

# مفاهیم دسته بندی و شبکه های بیزین

## تعاریف

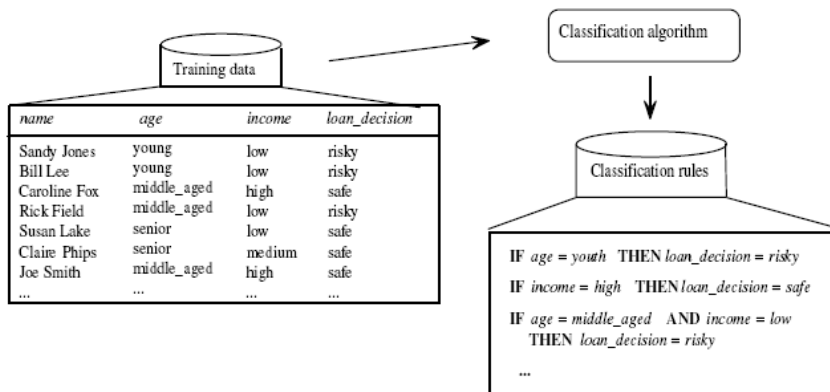
- دسته بندی یعنی تخصیص یک برچسب دسته به مجموعه های ازداده ها که هنوز دسته بندی نشده اند.
- داده ها دارای  $k$  ویژگی هستند که به صورت  $A_1, \dots, A_k$  نشان داده می شود. هر مورد یا مثال بوسیله مقادیر ویژگیها و یک برچسب دسته توصیف می شود.
- اغلب یکی از ویژگی های داده دسته آن را نشان می دهد.

## • تفاوت دسته بندی و خوشه بندی

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## فرایند دو مرحله ای در دسته بندی

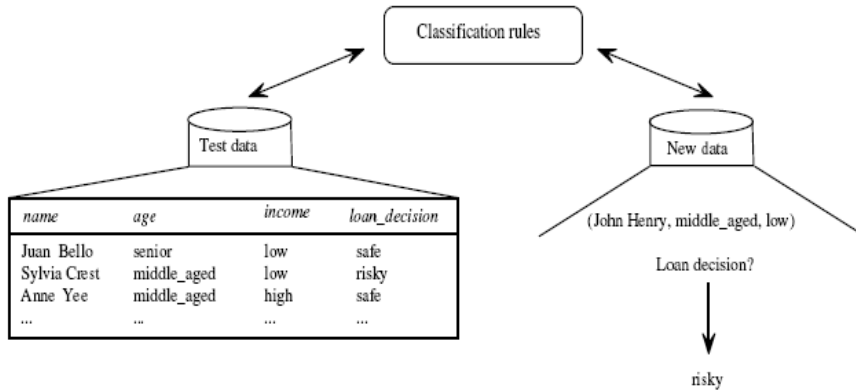
• ساخت مدل



سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

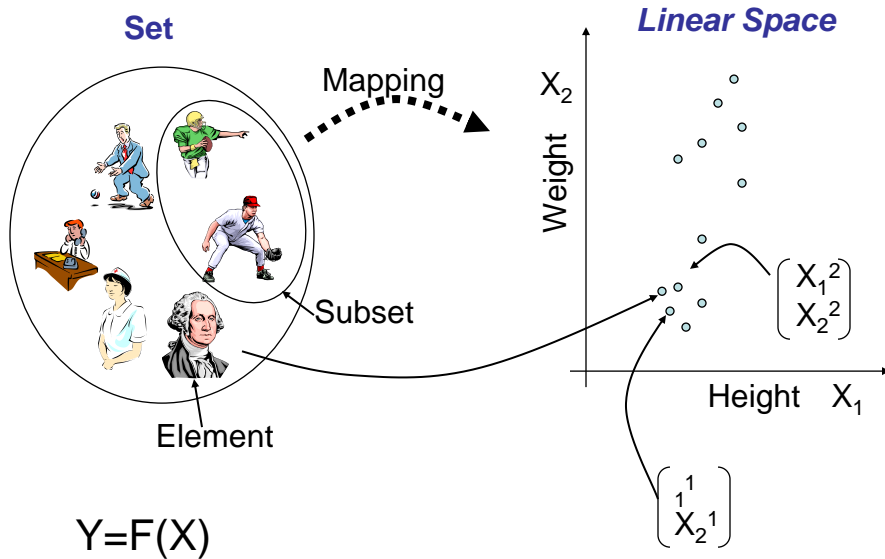
# فرایند دو مرحله ای در دسته بندی

• استفاده از مدل



سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## Mapping



سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## روشهای ارزیابی دسته بندی ها

- دقت پیش بینی (Accuracy)
- زمان لازم برای ساخت مدل و زمان لازم برای استفاده از مدل
- پایداری (Robustness):  
توانایی برخورد با داده های غیر معمول (noise) و یا مقادیر جا افتاده (missing values)
- قابلیت تفسیر (Interpretability):  
قابل فهم بودن مدل و فراهم سازی بینش لازم توسط آن
- جمع و جور بودن (Compactness) مدل:  
اندازه درخت یا تعداد قواعد و ...

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## شاخص دقت در دسته بندی

Classes	<i>buys_computer = yes</i>	<i>buys_computer = no</i>	Total	Recognition (%)
<i>buys_computer = yes</i>	6,954	46	7,000	99.34
<i>buys_computer = no</i>	412	2,588	3,000	86.27
Total	7,366	2,634	10,000	95.52

	$C_1$	$C_2$
$C_1$	true positives	false negatives
$C_2$	false positives	true negatives

$$\text{sensitivity} = \frac{t_{pos}}{pos}$$

$$\text{specificity} = \frac{t_{neg}}{neg}$$

$$\text{precision} = \frac{t_{pos}}{(t_{pos} + f_{pos})}$$

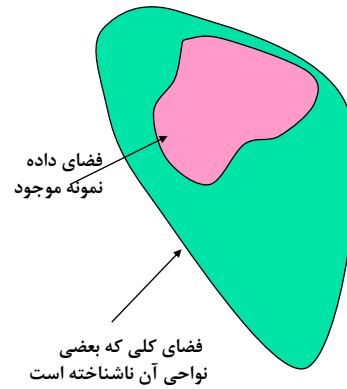
$$\text{accuracy} = \text{sensitivity} \frac{pos}{(pos + neg)} + \text{specificity} \frac{neg}{(pos + neg)}$$

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## مساله تعميم

### مساله تعميم

- کلاس بندي الگو فقط مي تواند براي نمونه هاي موجود مدل کلاس بندي را مشخص نمايد. اما در هر حال مي دانيم که معمولا همه داده ها در دسترس و موجود نيستند و مشکل از اينجا ناشي مي شود



سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## Simple Bayesian Formula

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

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

## Naïve Bayesian Classifier

Class:

C1:buys\_computer = 'yes'

C2:buys\_computer = 'no'

Data sample

X = (age <=30,

Income = medium,

Student = yes

Credit\_rating = Fair)

age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

- یادگیری احتمالی می تواند معادل محاسبه  $P(C=c | d)$  باشد, برای مثال برای یک داده نمونه  $d$  احتمال اینکه در کلاس  $C$  قرار گیرد چیست ( $C$  کلاس ویژگیهاست) چگونه محاسبات لازم انجام می شود ؟

فرض کنید  $A_1$  تا  $A_k$  ویژگیهایی با مقادیر گسسته باشند. این مقادیر برای پیش بینی یک کلاس گسسته  $C$  بکار می روند. فرض کنید یک مثال داریم با مقادیر ویژگی مشاهده شده  $a_1$  تا  $a_k$  کلاس بندی یا پیش بینی, انتخاب کلاس  $C$  است به نحوی که  $P(C=c | A_1=a_1 \cup \dots \cup A_k=a_k)$  ماکزیمم شود .

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

- $P(C=c|A_1=a_1 \cup \dots \cup A_k=a_k) =$
- $P(A_1=a_1 \cup \dots \cup A_k=a_k | C=c) P(C=c) / P(A_1=a_1 \cup \dots \cup A_k=a_k)$

•  $P(C=c)$  به سادگی از داده های تربیت مدل قابل استخراج است.

•  $P(A_1=a_1 \cup \dots \cup A_k=a_k)$  برای تصمیم گیری بی تاثیر است زیرا که برای همه مقادیر  $C$  یکسان است.

• پس لازم است که فقط مقدار زیر محاسبه شود

- $P(A_1=a_1 \cup \dots \cup A_k=a_k | C=c)$

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

- $P(A_1=a_1 \cup \dots \cup A_k=a_k | C=c)$

• فرضی که در **Bayes** ساده وجود دارد این است که ویژگیها به صورت شرطی از هم مستقل هستند.

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

$$= P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i).$$

- $P(A_1=a_1 \cup \dots \cup A_k=a_k | C=c) = P(A_1=a_1 | C=c) * \dots * P(A_k=a_k | C=c)$

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

- حال برای پیش بینی کلاس داده جدید اینگونه عمل میکنیم:

$$X = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})$$

$$P(\text{buys\_computer} = \text{yes}) = 9/14 = 0.643$$

$$P(\text{buys\_computer} = \text{no}) = 5/14 = 0.357$$

$$P(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{yes}) = 2/9 = 0.222$$

$$P(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{no}) = 3/5 = 0.600$$

$$P(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{yes}) = 4/9 = 0.444$$

$$P(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{no}) = 2/5 = 0.400$$

$$P(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{no}) = 1/5 = 0.200$$

$$P(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{no}) = 2/5 = 0.400$$

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی

$$\begin{aligned} P(X \mid \text{buys\_computer} = \text{yes}) &= P(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{yes}) \times \\ &P(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{yes}) \times \\ &P(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{yes}) \times \\ &P(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{yes}) \\ &= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044. \end{aligned}$$

همین طور برای کلاس دیگر داریم:

$$P(X \mid \text{buys\_computer} = \text{no}) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

$$P(X \mid C_i)P(C_i),$$

برای پیدا کردن کلاس داده جدید میبایستی عبارت روبرو ماکزیمم شود

بنابراین داریم:

$$P(X \mid \text{buys\_computer} = \text{yes})P(\text{buys\_computer} = \text{yes}) = 0.044 \times 0.643 = 0.028$$

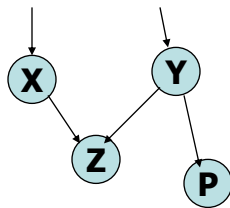
$$P(X \mid \text{buys\_computer} = \text{no})P(\text{buys\_computer} = \text{no}) = 0.019 \times 0.357 = 0.007$$

سمیه علیزاده هیات علمی دانشکده صنایع  
دانشگاه خواجه نصیر طوسی



## Bayesian Belief Networks

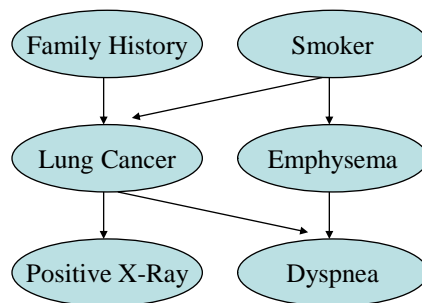
- Bayesian belief network allows a *subset* of the variables conditionally independent
- A graphical model of causal relationships
  - Represents dependency among the variables
  - Gives a specification of joint probability distribution



- Nodes: random variables
- Links: dependency
- X and Y are the parents of Z, and Y is the parent of P
- No dependency between Z and P
- Has no loops or cycles

## Example: Bayesian Network

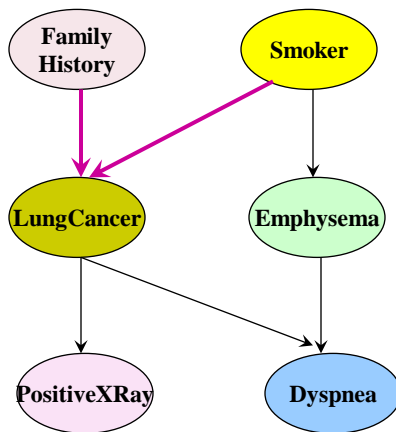
*Lung Cancer depends on Family History and Smoker*



	FH,S	FH,-S	-FH,S	-FH,-S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

*Lung Cancer is conditionally independent of emphysema. given Family History and Smoker*

## Bayesian Belief Network



The **conditional probability table (CPT)** for variable LungCancer:

	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

CPT shows the conditional probability for each possible combination of its parents

Derivation of the probability of a particular combination of values of  $\mathbf{X}$ , from CPT:

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{Parents}(Y_i))$$

## Training Bayesian Networks

- Several scenarios:
  - Given both the network structure and all variables observable: *learn only the CPTs*
  - Network structure known, some hidden variables: *gradient descent* (greedy hill-climbing) method, analogous to neural network learning
  - Network structure unknown, all variables observable: search through the model space to *reconstruct network topology*
  - Unknown structure, all hidden variables: No good algorithms known for this purpose
- Ref. D. Heckerman: Bayesian networks for data mining