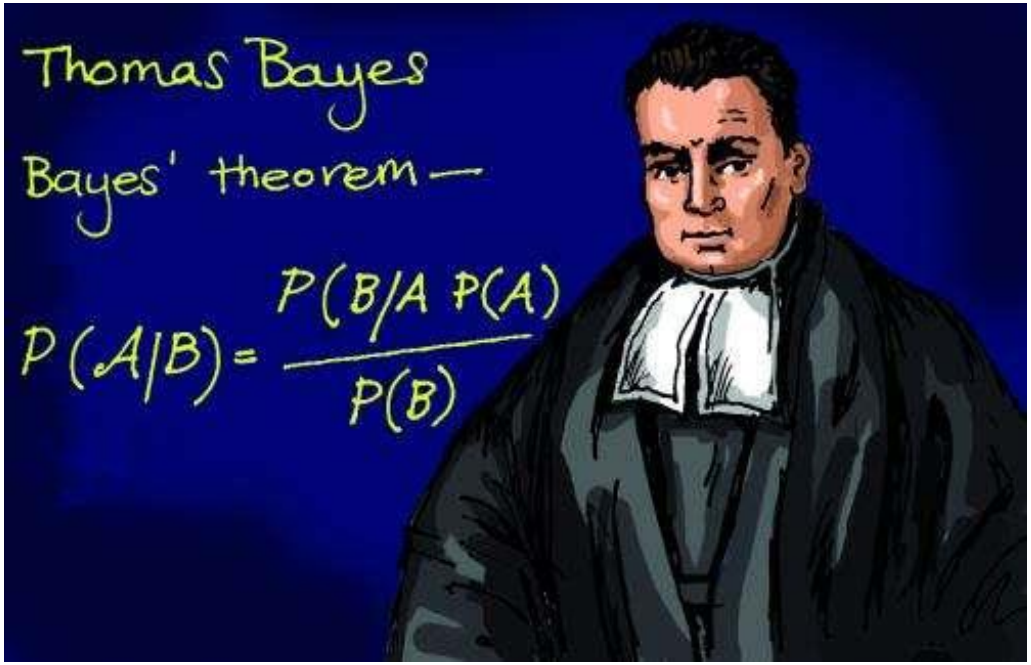
Bayes Classifier

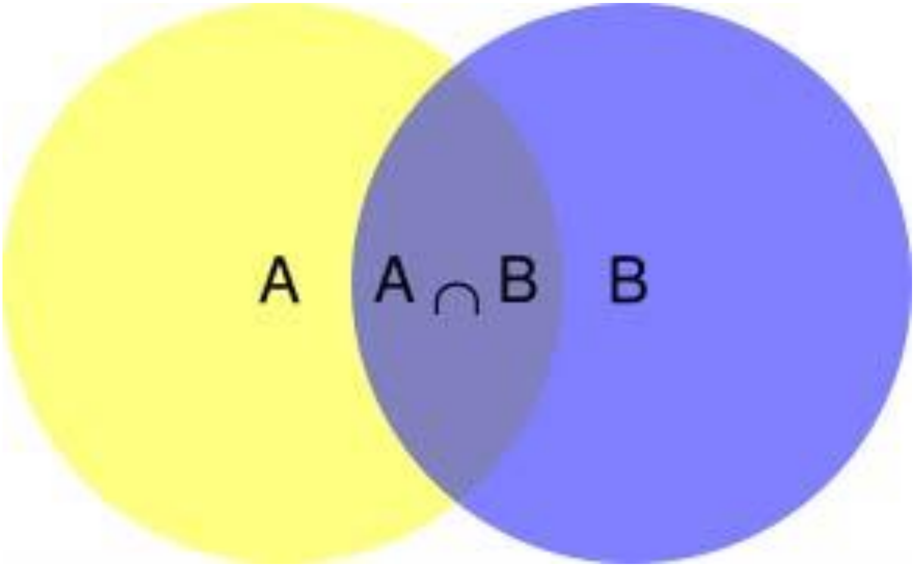
Bayes Classifier



An essay towards solving a problem in the doctrine of chances.

-- Thomas Bayes

Bayes Rule



Conditional probability

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

Bayes Rule

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

- Similarly

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

- Therefore

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

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

Conditional probability

Bayes Rule

Posterior probability

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

Prior probability

Conditional probability

Bayes Rule

Posterior probability

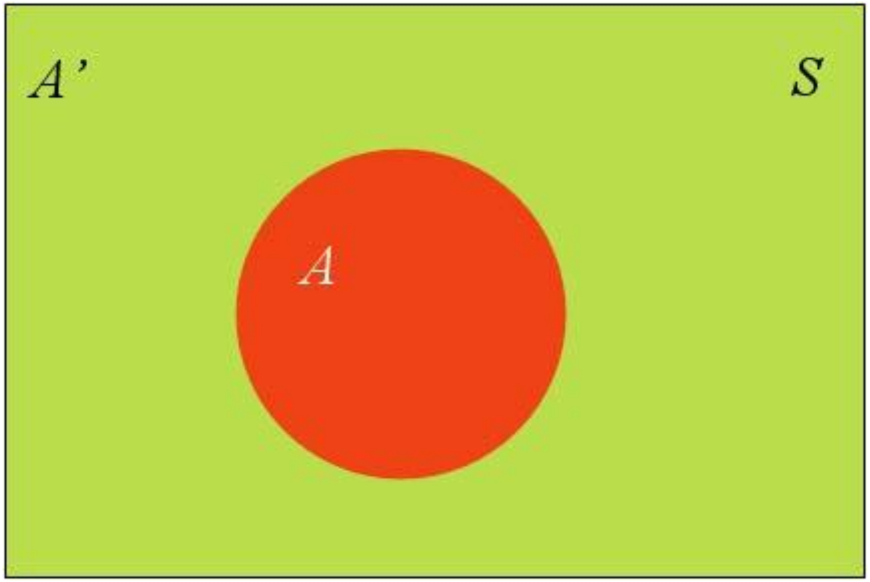
Class conditional probability

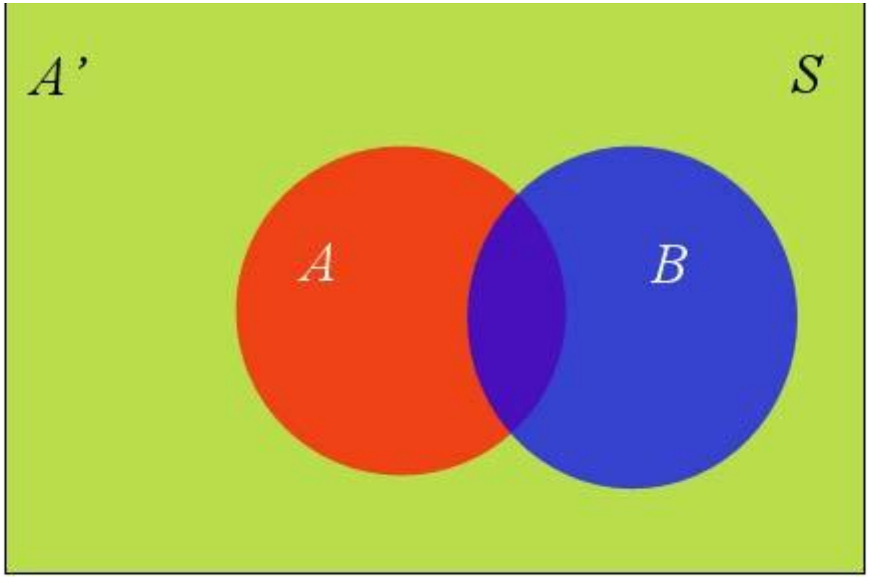
$$\boxed{P(c_i|x)} = \frac{\boxed{P(c_i)}\boxed{P(x|c_i)}}{\sum_{i=1}^{N} P(x|c_i)P(c_i)}$$

Prior probability

Bayes Rule

Incident: A / B





$$P(B) = P(B \cap A) + P(B \cap A')$$

Bayes Rule

$$P(B) = P(B \cap A) + P(B \cap A')$$

- Given

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

- So

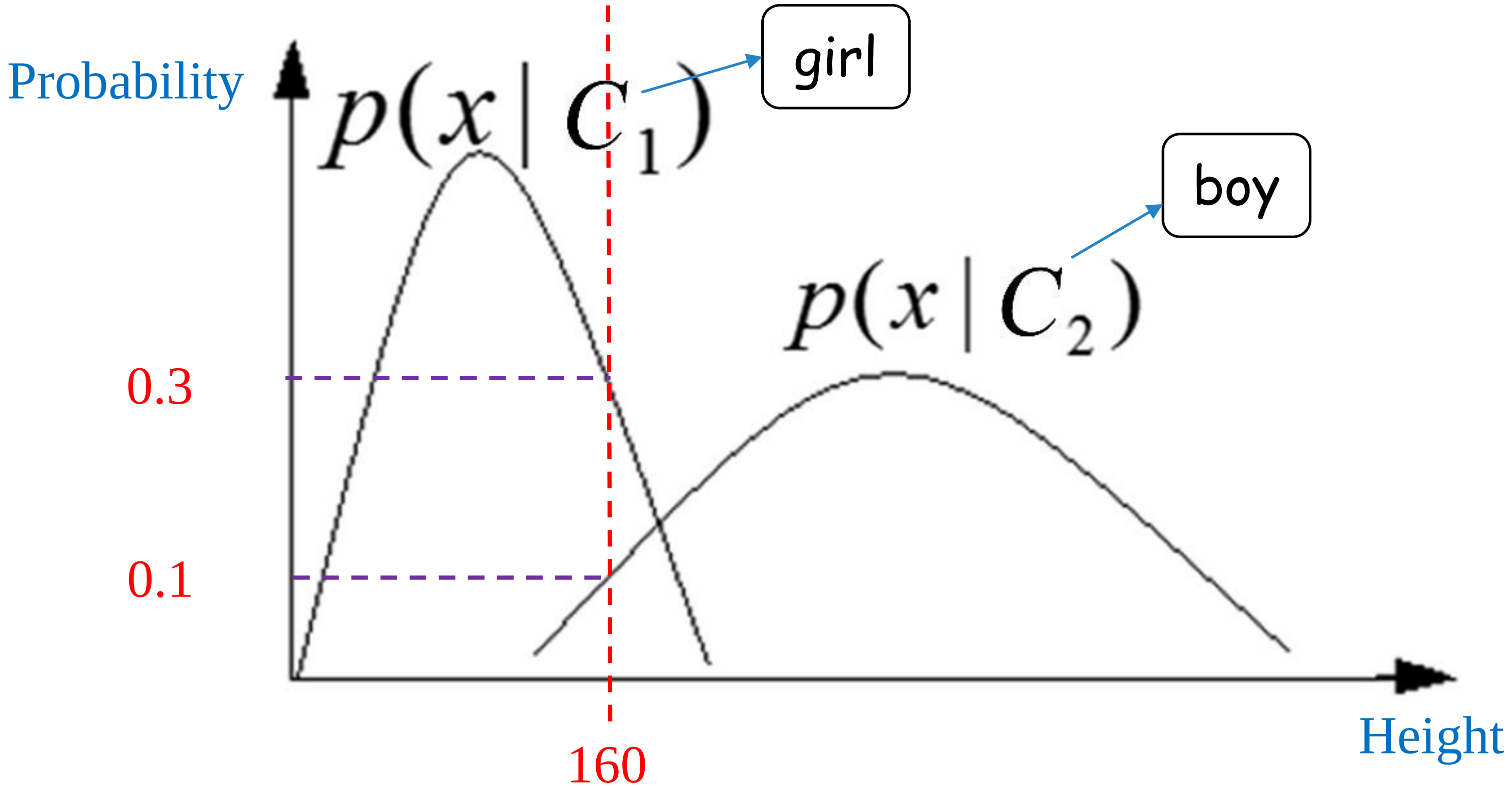
$$P(B) = P(B|A)P(A) + P(B|A')P(A')$$

Conditional probability

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

Bayes Rule

- Class conditional probability (类条件件概率)



Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Outline

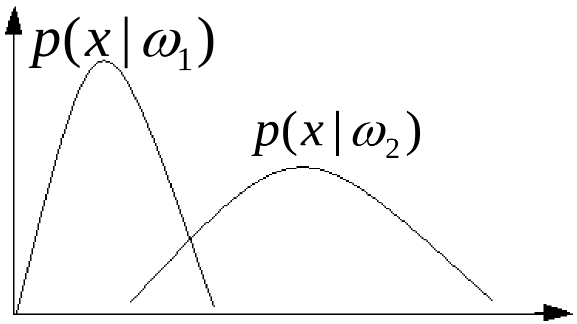
- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Bayesian Decision Based on Minimum Error Rate

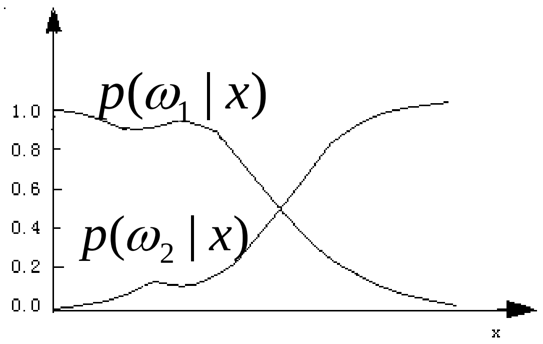
- In the problem of pattern classification, based on the Bayesian formula in probability theory to minimize the error of classification, the classification rule that minimizes the error rate can be obtained, which is called Bayesian decision based on minimum error rate.
- An example of cancer cell recognition illustrates the problem-solving process. Assuming that each cell to be identified has been preprocessed, d features representing the basic characteristics of the cell are extracted and become a vector x of the d-dimensional space. The purpose of the identification is to classify x as normal or abnormal cells.

Bayesian Decision Based on Minimum Error Rate

- Normal： $\omega = \omega_1$
- Abnormal： $\omega = \omega_2$
- Prior Probability： $p(\omega_1)$ $p(\omega_2)$
- Class Conditional Probability： $p(x \mid \omega_1)$ $p(x \mid \omega_2)$



Class Conditional Probability



Posterior Probability

Bayesian Decision Based on Minimum Error Rate

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{\sum_{j=1}^{2} p(x \mid \omega_j)P(\omega_j)}$$

- Bayesian Decision Based on Minimum Error Rate：
- If $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, x is classified as normal ω_1
- If $P(\omega_1 \mid x) < P(\omega_2 \mid x)$, x is classified as abnormal ω_2

Bayesian Decision Based on Minimum Error Rate

- Assuming that in a local area, the prior probabilities of normal and abnormal in cell recognition are:

  Normal: $P(c_1) = 0.9$

  Abnormal: $P(c_2) = 0.1$

- There is a cell to be identified, the observed value is x, from the class condition probability density distribution curve

$$p(x|c_1) = 0.2, \qquad p(x|c_2) = 0.4$$

- Try to judge whether the cell is normal or abnormal ?

Bayesian Decision Based on Minimum Error Rate

- The posterior probability of c_1 and c_2 is calculated by Bayesian formula :

$$P(c_1|x) = \frac{P(x|c_1)P(c_1)}{\sum_{j=1}^{2} P(x|c_j)P(c_j)} = \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.4 \times 0.1} = 0.818$$

$$P(c_2|x) = 1 - P(c_1|x) = 0.182$$

- According to Bayesian decision rules

$$P(c_1|x) = 0.818 > P(c_2|x) = 0.182$$

- Decision rules: Normal

Bayesian Decision Based on Minimum Error Rate

- From this example, it can be seen that the decision outcome depends on both the observed conditional probability density and the prior probability. In this example, because the prior probability of state 1 is several times greater than the prior probability of state 2, the prior probability plays a dominant role in making decisions.

Bayesian Decision Based on Minimum Error Rate

- Prior probability:

$$P(c_1) = 0.9 \qquad P(c_2) = 0.1$$

In fact, they are unknown.

- Class conditional probability:

$$\boxed{p(x|c_1) = 0.2, \qquad p(x|c_2) = 0.4}$$

- Bayesian model is a theoretical model. It is difficult to obtain prior probability and class conditional probability realistically, so their values need to be estimated.

Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Maximum Likelihood Estimation

to be estimated

$$P(c_i|x) = \frac{\boxed{P(c_i)P(x|c_i)}}{\sum_{i=1}^{N} P(x|c_i)P(c_i)}$$

Maximum Likelihood Estimation

- MLE (Maximum Likelihood Estimation)： A general method for estimating parameters in a model.

Likelihood

Given the distribution

$$\boxed{P(D_c : \theta_c)} = \prod_{x \in D_c} \boxed{P(x : \theta_c)}$$

Labeled Dataset

Parameter

$\hat{\theta}_c$: Choose θ_c that maximizes probability of observed data.

Maximum Likelihood Estimation

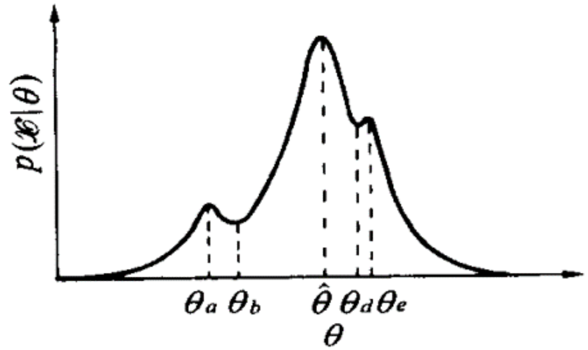
- Log-likelihood (对数似然)

$$P(D_c \mid \theta_c) = \prod_{x \in D_c} P(x \mid \theta_c)$$

$$LL(\theta_c) = \log P(D_c \mid \theta_c)$$

$$= \sum_{x \in D_c} \log P(x \mid \theta_c)$$

Then $$\hat{\theta}_c = \arg\max_{\theta_c} LL(\theta_c)$$



Maximum Likelihood Estimation

- Suppose $X_i \sim N(\mu, \sigma^2)$ and i.i.d.
- What is the likelihood function?

$$lik(\mu, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(X_i - \mu)^2}{2\sigma^2})\right]$$

- What is the log-likelihood function?

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \log[\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)]$$

$$= -\sum_{i=1}^{n} \log(\sigma) - \sum_{i=1}^{n} \log(\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}$$

$$= -n\log(\sigma) - n\log(\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

Maximum Likelihood Estimation

- Now there are two unknown parameters so we will need to find the separate partial derivatives:

$$\frac{\partial l(\mu,\sigma^2)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-nlog(\sigma) - nlog(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right)$$
$$= \frac{-1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}log(\sigma^2) - nlog(\sqrt{2\pi}) - \frac{1}{2}(\sigma^2)^{-1}\sum_{i=1}^{n}(X_i-\mu)^2\right)$$
$$= \frac{-n}{2\sigma^2} + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(X_i-\mu)^2$$

Maximum Likelihood Estimation

- Set the separate partial derivatives to zero and solve for the specific parameter:

$$\frac{\partial l(\mu,\sigma^2)}{\partial \mu} = \frac{-1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) \equiv 0$$

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}}$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}(X_i - \mu)^2 \equiv 0$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}$$

Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent

$$P(c_i \mid x) = \frac{P(c_i)\boxed{P(x \mid c_i)}}{P(x)} = \frac{P(c_i)}{P(x)}\boxed{\prod_{i=1}^{d} P(x_i \mid c_i)}$$

d The dimension of attributes

x_i The value of x on the attribute of i

Expression

Bayes optimal classifier

$$h*(x) = \underset{c \in y}{argmax} P(c|x)$$

$$P(c_i \mid x) = \frac{P(c_i)}{P(x)} \prod_{i=1}^{d} P(x_i \mid c_i)$$

NB optimal classifier

$$h_{nb}(x) = \arg\max_{c \in y} P(c_i) \prod_{i=1}^{d} P(x_i \mid c_i)$$

Naïve Bayes Classifier

Prior

$$P(C_i) = \frac{|D_c|}{|D|}$$

Class conditional probability

- For continuous attributes, suppose: $p(x_i \mid c) \sim N(\mu_{c,i}, \sigma_{c,i}^{2})$

$$p(x_i \mid c_i) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^{2}}\right)$$

- For discrete attributes

$$P(x_i|c) = \frac{|D_{c,x_i}|}{D_c}$$

Example

Test 1

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

Dataset

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

Example

$$P(\text{好瓜}=\text{是})=\frac{8}{17}\approx 0.471 \qquad P(\text{好瓜}=\text{否})=\frac{9}{17}\approx 0.529$$

$$P_{\text{青绿}|\text{是}}=\text{P}（\text{色泽}=\text{青绿}|\text{好瓜}=\text{是}）=\frac{3}{8}=0.375$$

$$P_{\text{青绿}|\text{否}}=\text{P}（\text{色泽}=\text{青绿}|\text{好瓜}=\text{否}）=\frac{3}{9}\approx 0.333$$

$$P_{\text{蜷缩}|\text{是}}=\text{P}（\text{根蒂}=\text{蜷缩}|\text{好瓜}=\text{是}）=\frac{5}{8}=0.625$$

$$P_{\text{蜷缩}|\text{否}}=\text{P}（\text{根蒂}=\text{蜷缩}|\text{好瓜}=\text{否}）=\frac{3}{9}\approx 0.333$$

$$P_{\text{浊响}|\text{是}}=\text{P}（\text{敲声}=\text{浊响}|\text{好瓜}=\text{是}）=\frac{6}{8}=0.750$$

$$P_{\text{浊响}|\text{否}}=\text{P}（\text{敲声}=\text{浊响}|\text{好瓜}=\text{否}）=\frac{4}{9}\approx 0.444$$

Example

$$P_{清晰|是} = \mathrm{P}（纹理 = 清晰|好瓜 = 是）= \frac{7}{8} = 0.875$$

$$P_{清晰|否} = \mathrm{P}（纹理 = 清晰|好瓜 = 否）= \frac{2}{9} \approx 0.222$$

$$P_{凹陷|是} = \mathrm{P}（脐部 = 凹陷|好瓜 = 是）= \frac{6}{8} = 0.750$$

$$P_{凹陷|否} = \mathrm{P}（脐部 = 凹陷|好瓜 = 否）= \frac{2}{9} \approx 0.222$$

$$P_{硬滑|是} = \mathrm{P}（触感 = 硬滑|好瓜 = 是）= \frac{6}{8} = 0.750$$

$$P_{硬滑|否} = \mathrm{P}（触感 = 硬滑|好瓜 = 否）= \frac{6}{9} \approx 0.667$$

$$\rho_{密度:0.697|是} = \rho(密度 = 0.697|好瓜 = 是) = \frac{1}{\sqrt{2\pi} \bullet 0.129} \exp\left(-\frac{(0.697-0.574)^2}{2 \bullet 0.129^2}\right) \approx 1.959$$

Example

$$\rho_{密度:0697|否} = \rho(密度 = 0.697|好瓜 = 否) = \frac{1}{\sqrt{2\pi} \bullet 0.195} \exp(-\frac{(0.697-0.496)^2}{2 \bullet 0.195^2}) \approx 1.203$$

$$\rho_{含糖:0.460|是} = \rho(含糖率 = 0.460|好瓜 = 是) = \frac{1}{\sqrt{2\pi} \bullet 0.101} \exp(-\frac{(0.460-0.279)^2}{2 \bullet 0.101^2}) \approx 0.788$$

$$\rho_{含糖:0.460|否} = \rho(含糖率 = 0.460|好瓜 = 否) = \frac{1}{\sqrt{2\pi} \bullet 0.108} \exp(-\frac{(0.460-0.154)^2}{2 \bullet 0.108^2}) \approx 0.066$$

Posterior Probability:

$$P(好瓜 = 是) \times P_{青绿|是} \times P_{蜷缩|是} \times P_{浊响|是} \times P_{清晰|是} \times P_{凹陷|是} \times P_{硬滑|是} \times \rho_{密度:0.697|是} \times \rho_{含糖:0.460|是} \approx 0.038,$$

$$P(好瓜 = 否) \times P_{青绿|否} \times P_{蜷缩|否} \times P_{浊响|否} \times P_{清晰|否} \times P_{凹陷|否} \times P_{硬滑|否} \times \rho_{密度:0.697|否} \times \rho_{含糖:0.460|否} \approx 6.80 \times 10^{-5}.$$

$0.038 > 6.80 \times 10^{-5}$, test 1 is classified as “好瓜”.

Laplacian Correction

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测2	青绿	蜷缩	清脆	清晰	凹陷	硬滑	0.697	0.460	?

$$P_{\text{清脆}|\text{是}} = \text{P（敲声}=\text{清脆}|\text{好瓜}=\text{是）}=\frac{0}{8}=0$$

$$\hat{P}(c_i)=\frac{|D_c|+1}{|D|+N}$$

$$\hat{P}(x_i \mid c_i)=\frac{|D_{c,x_i}|+1}{|D_c|+N_i}$$

N : the number of class in training set D.

N_i : the number of values for the i-th attribute.

Example

$$P(\text{好瓜}=\text{是})=\frac{8+1}{17+2}\approx 0.474 \qquad P(\text{好瓜}=\text{否})=\frac{9+1}{17+2}\approx 0.526$$

$$P_{\text{青绿|是}}=\mathrm{P}(\text{色泽}=\text{青绿|好瓜}=\text{是})=\frac{3+1}{8+3}=0.364$$

$$P_{\text{青绿|否}}=\mathrm{P}(\text{色泽}=\text{青绿|好瓜}=\text{否})=\frac{3+1}{9+3}\approx 0.333$$

$$P_{\text{清脆|是}}=\mathrm{P}（\text{敲声}=\text{清脆|好瓜}=\text{是}）=\frac{0+1}{8+3}=0.091$$

Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

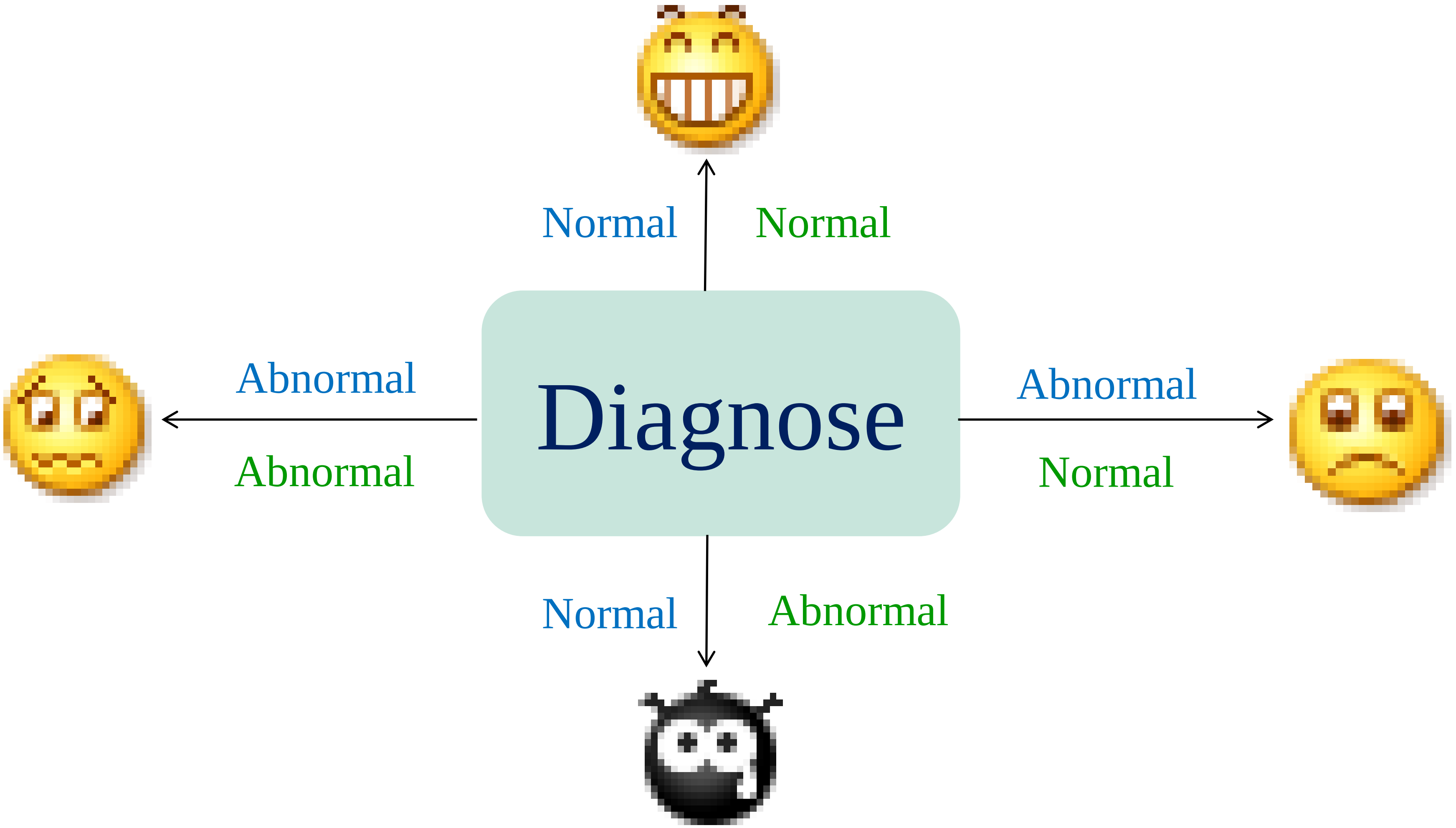
Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Bayes Classifier Extension

- Bayesian Decision Based on Minimal Risk
- Parameter Estimation

Bayesian Decision Based on Minimal Risk



Bayesian Decision Based on Minimal Risk

- $y = \{c_1, c_2, \ldots, c_N\}$: the finite set of N states of labels
- λ_{ij} : the loss incurred by mistaking c_i for c_j
- Given $x = [x_1, x_2, \cdots x_d]^T$
- Conditional risk:

$$\boxed{R(c_i|x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|x)}$$

Bayesian Decision Based on Minimal Risk

- Find a **decision rule** $h: X \mapsto Y$ to minimize the overall risk.

$$R(h) = E_x[R(h(x)|x)]$$

- Choose the label which can minimize the conditional risk for each

$$h*(x) = \underset{c \in y}{argmin} R(c|x)$$

Bayes optimal classifier

Bayesian Decision Based on Minimal Risk

- Assuming that in a local area, the prior probabilities of normal and abnormal in cell recognition are:

  Normal: $P(c_1) = 0.9$

  Abnormal: $P(c_2) = 0.1$

- There is a cell to be identified, the observed value is x, from the class condition probability density distribution curve

$$p(x|c_1) = 0.2, \qquad p(x|c_2) = 0.4$$

- Try to judge whether the cell is normal or abnormal ?

Bayesian Decision Based on Minimal Risk

- The posterior probability of c_1 and c_2 is calculated by Bayesian formula :

$$P(c_1|x) = \frac{P(x|c_1)P(c_1)}{\sum_{j=1}^{2} P(x|c_j)P(c_j)} = \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.4 \times 0.1} = 0.818$$

$$P(c_2|x) = 1 - P(c_1|x) = 0.182$$

- According to Bayesian decision rules

$$P(c_1|x) = 0.818 > P(c_2|x) = 0.182$$

- Decision rules: Normal

Bayesian Decision Based on Minimal Risk

Loss / Real State / Prediction Result	c_1	c_2
c_1	0	6
c_2	1	0

$$\begin{cases} R(c_1|x) = \sum_{j=1}^{N} \lambda_{1,j} P\left(c_j \middle| x\right) = 6 * 0.182 = 1.092 \\ R(c_2|x) = \sum_{j=1}^{N} \lambda_{2,j} P\left(c_j \middle| x\right) = 1 * 0.818 = 0.818 \end{cases}$$

$$\boxed{R(c_1|x) = 1.092 > R(c_2|x) = 0.818}$$

- Decision rules: Abnormal

Bayesian Decision Based on Minimal Risk

- The result of the classification is just the opposite. This is because there is one more factor affecting the decision-making result, namely "loss". And the losses caused by the two types of wrong decisions are very different, so "loss" has played a leading role.

Bayesian Decision Based on Minimal Risk

Prediction Result \ Loss \ Real State	c_i	c_j
c_i	0	1
c_j	1	0

$$\lambda_{i,j}\begin{cases} 0, & if\ i = j; \\ 1, & otherwise, \end{cases}$$

An example (Suppose N=4, i=3)

$$R(c_i|x) = \sum_{j=1}^{N} \lambda_{i,j} P(c_j|x)$$

$$= \lambda_{31}P(c_1|x) + \lambda_{32}P(c_2|x) + \lambda_{33}P(c_3|x) + \lambda_{34}P(c_4|x)$$

$$= P(c_1|x) + P(c_2|x) + P(c_4|x)$$

$$= 1 - P(c_3|x)$$

Bayesian Decision Based on Minimal Risk

- Conditional risk: $R(c|x) = 1 - P(c|x)$
- Recall: $h*(x) = \underset{c \in y}{argmin} R(c|x)$
- Bayes optimal classifier : $\boxed{h*(x) = \underset{c \in y}{\mathrm{argmax}} P(c|x)}$

An example (y=2)

$P(c_1|x) > P(c_2|x) \rightarrow h*(x) = c_1$

$P(c_1|x) < P(c_2|x) \rightarrow h*(x) = c_2$

Bayes Classifier Extension

- Bayesian Decision Based on Minimal Risk
- Parameter Estimation

Parameter Estimation

The method of estimating the overall probability distribution from the sample set can be summarized as follows:

- ✓ Supervised Parameter Estimation (*)
- ✓ Unsupervised Parameter Estimation
- ✓ Non-parametric Estimation

Parameter Estimation

- Supervised Parameter Estimation:

The categories and conditions to which the sample belongs are known in the form of the overall probability density function, and some of the parameters that characterize the probability density function are unknown.

For example, only the overall distribution of the sample is known, and the parameters of the normal distribution are unknown. Our goal is to statistically judge some of the population distribution from a set of samples of a known category. The estimate in this case is called the parameter estimation under supervision.

Parameter Estimation

- Unsupervised Parameter Estimation:

The overall probability density function is known, but the category to which the sample belongs is unknown and it is required to determine some parameters of the probability density function.

Supervised and unsupervised means whether the category to which the sample belongs is known or unknown. There are two commonly used methods, one is the maximum likelihood estimation, and the other is Bayesian estimation.

Parameter Estimation

- The maximum likelihood estimation is that the parameters are regarded as definite and unknown, and the best estimate is obtained under the condition that the probability of obtaining the actual observed sample is maximum.
- The Bayesian estimation treats the unknown parameters as a random variable with a certain distribution. The observed result of the sample transforms the prior distribution into a posterior distribution, and then corrects the original estimate of the parameter based on the posterior distribution.

Parameter Estimation

- Non-parametric Estimation：

The category to which the sample belongs is known, but the form of probability density function is unknown, so it requires us to directly infer the probability density function itself.

Parameter Estimation

The method of estimating the overall probability distribution from the sample set can be summarized as follows:

- ✓ Supervised Parameter Estimation (*)
- ✓ Unsupervised Parameter Estimation
- ✓ Non-parametric Estimation

Kn Nearest Neighbor

Parzen Window

Parzen Window

The number of samples in each region

The total number of samples

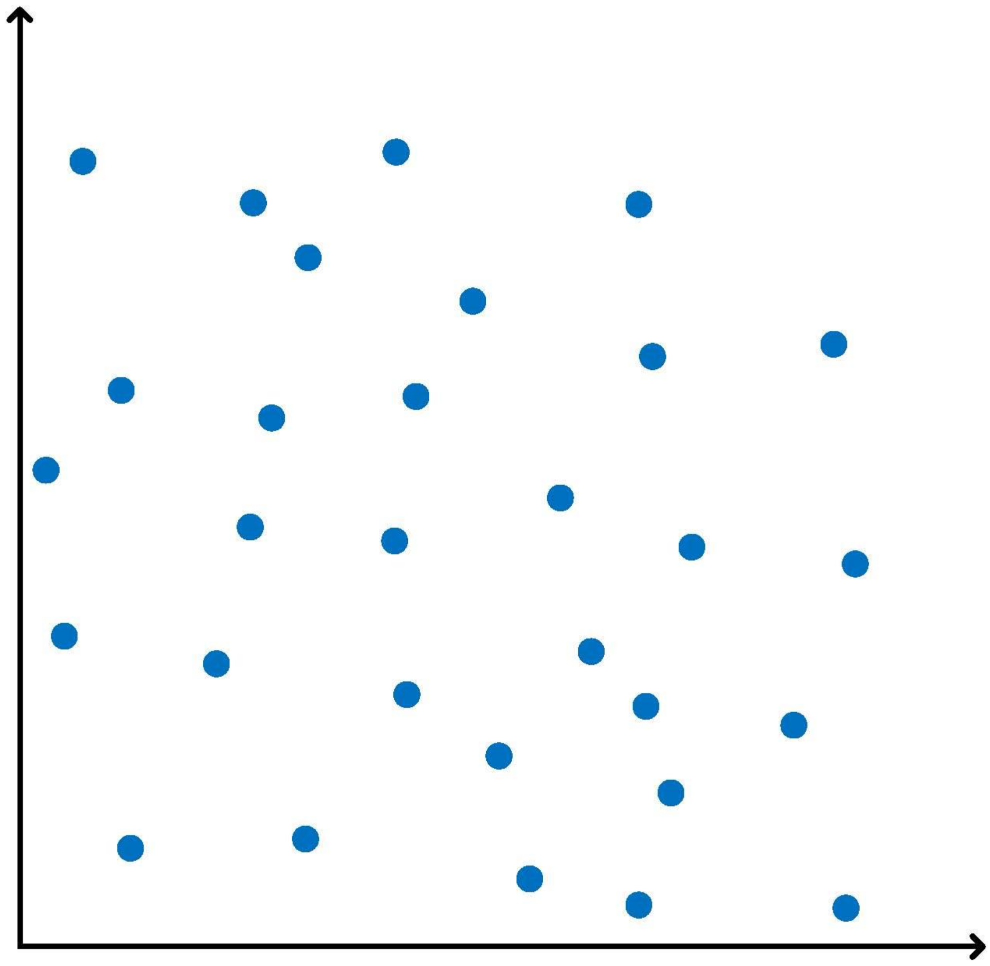
$$p(x) = \frac{k_N/N}{V}$$

Probability density estimation

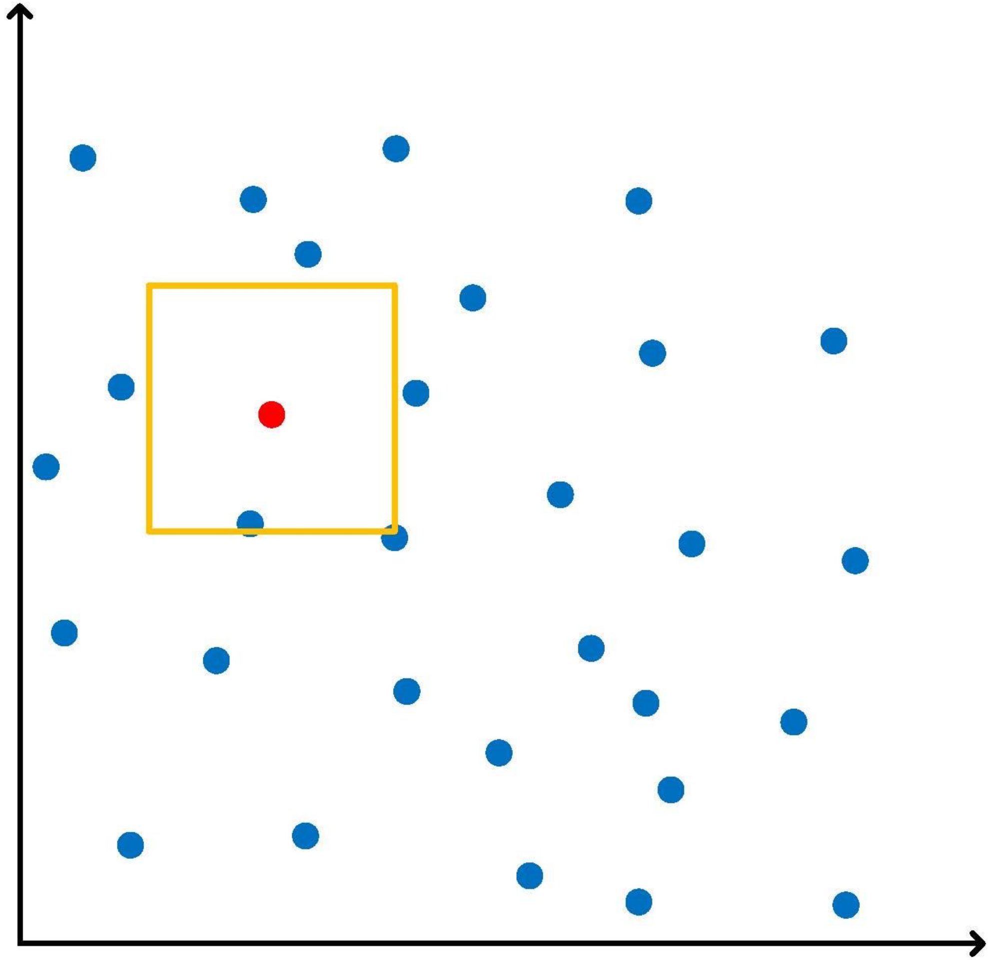
固定$\boldsymbol{V}$，找$\boldsymbol{k_N}$

Volume

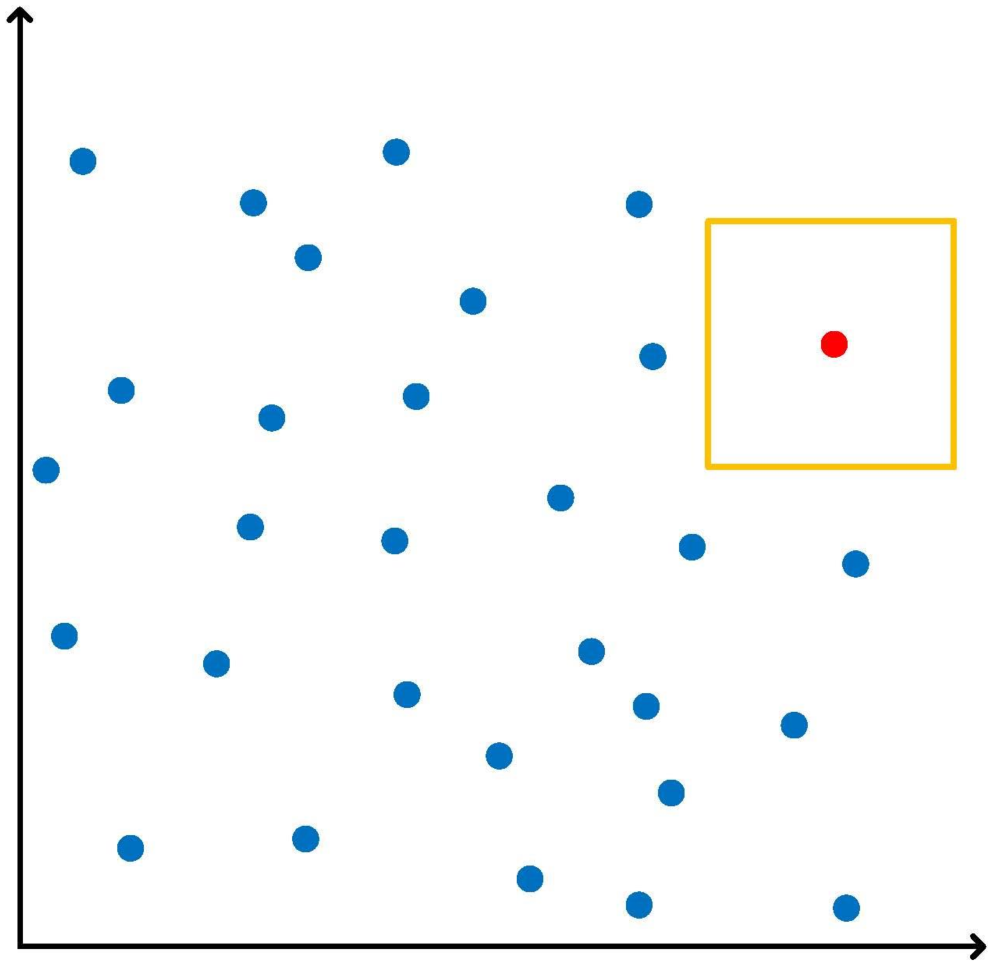
Parzen Window



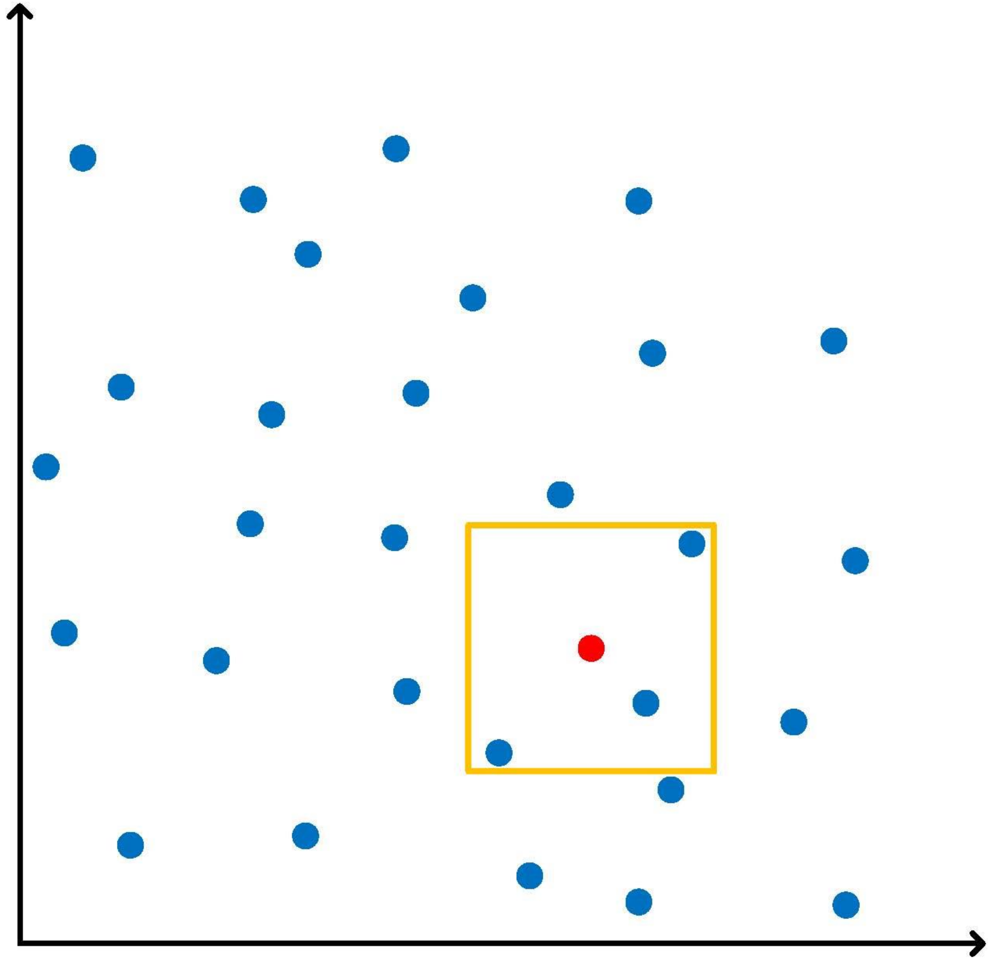
Parzen Window



Parzen Window



Parzen Window



Parzen Window

- Two-dimensional plane :

  Square

- Three-dimensional plane :

  Cube

- N-dimensional plane :

  Hypercube

Parameter Estimation

The method of estimating the overall probability distribution from the sample set can be summarized as follows:

- ✓ Supervised Parameter Estimation (*)
- ✓ Unsupervised Parameter Estimation
- ✓ Non-parametric Estimation

Kn Nearest Neighbor

Parzen Window

Kn Nearest Neighbor

The number of samples in each region

The total number of samples

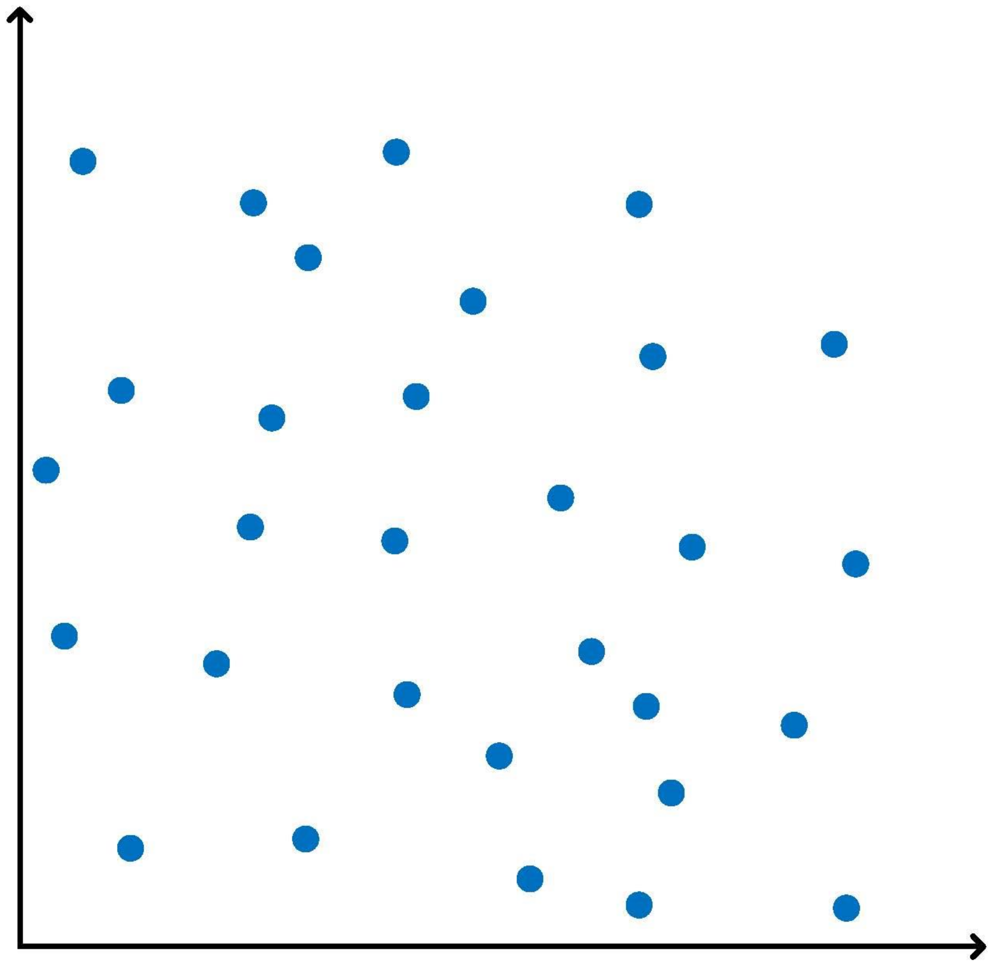
$$p(x) = \frac{\boxed{k_N} / \boxed{N}}{\boxed{V}}$$

Probability density estimation

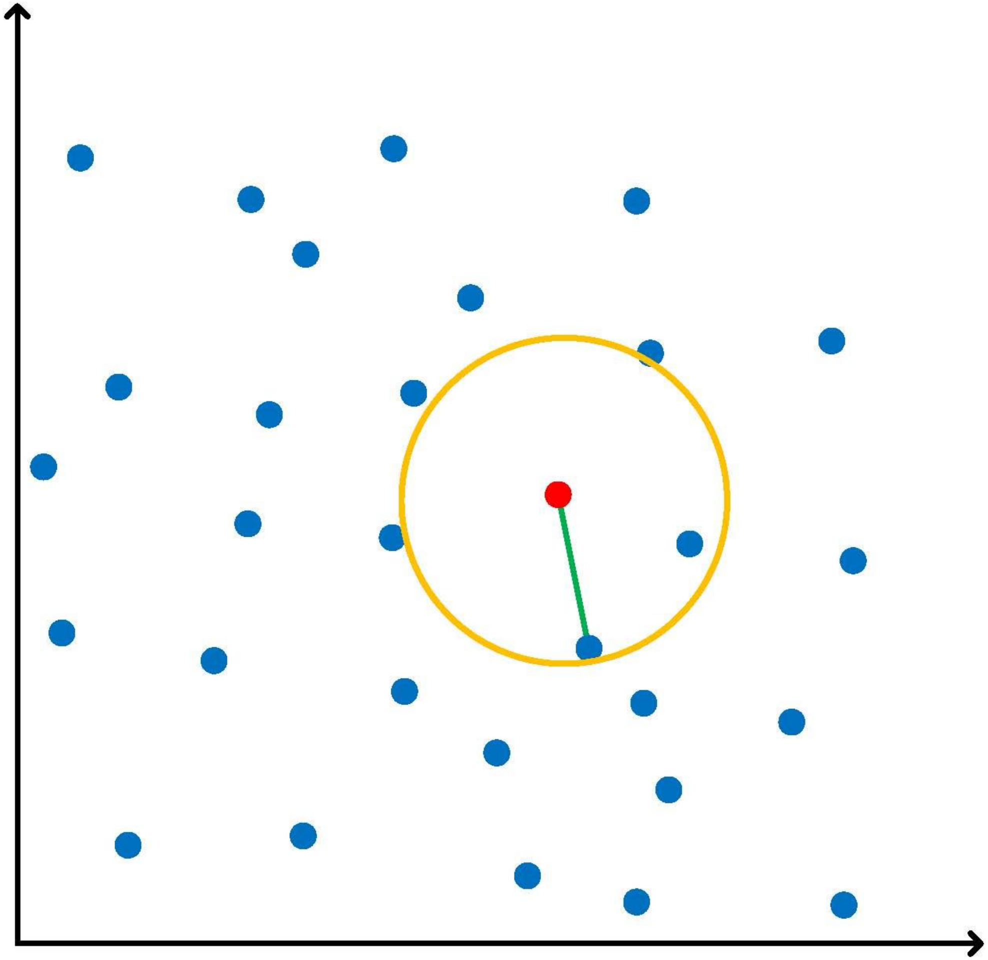
固定k_N，找V

Volume

Kn Nearest Neighbor

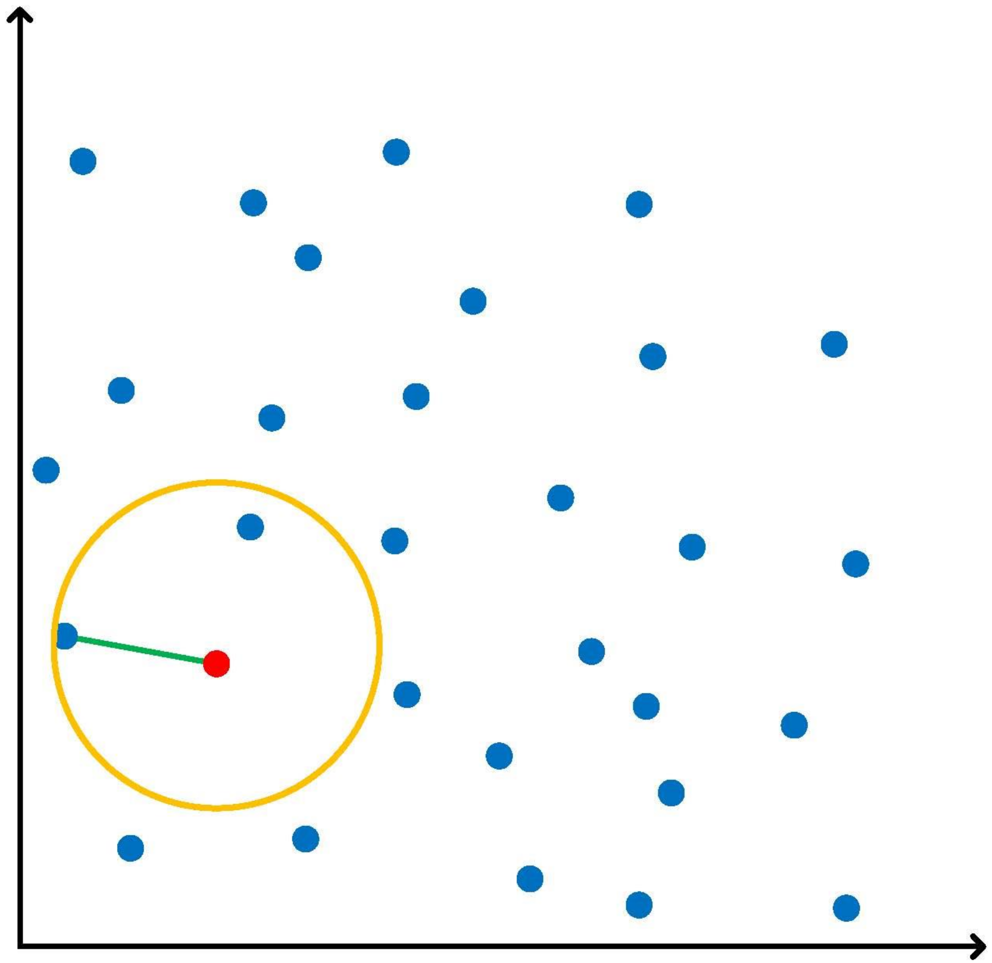


Kn Nearest Neighbor



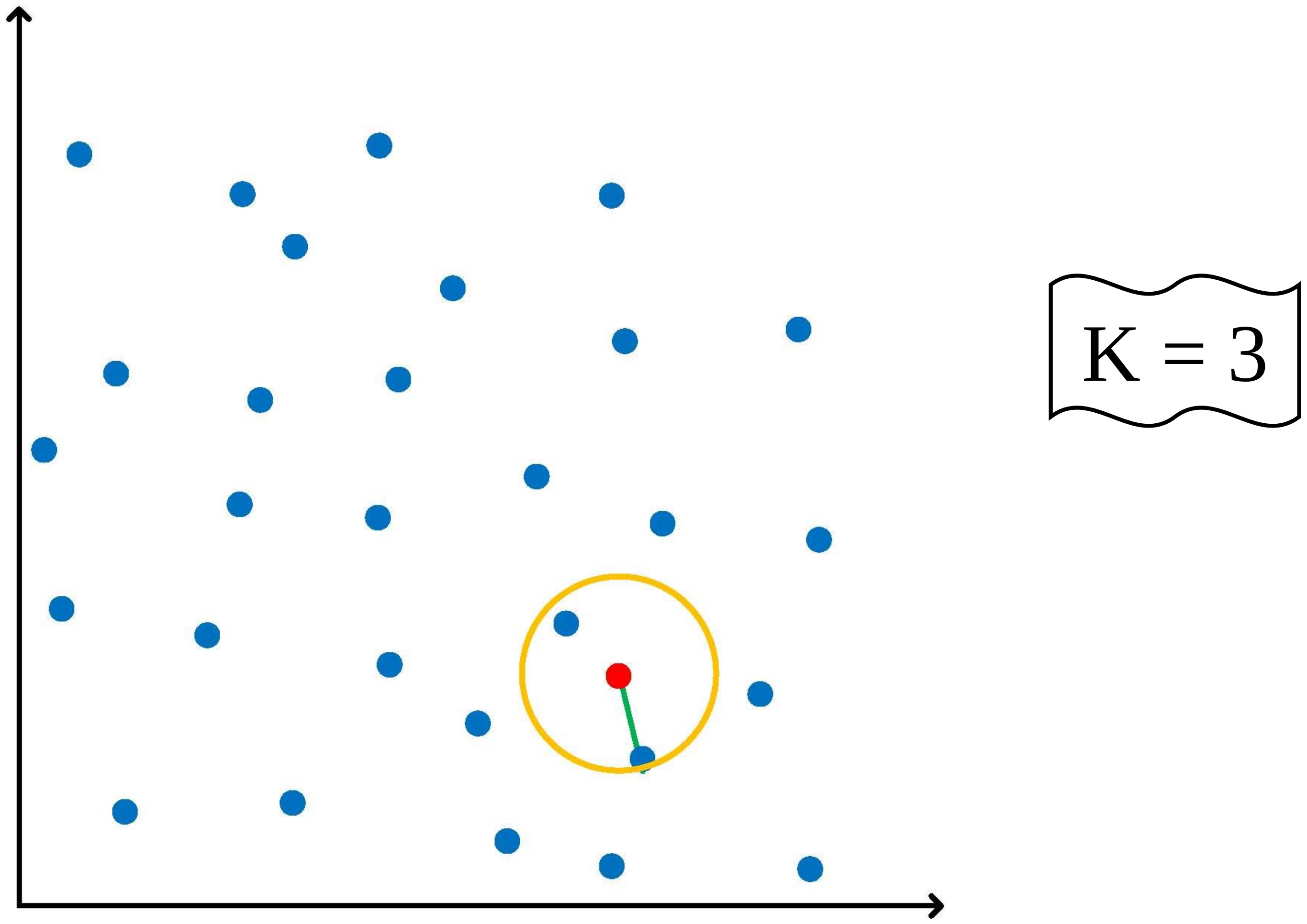
K = 3

Kn Nearest Neighbor



K = 3

Kn Nearest Neighbor



Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

Outline

- Bayesian Decision Based on Minimum Error Rate
- Maximum Likelihood Estimation
- Naïve Bayes Classifier
- Bayes Classifier Extension
- Semi-naïve Bayes Classifier
- Bayesian Application Examples

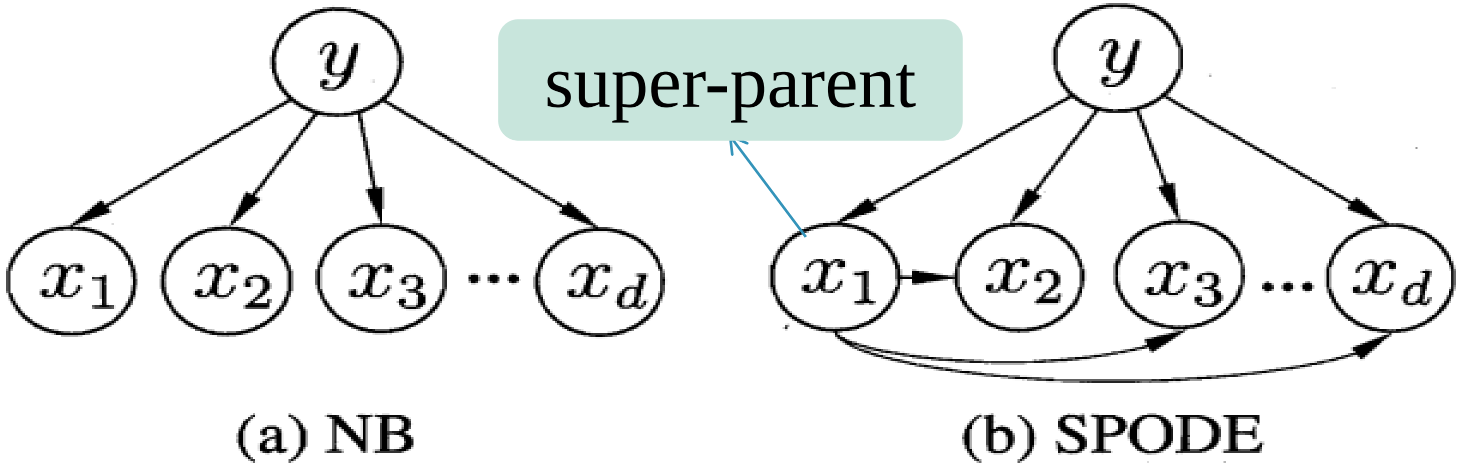
Semi-naïve Bayes Classifier

- In practical scenarios, the attribute independence assumption is often violated!
- ODE (One-Dependent Estimator): Suppose an attribute only relies on a most other properties.

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

pa_i : the parent-node of x_i

Super-Parent

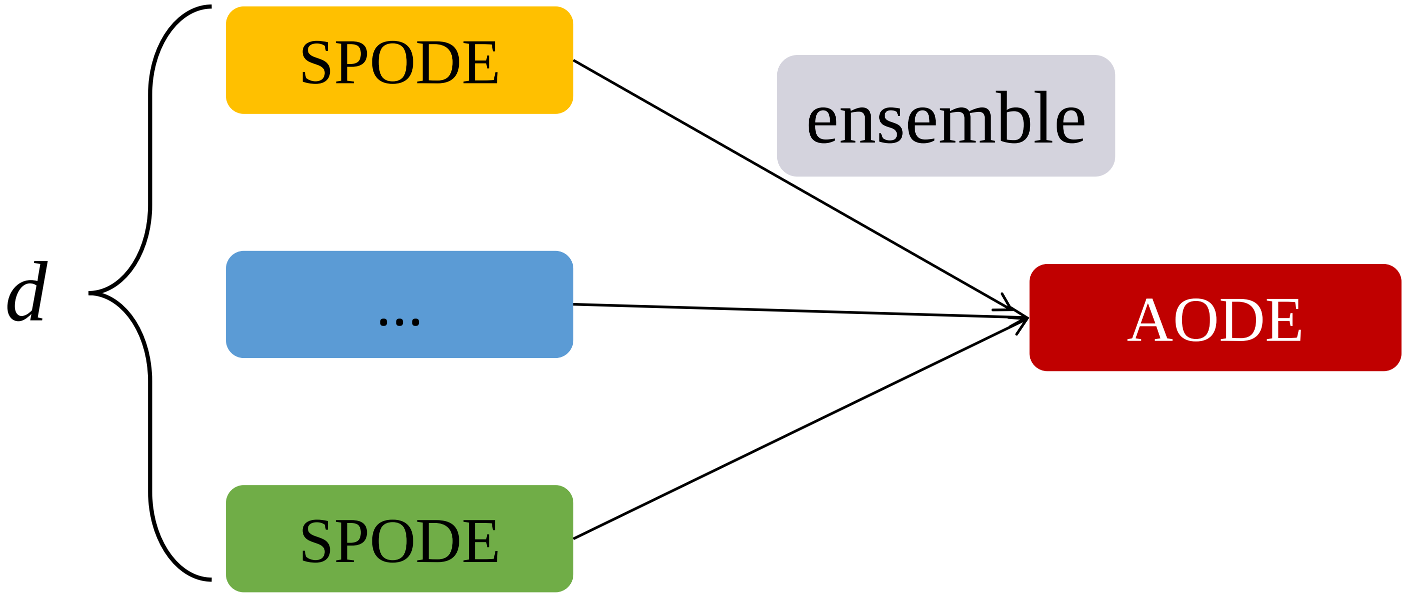


SPODE (Super-Parent ODE): assume that all attributes depend on the same attribute.

$$P(c \mid x) \propto P(c) \prod_{i=1}^{d} P(x_i \mid c, x_1)$$

AODE

- AODE: Averaged One-Dependent Estimator



SPODE which has enough statistical data to support.

AODE

$$P(c \mid x) \propto \sum_{\substack{i=1 \\ |D_{x_i}| \geq m'}}^{d} \boxed{P(c, x_i)} \prod_{j=1}^{d} \boxed{P(x_j \mid c, x_i)}$$

Laplacian Correction

$$\hat{P}(c, x_i) = \frac{|D_{c,x_i}| + 1}{|D| + N_i}$$

$$\hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j}$$

m : threshold constant.

D_{x_i} : a set of samples that take value x_i on the i - th attribute.

N_i : the number of values for the i - th attribute.

Example

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

$$P_{是,浊响} = P(好瓜=是, 敲声=浊响) = \frac{6+1}{17+3} = 0.350,$$

$$P_{凹陷|是,浊响} = P(脐部=凹陷|好瓜=是, 敲声=浊响) = \frac{3+1}{6+3} \approx 0.444.$$

Conclusions

- $h^*(x) = \arg\max_{c \in y} P(c \mid x)$
- MLE (parameter estimate)
- Naïve Bayes Classifier (attribute conditional independence assumption)
- Bayes Extension
- Semi-naïve Bayes Classifiers (ODE)

How to get the P(c|x)?

- Discriminative models (判别式模型)

eg: Support Vector Machines, Decision Tree, BP Neural Network

- Generative models (生成式模型)

eg: Naïve Bayes, AODE, Restricted Boltzmann Machine

$$P(c_i|x) = \frac{P(x, c_i)}{P(x)} = \frac{P(c_i)P(x|c_i)}{\sum_{i=1}^{N} P(x|c_i)P(c_i)}$$

Extension

- Bayesian network
- Expectation-Maximization

Extension

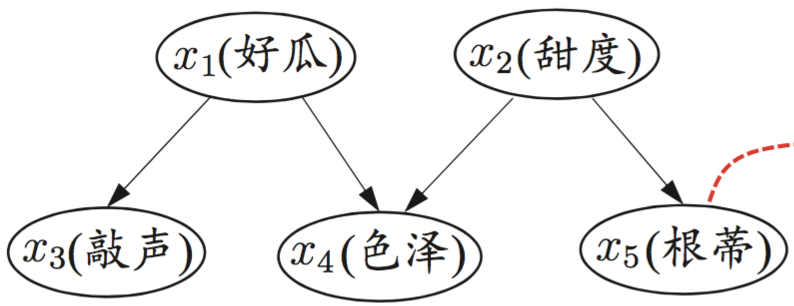
- Bayesian network
- Expectation-Maximization

Bayesian network

Definition: $B = < G, \Theta >$

G : directed acyclic graph (BN's **structure**)

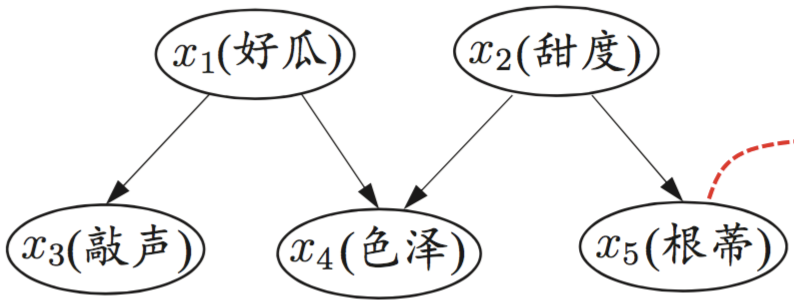
Θ: Conditional Probability Table(CPT)，$\theta_{x_i|\pi_i} = P_B(x_i \mid \pi_i)$



		根蒂: 硬挺	根蒂: 蜷缩
甜度	高	0.1	0.9
	低	0.7	0.3

G & CPT

Example



		根蒂	
		硬挺	蜷缩
甜度	高	0.1	0.9
	低	0.7	0.3

G & CPT

From G -> “色泽”直接依赖于“好瓜”和“甜度”

From CPT -> “根蒂”对“甜度”的量化依赖关系 P(根蒂＝硬挺|甜度＝高)＝0. 1

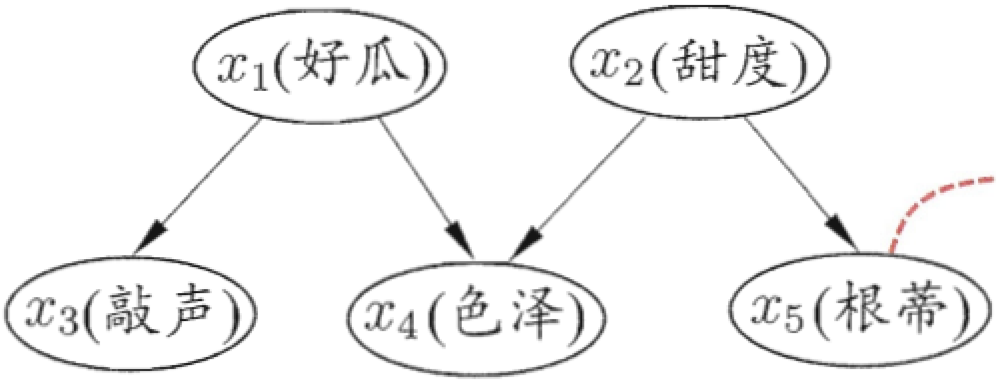
Bayesian network

The joint probability distribution of $x_1, x_2, \ldots, x_d$

$$P_B(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i \mid \pi_i}$$

Given parent-node set, an attribute is independent with its non-descendant attribute.

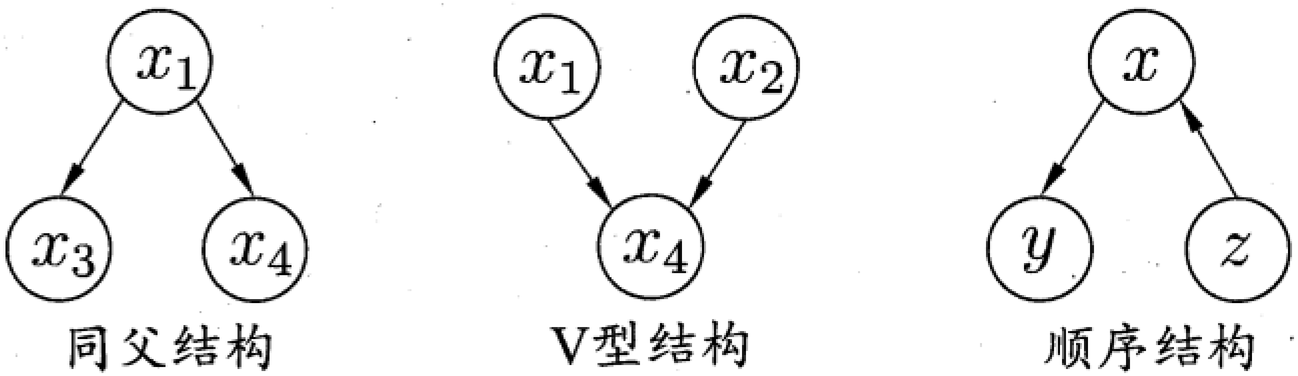
Example



$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3 \mid x_1)P(x_4 \mid x_1, x_2)P(x_5 \mid x_2)$$

Given x_1,x_3 is independent with x_4 , given x_2 ,x_4 is independent with x_5. $x_3 \perp x_4 \mid x_1$ $\quad x_4 \perp x_5 \mid x_2$

Structure



Common parent structure: Given x_1 , $x_3 \perp x_4 \mid x_1$

Sequential structure: Given x , $y \perp z$

V-structure

V-structure(collision structure), given x_4 , x_1 is dependent with x_2.

Surprisely, when the value of x_4 is unknow , x_1 is independent with x_2.

$$
\begin{aligned}
P(x_1, x_2) &= \sum_{x_4} P(x_1, x_2, x_4) \\
&= \sum_{x_4} P(x_4 \mid x_1, x_2) P(x_1) P(x_2) \\
&= P(x_1) P(x_2) .
\end{aligned}
$$

Extension

- Bayesian network
- Expectation-Maximization

Extension

- Bayesian network
- Expectation-Maximization

Expectation-Maximization

X : observed variable

Θ : model parameter

Z : latent variable (隐变量)

EM is a method to find θ_{ML} where

$$LL(\Theta \mid X, Z) = \ln P(X, Z \mid \Theta)$$

marginal likelihood (最大化边际似然）

$$LL(\Theta \mid X) = \ln P(X \mid \Theta) = \ln \sum_Z P(X, Z \mid \Theta)$$

Basic idea

- E-step: calculate(基于Θ^t推断隐变量Z的期望)

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^t) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\Theta}^t} LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z})$$

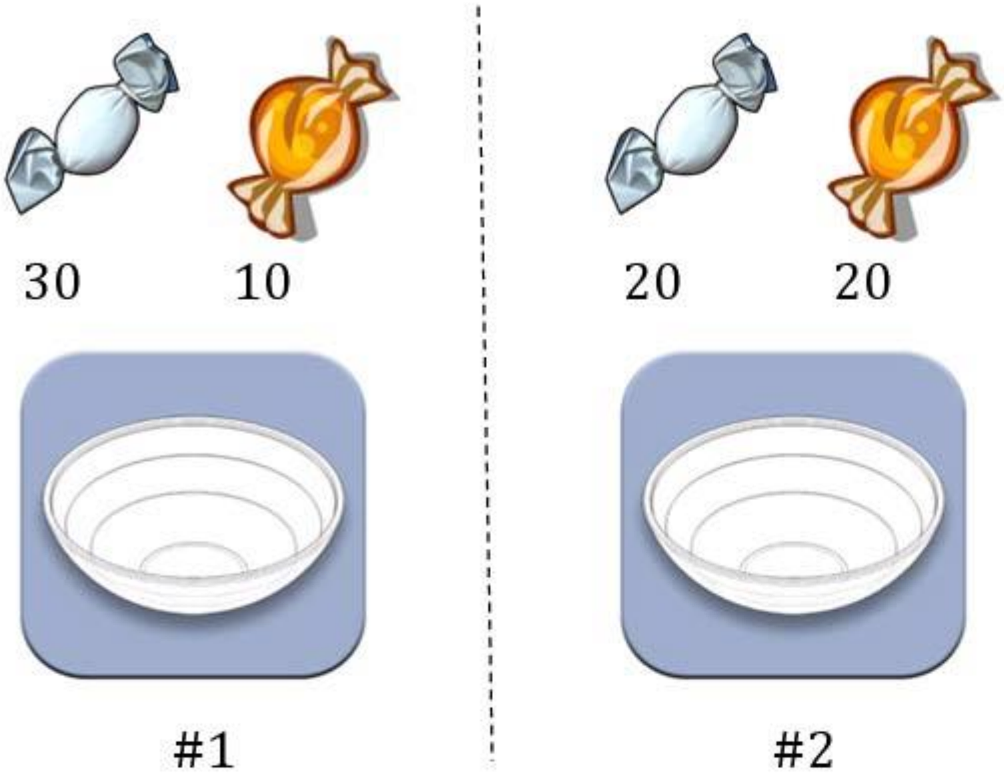
- M-step: find (基于已观测到变量X和Z^t对参数$\boldsymbol{\Theta}$做极大似然估计，记为 $\boldsymbol{\Theta}^{t+1}$;)

$$\boldsymbol{\Theta}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^t)$$

Homework

- ✓ Naïve Bayesian advantages and disadvantages?
- ✓ Three conditions of Naïve Bayesian?
- ✓ What is MLE?
- ✓ What is Naïve Bayes?
- ✓ What is EM?

Homework



两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大？

-- 水果糖问题

Homework

假阳性问题

已知某种疾病的发病率是0.001，即1000人中会有1个人得病。现有一种试剂可以检验患者是否得病，它的准确率是0.99，即在患者确实得病的情况下，它有99%的可能呈现阳性。它的误报率是5%，即在患者没有得病的情况下，它有5%的可能呈现阳性。现有一个病人的检验结果为阳性，请问他确实得病的可能性有多大？

Homework

8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的抢射击，中靶概率为0.3；现从8支抢中随机取一支射击，结果中靶。求该枪是已校准过的概率。

-- 射击问题

Thanks !