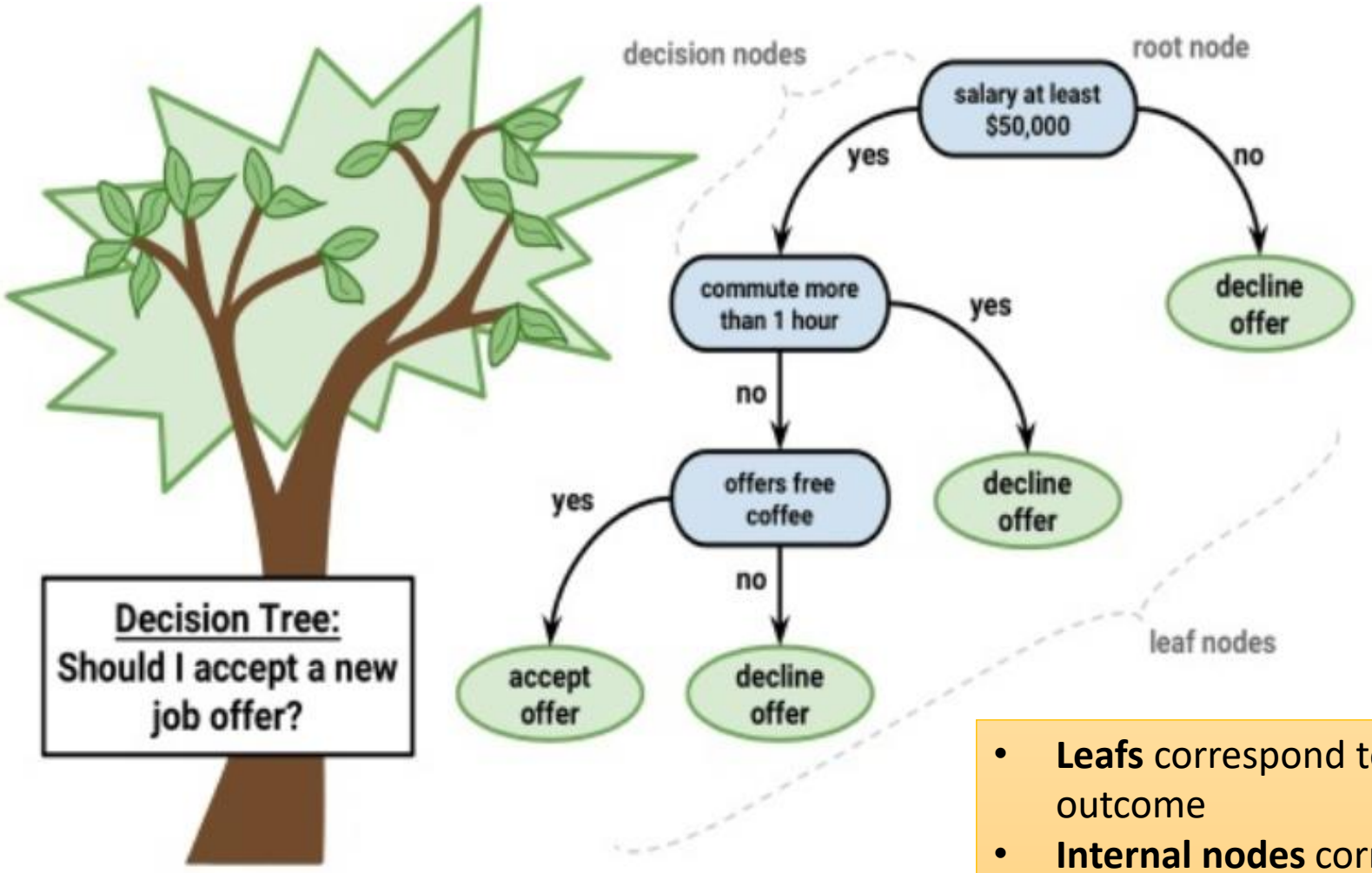
Machine Learning

Bayes Classifier

Dr. Shuang LIANG

Recall: Decision Tree



- **Leafs** correspond to classification outcome
- **Internal nodes** correspond to attributes (features)
- **Edges** denote assignment

Recall: Identify the best attribute

- Entropy

$$Ent\,(X) = \sum_{c} -p(X = c) log_2 p(X = c)$$

- Information Gain

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

Today's Topics

- Type of classifiers
- Bayesian decision theory
- Naïve Bayes Classifier
- Bayesian Network

Today's Topics

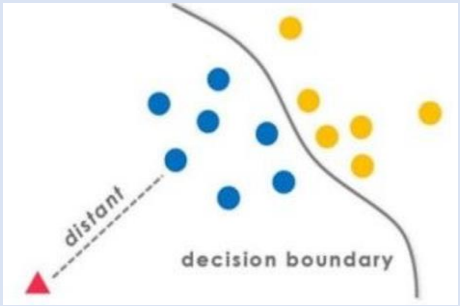
- ***Type of classifiers***
- Bayesian decision theory
- Naïve Bayes Classifier
- Bayesian Network

Types of Classifiers

Model-based

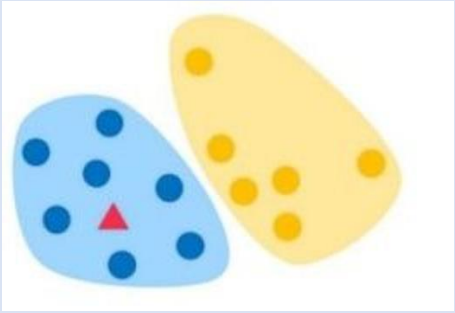
No Model

Discriminative
directly estimate a decision rule/boundary



Logistic regression
Decision tree
Neural network
……

Generative
build a generative statistical model



Naïve Bayes
Bayesian Networks
HMM
……

Instance-based
Use observation directly

KNN

Discriminative

- Only care about estimating the conditional probabilities $P(y|x)$
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

Generative

- Model observations (x, y) first ($P(x, y)$), then infer $P(y|x)$
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

Today's Topics

- Type of classifiers
- ***Bayesian decision theory***
- Naïve Bayes Classifier
- Bayesian Network

Bayesian decision theory

- The basic method for implementing decision-making in a probabilistic framework
- In classification tasks, bayesian decision theory selects optimal class labels based on known relevant probabilities and misclassification losses

Bayesian decision theory

- Suppose there are N possible class labels

$$y = \{c_1, c_2, \dots, c_N\}$$

- Then the expected loss of classifying sample x as c_i is

$$\underbrace{R(c_i|x)}_{\textbf{Conditional Risk}} = \sum_{j=1}^{N} \lambda_{ij} P(c_j|x)$$

λ_{ij} : Loss for misclassifying a true sample labeled c_j as c_i

$P(c_j|x)$: Posterior probability

- Our task is to find a decision rule $h: X \to Y$ that minimizes the overall risk

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

- For each sample x, if h can minimize the conditional risk $R(h(x)|x)$, then the overall risk $R(h)$ will also be minimized

Bayesian decision rule

- To minimize the overall risk, we can simply select the class that minimizes the conditional risk $R(c|x)$ on each sample

$$h^*(x) = argmin\, R(c|x) \;\; (c \in y)$$

- $h^*(x)$: Bayes optimal classifier
- $R(h^*)$: Bayes risk
- $1 - R(h^*)$: The best performance the classifier can achieve

Case Study

- We need to minimize *the classification error rate*, then
- Misclassification loss

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

- Conditional risk

$$R(c|x) = 1 - \ P(c|x)$$

- Bayes optimal classifier

$$h^*(x) = argmax \ P(c|x) \ \ (c \in y)$$

For each sample, select the class that **maximizes the posterior probability**

Thinking

- If we use the Bayes decision rule to minimize the decision risk, the first step is to obtain the posterior probability $P(c|x)$
- Is $P(c|x)$ easily and directly obtainable in reality? 

Explaining Machine Learning from a **Probabilistic View**

What machine learning wants to achieve is to ***estimate*** the posterior probability as accurately as possible based on limited training samples

Strategies

- There are two main strategies for estimating the posterior probability
- Discriminative models
  - Given x, predict c by directly modeling $P(c|x)$
  - Decision tree, Neural Network(based on BP), SVM…
- Generative Models
  - First model the joint probability distribution $P(x, c)$, and then obtain $P(c|x)$ from it

Generative Models

$$P(c|x) = \frac{P(x,c)}{P(x)}$$

Class-conditional probability of sample x relative to class c

- According to Bayes' theorem, $P(c|x)$ can be written as

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

Prior Probability
can be estimated by the frequency of occurrence of various samples

"Evidence" factor, independent of the class

Today's Topics

- Type of classifiers
- Bayesian decision theory
- ***Naïve Bayes Classifier***
- Bayesian Network

Generative Models

$$P(c|x) = \frac{P(x,c)}{P(x)}$$

- According to Bayes' theorem, $P(c|x)$ can be written as

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

- How to get $P(x|c)$?

Naïve Bayes Classifier

- $P(x|c)$ is the joint probability over all attributes, which is difficult to estimate directly from limited training samples
- **Attribute conditional independence assumption**: For known classes, all attributes are assumed to be independent of each other
- Based on this assumption, we can get:

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

d is the number of attributes, x_i is the value of x on the i-th attribute

Naïve Bayes Classifier

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

- $P(x)$ is the same for all classes, so according to the Bayesian decision rule (on page 11):

$$h_{nb}(x) = argmax\ P(c) \prod_{i=1}^{d} P(x_i|c) \quad (c \in y)$$

- This is the expression for the **Naive Bayes classifier**

Train a Naïve Bayes Classifier

- Estimate the class prior probability $P(c)$ based on the training set and estimate the conditional probability $P(x_i \mid c)$ for each attribute.
- How to get $P(c)$?
- Let D_c denote the set of samples of class c in the training set D, If there are sufficient $i.i.d$ samples

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

Train a Naïve Bayes Classifier

- Estimate the class prior probability $P(c)$ based on the training set and estimate the conditional probability $P(x_i \mid c)$ for each attribute.
- How to get $P(x_i \mid c)$?
- For ***discrete attributes***, Let D_{c,x_i} denote the set of samples in D_c with the value x_i on the i-th attribute, then

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

Train a Naïve Bayes Classifier

- Estimate the class prior probability $P(c)$ based on the training set and estimate the conditional probability $P(x_i \mid c)$ for each attribute.
- How to get $P(x_i \mid c)$?
- For ***continuous attributes***, assume that $p(x_i|c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$, where $\mu_{c,i}$ and $\sigma_{c,i}^2$ are the mean and variance of the value of the c-class sample on the i-th attribute, respectively. Then

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

A watermelon case study

- We have a watermelon data set

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

A watermelon case study

- Now we want to classify test sample 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	?

A watermelon

- Step 1: Get $P(c)$

编号	色泽	根蒂	敲声
测 1	青绿	蜷缩	浊响

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

$$P(\text{好瓜} = \text{是}) =$$

$$P(\text{好瓜} = \text{否}) =$$

?

A watermelon

- Step 2: Get $P(x_i|c)$ for e

编号	色泽	根蒂	敲声	纹
测 1	青绿	蜷缩	浊响	清

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

$$P_{青绿|是} = P(色泽 = 青绿 \mid 好瓜 = 是) =$$

$$P_{青绿|否} = P(色泽 = 青绿 \mid 好瓜 = 否) =$$

$$P_{蜷缩|是} = P(根蒂 = 蜷缩 \mid 好瓜 = 是) =$$

$$P_{蜷缩|否} = P(根蒂 = 蜷缩 \mid 好瓜 = 否) =$$

$$P_{浊响|是} = P(敲声 = 浊响 \mid 好瓜 = 是) =$$

$$P_{浊响|否} = P(敲声 = 浊响 \mid 好瓜 = 否) =$$

?

$$P_{清晰|是} = P(纹理 = 清晰 \mid 好瓜 = 是) =$$

$$P_{清晰|否} = P(纹理 = 清晰 \mid 好瓜 = 否) =$$

$$P_{凹陷|是} = P(脐部 = 凹陷 \mid 好瓜 = 是) =$$

$$P_{凹陷|否} = P(脐部 = 凹陷 \mid 好瓜 = 否) =$$

$$P_{硬滑|是} = P(触感 = 硬滑 \mid 好瓜 = 是) =$$

$$P_{硬滑|否} = P(触感 = 硬滑 \mid 好瓜 = 否) =$$

?

A watermelon

- Step 2: Get $P(x_i|c)$ for e

编号	色泽	根蒂	敲声	纹
测 1	青绿	蜷缩	浊响	清

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

$$p_{\text{密度: 0.697}|\text{是}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{是})$$

?

$$p_{\text{密度: 0.697}|\text{否}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{否})$$

?

A watermelon case study

- Step 2: Get $P(x_i|c)$ for every attribute (*continuous*)

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	?

$$p_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是})$$

?

$$p_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否})$$

?

A watermelon case study

- Step 3: Make decision

$$h_{nb}(x) = argmax\ P(c) \prod_{i=1}^{d} P(x_i|c) \quad (c \in y)$$

$$P(\text{好瓜} = \text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}}$$
$$\times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \quad =?$$
$$P(\text{好瓜} = \text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}}$$
$$\times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \quad =?$$

- $0.038 > 6.8 \times 10^{-5} \rightarrow$ 好瓜

Special case

- What if an attribute value does not appear at the same time with a class in the training set?
- For example, for a test sample with attribute value “敲声=清脆",

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

- Then we will find that $P(c)\prod_{i=1}^{d} P(x_i|c) = 0$!
- The classification result will be “好瓜=否” even if it is obviously like a good melon in other attributes.

Unreasonable!

Laplacian correction

- “Smoothing" is usually done when estimating probabilities
- Laplacian correction

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

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

- N: The number of possible classes in the training set D
- N_i: The number of possible values of the i-th attribute
- Example: For $P(c)$ in the watermelon case,

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

Have a try!

- Try estimating the following probabilities using Laplacian correction

$$\hat{P}_{\text{青绿}|\text{是}} = \hat{P}(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{是})$$

$$\hat{P}_{\text{青绿}|\text{否}} = \hat{P}(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{否})$$

$$\hat{P}_{\text{清脆}|\text{是}} = \hat{P}(\text{敲声} = \text{清脆} \mid \text{好瓜} = \text{是})$$

$$\frac{4}{11}, \frac{4}{12}, \frac{1}{11}$$

Have a try!

- Try estimating the follo
  correction

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

$$\hat{P}_{青绿|是} = \hat{P}(色泽 = 青绿 \mid 好瓜 = 是)$$

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

- $|D_c| = 8$: 有8个好瓜
- $N_i = 3$ ： 色泽属性有3个取值
- $|D_{c,x_i}| = 3$ ： 色泽青绿的好瓜有3个

故结果为:

$$\frac{3+1}{8+3} = \frac{4}{11}$$

Have a try!

- Try estimating the follo
  correction

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

$$\hat{P}_{青绿|否} = \hat{P}(色泽 = 青绿 \mid 好瓜 = 否)$$

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

- $|D_c| = 9$: 有9个坏瓜
- $N_i = 3$: 色泽属性有3个取值
- $|D_{c,x_i}| = 3$: 色泽青绿的坏瓜有3个

故结果为:

$$\frac{3+1}{9+3} = \frac{4}{12}$$

Have a try!

- Try estimating the follo
  correction

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

$$\hat{P}_{清脆|是} = \hat{P}(敲声 = 清脆 \mid 好瓜 = 是)$$

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

- $|D_c| = 8$: 有8个好瓜
- $N_i = 3$: 敲声属性有3个取值
- $|D_{c,x_i}| = 0$: 敲声清脆的好瓜有0个

故结果为:

$$\frac{0+1}{8+3} = \frac{1}{11}$$

How to use Naïve Bayes Classifier

- There are many ways to use Naïve Bayes Classifier in real-world tasks
- When the task requires high prediction speed
  - Calculate and store the probability first, and directly “look up the table” when predicting
- When data changes frequently
  - Lazy learning
- When data keeps increasing
  - Incremental learning

Practice

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Classify sample X

X=(Outlook = Sunny,
Temperature = Cool,
Humidity = High,
Wind = Strong)

Play Tennis = ?

No

Practice

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Classify sample X

X=(Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)

Play Tennis = ?

Step 1: Get $P(c)$

$$P(Play\ Tennis = Yes) = \frac{9}{14}$$

$$P(Play\ Tennis = No) = \frac{5}{14}$$

Practice

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Classify sample X

X=(Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)

Play Tennis = ?

Step 2: Get $P(x_i|c)$ for every attribute

$$P(Outlook = Sunny \mid Play\ Tennis = Yes) = \frac{2}{9}$$

$$P(Outlook = Sunny \mid Play\ Tennis = No) = \frac{3}{5}$$

$$P(Temp = Cool \mid Play\ Tennis = Yes) = \frac{3}{9}$$

$$P(Temp = Cool \mid Play\ Tennis = No) = \frac{1}{5}$$

$$P(Humidity = High \mid Play\ Tennis = Yes) = \frac{3}{9}$$

$$P(Humidity = High \mid Play\ Tennis = No) = \frac{4}{5}$$

$$P(Wind = Strong \mid Play\ Tennis = Yes) = \frac{3}{9}$$

$$P(Wind = Strong \mid Play\ Tennis = No) = \frac{3}{5}$$

Practice

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Classify sample X

X=(Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)

Play Tennis = ?

Step 3: Make decision

$$h_{nb}(x) = argmax\ P(c) \prod_{i=1}^{d} P(x_i|c) \quad (c \in y)$$

$$P(Play = Yes) \prod_{i=1}^{d} P(x_i|Play = Yes)$$

$$= \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \approx 0.0053$$

$$P(Play = No) \prod_{i=1}^{d} P(x_i|Play = No)$$

$$= \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \approx 0.02$$

Prediction: Play Tennis = No

Today's Topics

- Type of classifiers
- Bayesian decision theory
- Naïve Bayes Classifier
- ***Bayesian Network***

The problem of Naïve Bayes

- In most cases, the assumption of conditional independence given the class label is violated.
- For example, much more likely to find the word Barack if we saw the word Obama regardless of the class
- There are models that can improve upon this assumption without using the full conditional model
  - **Semi-naïve Bayes Classifiers**: appropriately considers the interdependence information among some attributes.
  - ***Bayesian Network***

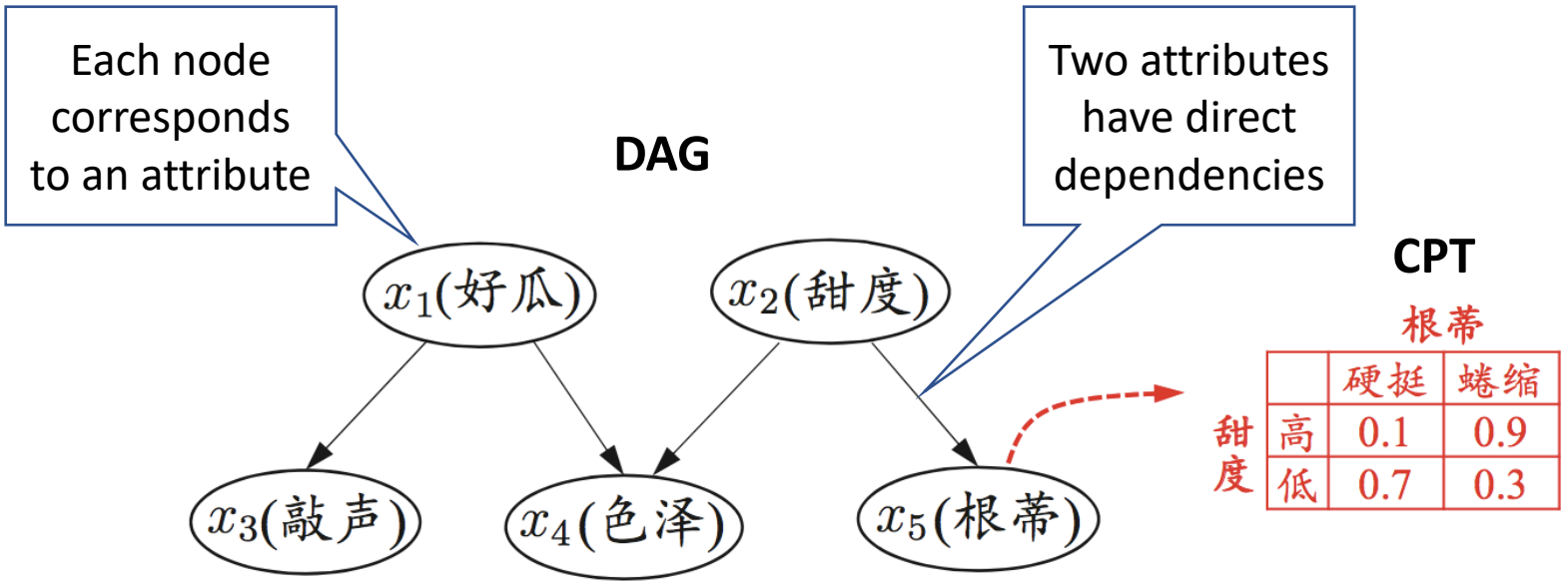
Bayesian Network

- Also called belief network
- It uses the **Directed Acyclic Graph (DAG)** to describe the dependencies between attributes.
- It uses **Conditional Probability Table (CPT)** to describe the joint probability distribution of attributes

Bayesian Network

- A Bayesian network B consists of two parts, the structure G and the parameter Θ, i.e. $B =< G, \Theta >$
- The network structure G is a DAG
- The parameter Θ quantitatively describes the direct dependencies between attributes
- Assuming that the parent node set of attribute x_i in G is π_i, then Θ contains the CPT of each attribute: $\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$

A watermelon case



- What we can learn from the DAG
  - “色泽”直接依赖于“好瓜”和“甜度”
  - “根蒂”直接依赖于“甜度”
- What we can learn from the CPT
  - P(根蒂=硬挺|甜度=高)=0.1

Structure

- The Bayesian network structure effectively expresses the conditional independence between attributes
- Given a set of parent nodes, the Bayesian network assumes that each attribute is independent of its non-descendant attributes, then $B =< G, \Theta >$ defines the joint probability distribution of attributes $x_1, x_2, \dots, x_d$ as:

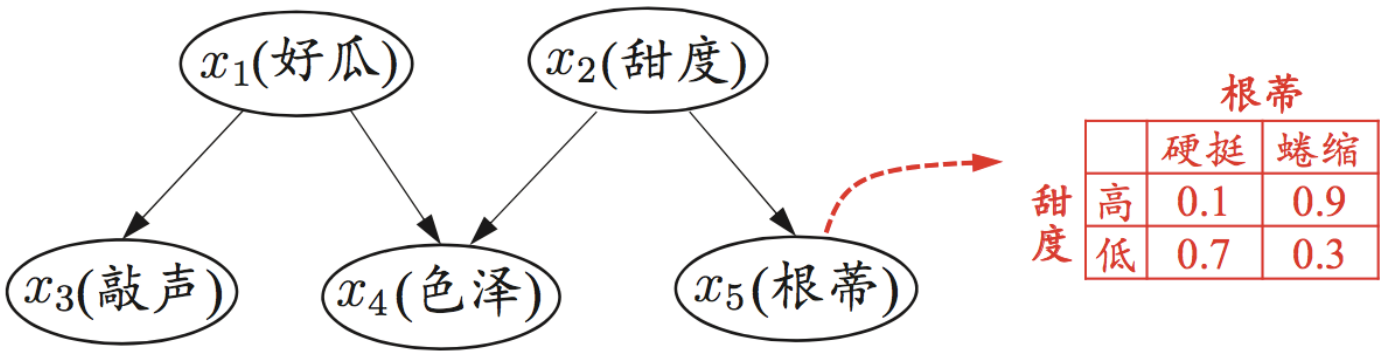
$$P_B(x_1, x_2, \dots, x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$

Structure

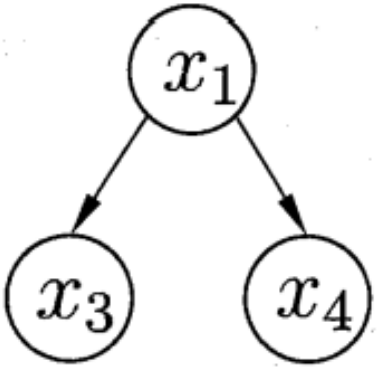
- Example: The joint probability distribution of the watermelon case is defined as

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

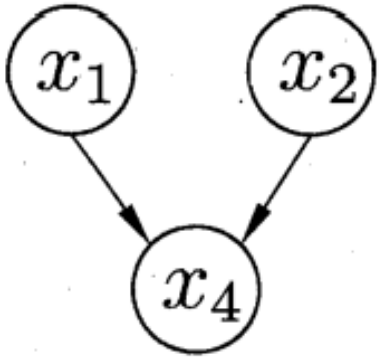
- x_3 and x_4 are independent when given x_1 => $x_3 \perp x_4 \,|x_1$
- x_4 and x_5 are independent when given x_2 => $x_4 \perp x_5 \,|x_2$



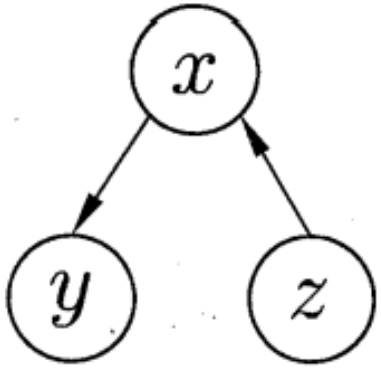
Typical dependencies



Common parent
同父结构
Given x_1, x_3 and x_4 are independent



V structure
V型结构
Given x_4, x_1 and x_2 are not independent.
But when x_4 is unknown, x_1 and x_2 are independent.



Sequential structure
顺序结构
Given x, y and z are independent

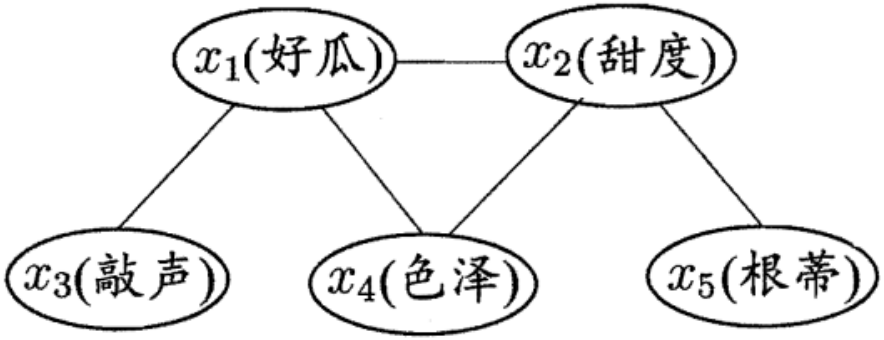
Marginal independence, $x_1 \perp\!\!\!\perp x_2$

Moral Graph

For fun

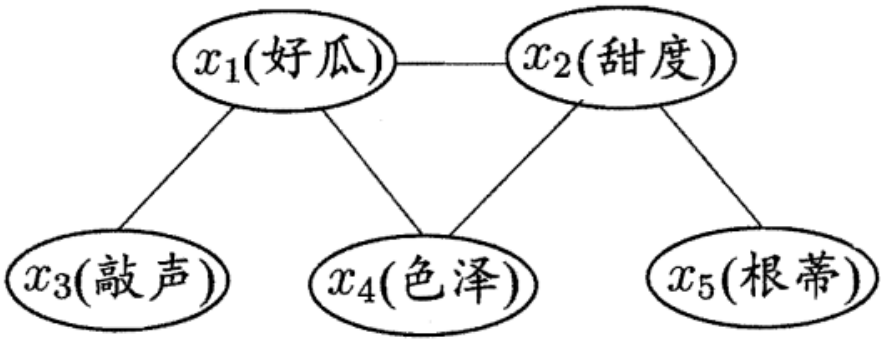
The meaning of "moralization":
The child's parents should
have a solid relationship,
otherwise it is immoral.

- We use D-separation(有向分离) to analyze conditional independence among variables in a directed graph
  - Find all V-structures in the directed graph and add a undirected edge between the two parent nodes of the V-structure
  - Change all directed edges to undirected edges
- The resulting undirected graph is called a **moral graph**, and the process of connecting parent nodes is called **moralization**
- Try to create a moral graph corresponding to the watermelon case!



Moral Graph

- Based on the moral graph, the conditional independence between variables can be found intuitively and quickly
- Assuming that there are variables x, y and variable set $\boldsymbol{z} = \{z_i\}$ in the moral graph. If x, y can be separated by $\boldsymbol{z}$ on the graph, then $x \perp y \mid \boldsymbol{z}$
- Try to find as many conditional independence relationships as possible in the moral graph of the watermelon case



Example:

$x_3 \perp x_4 \mid x_1$

$x_4 \perp x_5 \mid x_2$

$x_3 \perp x_2 \mid x_1$

$x_3 \perp x_5 \mid x_1$

$x_3 \perp x_5 \mid x_2$

Learning

- In practical applications, we often do not know the network structure.
- Therefore, the primary task of Bayesian network learning is to find the most "appropriate" Bayesian network structure based on the training data set.
- A common method is score searching(评分搜索): Define a score function and find the optimally structured Bayesian network based on this function
- Commonly used scoring functions are usually based on information theory, like ***Minimal Description Length (MDL).***

Inference

- Infer the values of other attribute variables through the observed values of some attribute variables
- Ideally, the posterior probability is accurately calculated directly from the joint probability distribution defined by the Bayesian network. But **it's NP hard**.
- We need "approximate inference" to obtain approximate solutions in limited time by reducing precision requirements.
- Common method: ***Gibbs sampling***

Summary

- **Bayesian decision theory**
  - Conditional Risk
  - Generative Models
- **Naïve Bayes Classifier**
  - Training steps
  - Laplacian correction
- **Bayesian Network**
  - Structure
  - Moral Graph

Thinking

- The attribute conditional independence assumption of Naive Bayes classifiers is difficult to hold in real-world applications. Does this mean that its classification performance is poor?