# داده کاوی

## فصل ۹

تهیه شده توسط
محمد پورزعفرانی
Pourzaferani.ir
دانشگاه اصفهان

دانشگاه اصفهان

# مترجمان

- خانم بخشیانی
- خانم سلطانی
- خانم دهقانی
- خانم تیماس
- خانم طرقی
- خانم امین الرعایا
- خانم مولایی
- خانم شمسی زاده

- ویرایش : جزینی

# فصل نهم: روش‌های دسته‌بندی پیشرفته

- شبکه های باور بیزین Bayesian Belief **Network**
- دسته بندی به کمک شبکه‌های عصبی
- ماشین های بردار پشتیبان(SVM)
- دسته بندی با استفاده از الگوهای پرتکرار
- یادگیری تنبل (**Lazy learner**)
- خلاصه

## شبکه های باور بیزی

- **شبکه های باور بیزی**( سایر اسامی **شبکه های بیزین** یا **شبکه های احتمالی**)
- تعریف : یک مدل (غیرمدور و جهتدار) گرافیکی از روابط علت و معلولی
  - نشان دهنده وابستگی بین متغیرها
  - نشان دهنده توزیع احتمال ترکیبی

✓ نحوه ترسیم گراف

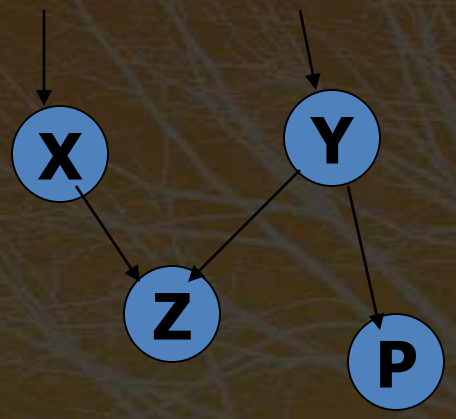
✓ گره‌ها : متغیرهای تصادفی

✓ لینکها : وابستگی

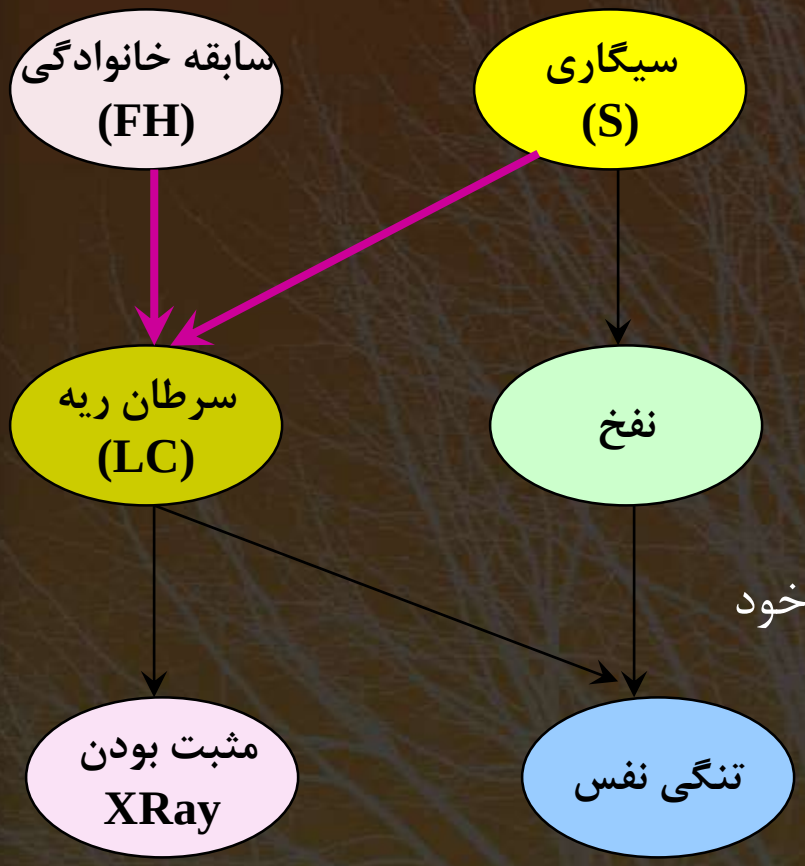
✓ X و Y والدین Z هستند ، و Y والد P می‌باشد.

✓ هیچ وابستگی بین Z و P وجود ندارد.

✓ بدون حلقه / چرخه



# شبکه های باور بیزی : مثال




**Bayesian Belief Network**

**CPT (Conditional Probability Table)**

جدول احتمال شرطی برای متغیر سرطان ریه

| | (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 |

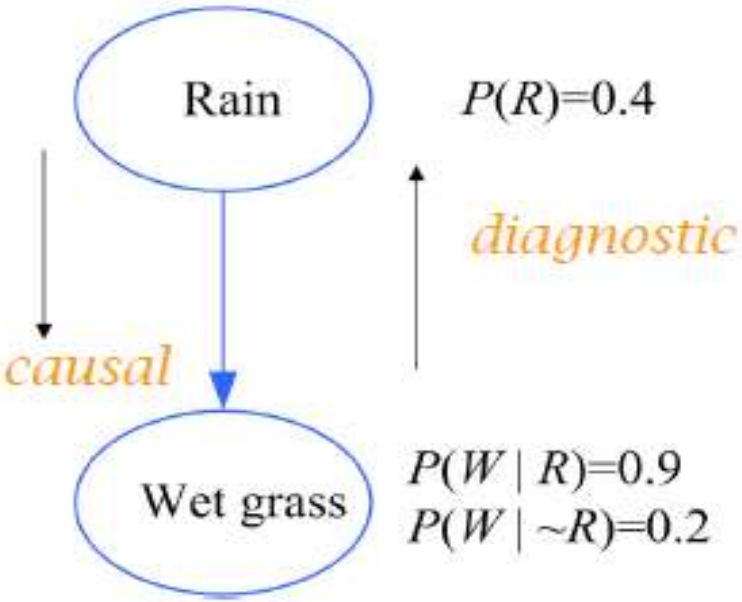
نشان دهنده احتمال شرطی برای هر ترکیب ممکن از والدین خود

نحوه بدست آوردن احتمال **X** از CPT :

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(Y_i))$$

# درنظر گرفتن هر دو ساختار شبکه و تمام متغیرهای قابل مشاهده:

## Causes and Bayes' Rule




**استنتاج تشخیصی :**

**با علم به اینکه چمن خیس است ، چقدر احتمال این وجود دارد که علت مرطوب بودن آن باران است؟**

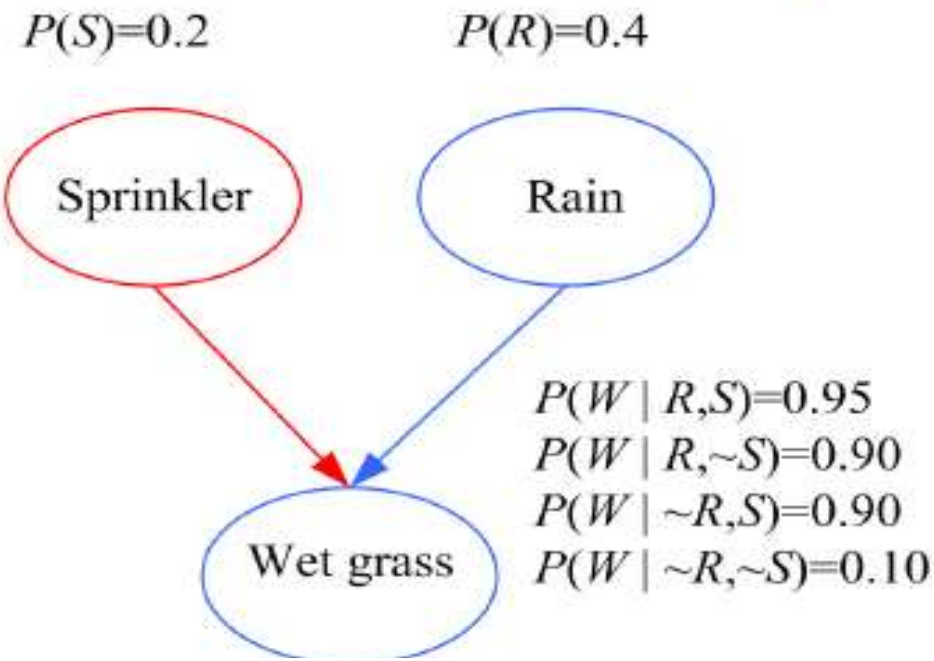
$$P(R \mid W) = \frac{P(W \mid R)P(R)}{P(W)}$$

$$= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)}$$

$$= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75$$

## درنظر گرفتن هر دو ساختار شبکه و تمام متغیرهای قابل مشاهده:

# Causal vs Diagnostic Inference




استنتاج علی :

اگر آب پاش روشن باشد ،چقدر احتمال این وجود دارد که چمن مرطوب است؟ ?

$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|\sim R,S)\,P(\sim R|S)$$
$$= P(W|R,S)\,P(R) + P(W|\sim R,S)\,P(\sim R)$$
$$= 0.95\ 0.4 + 0.9\ 0.6 = 0.92$$

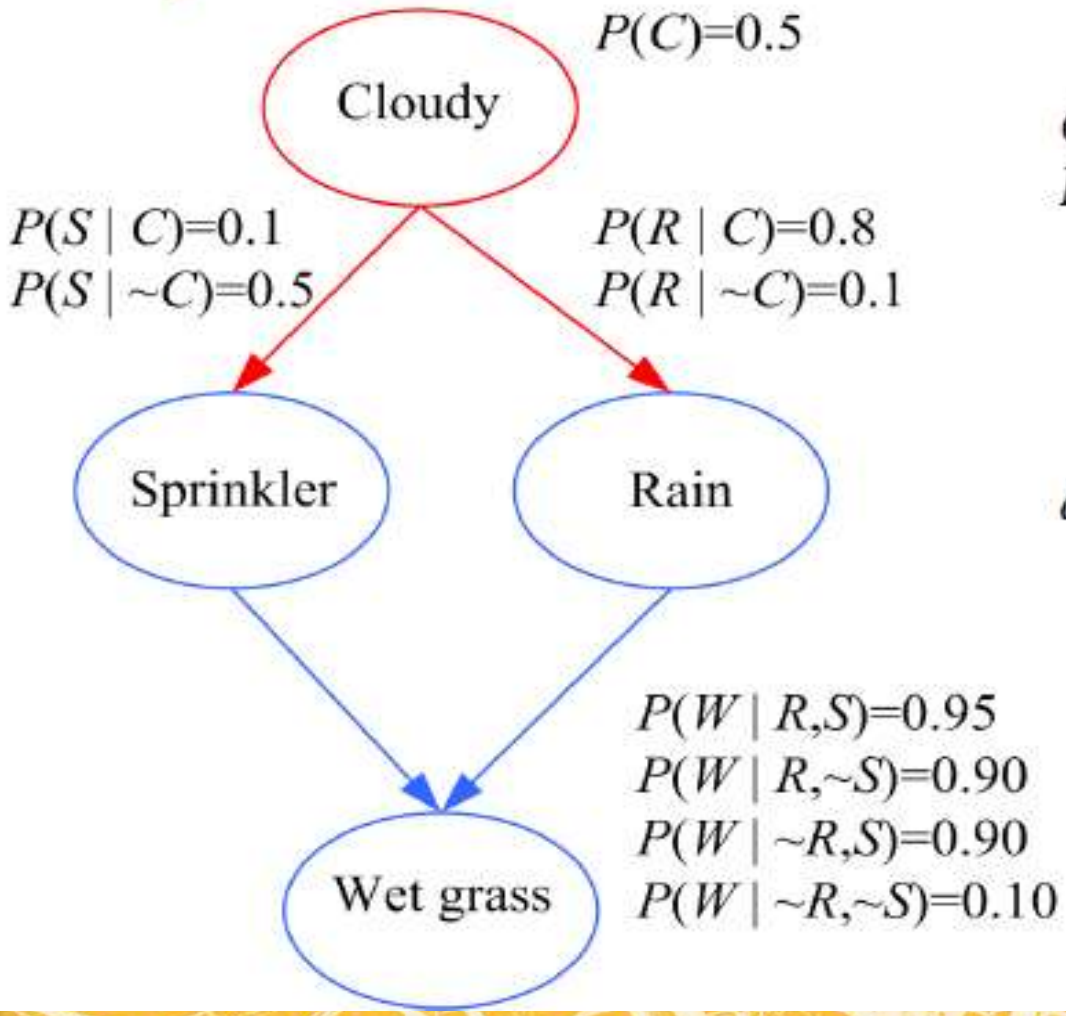
**استنتاج تشخیصی :** اگر این چمن مرطوب است ، احتمال اینکه آب پاش روشن باشد چقدر است ؟

$P(S|W) = 0.35 > 0.2\ P(S)$

**توضیح دور :** دانستن اینکه باران باریده است احتمال اینکه آب پاش روشن باشد را $P(S|R,W) = 0.21$ کاهش میدهد.

## درنظر گرفتن هر دو ساختار شبکه و تمام متغیرهای قابل مشاهده:

# Bayesian Networks: Causes




*Causal inference:*

$$P(W|C) = P(W|R,S)\,P(R,S|C) + P(W|\sim R,S)\,P(\sim R,S|C) + P(W|R,\sim S)\,P(R,\sim S|C) + P(W|\sim R,\sim S)\,P(\sim R,\sim S|C)$$

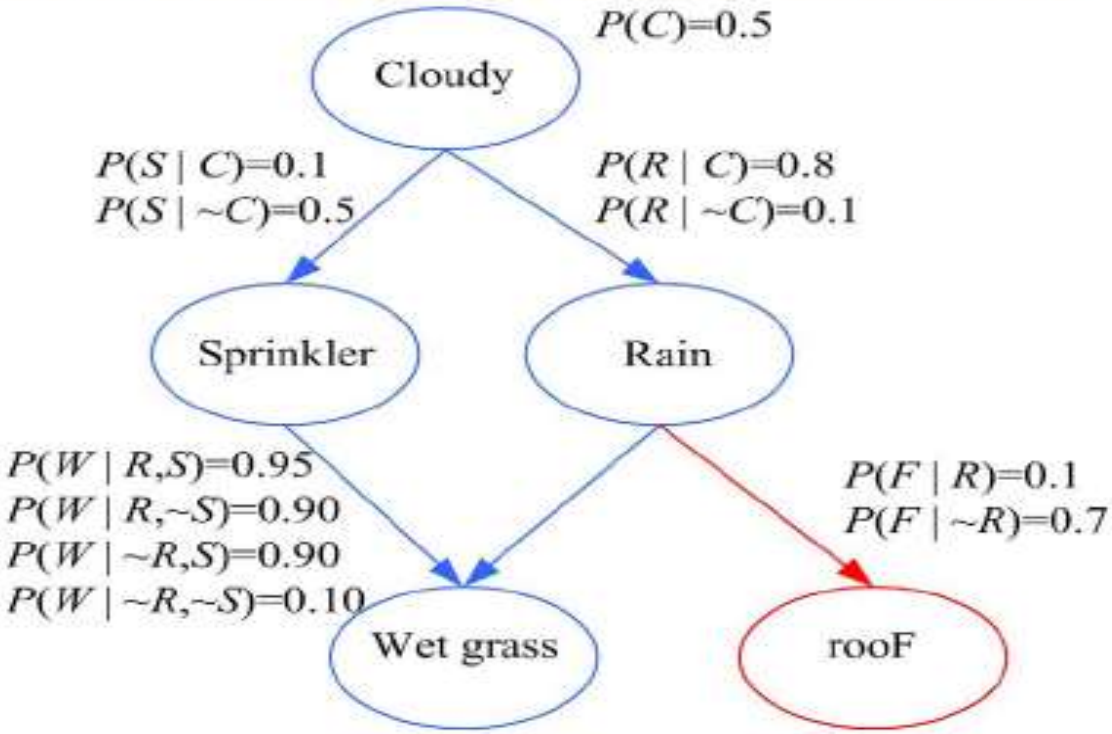
*and use the fact that*

$$P(R,S|C) = P(R|C)\,P(S|C)$$

*Diagnostic:* $P(C|W) = ?$

## درنظر گرفتن هر دو ساختار شبکه و تمام متغیرهای قابل مشاهده:

# Bayesian Nets: Local structure




$$P(F \mid C) = ?$$

$$P(C,S,R,W,F) = P(C)P(S \mid C)P(R \mid C)P(W \mid S,R)P(F \mid R)$$

$$P(X_1,\ldots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents}(X_i))$$

# Naïve Bayes Classifier

- روش :
  - محاسبه احتمال خلف $P(C \mid A_1, A_2, \ldots, A_n)$ برای تمام مقادیر C با استفاده از قضیه بیز.

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - C به گونه ای انتخاب شود که ماکزیمم شود

$$P(C \mid A_1, A_2, \ldots, A_n)$$

  - معادل است با انتخاب ماکزیمم C

$$P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$$

- چگونه می‌توان $P(A_1, A_2, \ldots, A_n \mid C)$ را تخمین زد؟
- فرض کنید هنگامی که کلاس داده شده ویژگی های $A_i$ مستقل هستند :
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - برای تمامی $A_i$ و $C_j$ میتوان $P(A_i \mid C_j)$ را تخمین زد
  - نقطه جدید در $C_j$ دسته بندی میشود اگر $P(C_j) \prod P(A_i \mid C_j)$ ماکزیمم باشد.

با توجه به دو ساختار شبکه و تمام متغیرهای مشاهده شده

# Bayesian Networks: Inference

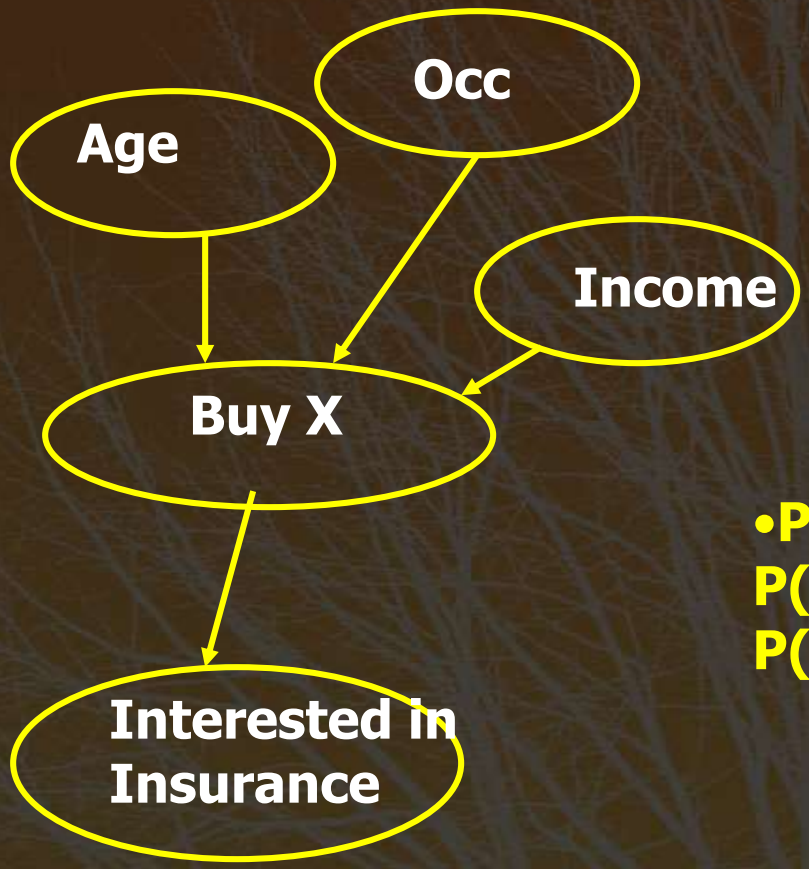
$$P(C,S,R,W,F) = P(C)\ P(S|C)\ P(R|C)\ P(W|R,S)\ P(F|R)$$

$$P(C,F) = \Sigma_S\ \Sigma_R\ \Sigma_W\ P(C,S,R,W,F)$$

$$P(F|C) = P(C,F)\ /\ P(C)$$ *Not efficient!*

Belief propagation (Pearl, 1988)
Junction trees (Lauritzen and Spiegelhalter, 1988)

# شبکه‌های باور بیزی: مثال




- سن، شغل و درآمد و تعیین اینکه آیا مشتری این محصول را خریداری می کند.
- با توجه به اینکه مشتری محصولات را خریداری می کند، آیا علاقه به بیمه ای که مستقل از سن، شغل و درآمد است می باشد.

- **P(Age, Occ, Inc, Buy, Ins ) = P(Age)P(Occ)P(Inc) P(Buy|Age,Occ,Inc)P(Int|Buy)**

# فصل نهم:دسته بندی : روش های پیشرفته

- شبکه های باور بیزیBayesian Belief
- دسته بندی بر اساس پس انتشار(Backpropagation)
- ماشین های بردار پشتیبان(SVM)
- دسته بندی با استفاده از الگوهای پرتکرار
- کندآموزان ( یا یادگیری از همسایگان خود)
- سایر روشهای دسته بندی
- مباحث اضافی در مورد دسته بندی
- خلاصه

# فصل نهم:دسته بندی : روش های پیشرفته

- شبکه های باور بیزیBayesian Belief
- دسته بندی بر اساس پس انتشار(Backpropagation)
- ماشین های بردار پشتیبان(SVM)
- دسته بندی با استفاده از الگوهای پرتکرار
- کندآموزان ( یا یادگیری از همسایگان خود)
- سایر روشهای دسته بندی
- مباحث اضافی در مورد دسته بندی
- خلاصه

# دسته‌بندی با استفاده از الگوهای پرتکرار

For example, the following is an association rule mined from a data set, $D$, shown with its confidence and support:

$$age = youth \wedge credit = OK \Rightarrow buys_computer = yes \; [support = 20\%, confidence = 93\%], \qquad (9.21)$$

where $\wedge$ represents a logical "AND." We will say more about confidence and support later.

More formally, let $D$ be a data set of tuples. Each tuple in $D$ is described by $n$ attributes, $A_1, A_2, \ldots, A_n$, and a class label attribute, $A_{class}$. All continuous attributes are discretized and treated as categorical (or nominal) attributes. An **item**, $p$, is an attribute–value pair of the form $(A_i, v)$, where $A_i$ is an attribute taking a value, $v$. A data tuple $X = (x_1, x_2, \ldots, x_n)$ satisfies an item, $p = (A_i, v)$, if and only if $x_i = v$, where $x_i$ is the value of the $i$th attribute of $X$. Association rules can have any number of items in the rule antecedent (left side) and any number of items in the rule consequent (right side). However, when mining association rules for use in classification, we are only interested in association rules of the form $p_1 \wedge p_2 \wedge \ldots p_l \Rightarrow A_{class} = C$, where the rule antecedent is a conjunction of items, $p_1, p_2, \ldots, p_l$ $(l \leq n)$, associated with a class label, $C$. For a given rule, $R$, the percentage of tuples in $D$ satisfying the rule antecedent that also have the class label $C$ is called the **confidence** of $R$.

# دسته بندی با استفاده از الگوهای پر تکرار

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| middle | high | no | fair | yes |
| senior | medium | no | fair | yes |
| senior | low | yes | fair | yes |
| senior | low | yes | excellent | no |
| middle | low | yes | excellent | yes |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| senior | medium | yes | fair | yes |
| youth | medium | yes | excellent | yes |
| middle | medium | no | excellent | yes |
| middle | high | yes | fair | yes |
| senior | medium | no | excellent | no |

p1=(age,youth),P2=(age,middle),P3=(age,senior)

P4=(income,low),P5=(income,medium),

P6=(income,high)

P7=(credit_rating,fair),P8=(credit_rating,excellent)

P9=(student,no), P10=(student,yes)

I1=(buys_computer,no),I2=(buys_computer,yes)

| | | | | |
|---|---|---|---|---|
| P1 | P6 | P9 | P7 | I1 |
| P1 | P6 | P9 | P8 | I1 |
| P2 | P6 | P9 | P7 | I2 |
| P3 | P5 | P9 | P7 | I2 |
| P3 | P4 | P10 | P7 | I2 |
| P3 | P4 | P10 | P8 | I1 |
| P2 | P4 | P10 | P8 | I2 |
| P1 | P5 | P9 | P7 | I1 |
| P1 | P4 | P10 | P7 | I2 |
| P3 | P5 | P10 | P7 | I2 |
| P1 | P5 | P10 | P8 | I2 |
| P2 | P5 | P9 | P8 | I2 |
| P2 | P6 | P10 | P7 | I2 |
| P3 | P5 | P9 | P8 | I1 |

# دسته بندی با استفاده از الگوهای پرتکرار

قدم‌های کلی این روش به صورت زیر هستند

۱. تجزیه و تحلیل مجموعه اقلام پرتکرار به منظور تولید قوانین انجمنی در هر کلاس ، تا بتواند معیار confidence و support را برآورده سازد.

۲. قوانین انجمنی قوی که از روی الگوهای پرتکرار ساخته می‌شوند به صورت زیر هستند

۳. در مقدم قوانین، تعدادی جفت ویژگی-مقدار با یکدیگر ترکیب شده و در نتیجه قانون نتیجه کلاس را داریم

$$P_1 \wedge p_2 \ldots \wedge p_l \rightarrow \text{“}A_{class} = C\text{”} \ (conf, sup)$$

دلیل مناسب بودن این روش

- این روش قوانین را با در نظر گرفتن چند ویژگی در یک زمان لحاظ میکند در حالی که در درخت تصمیم در هر بار یک ویژگی انتخاب می‌شود

# فصل نهم:دسته بندی : روش های پیشرفته

- شبکه های باور بیزیBayesian Belief
- دسته بندی بر اساس پس انتشار(Backpropagation)
- ماشین های بردار پشتیبان(SVM)
- دسته بندی با استفاده از الگوهای پرتکرار
- کندآموزان ( یا یادگیری از همسایگان خود)
- سایر روشهای دسته بندی
- مباحث اضافی در مورد دسته بندی
- خلاصه

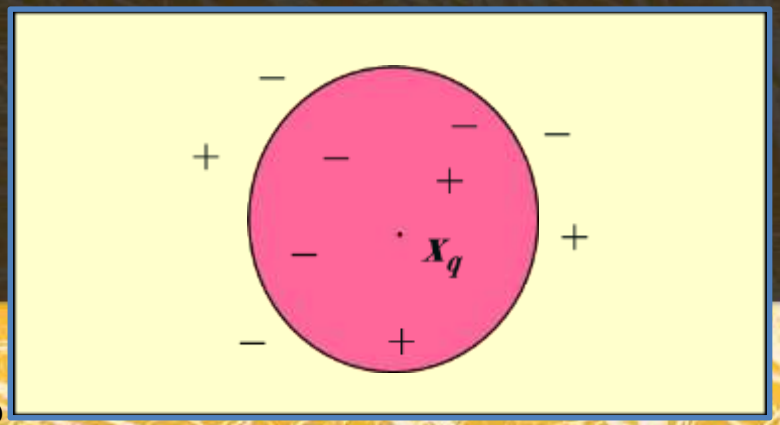
# یادگیری تنبل در مقابل مشتاق

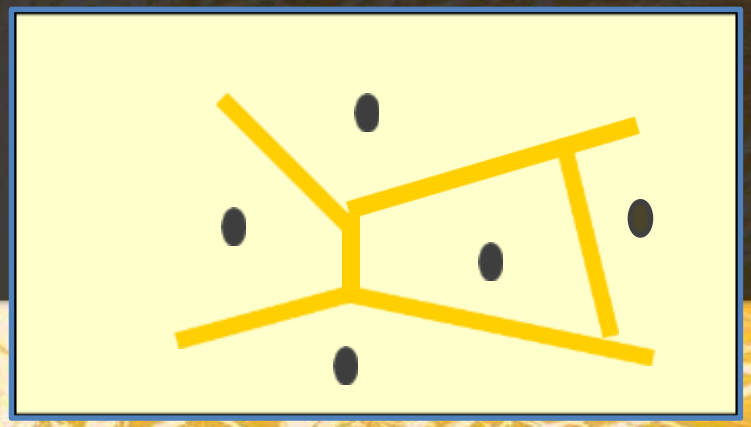
- یادگیری تنبل در مقابل مشتاق
  - یادگیری تنبل : به طور ساده داده‌های آموزشی را ذخیره کرده و تا زمانی که یک سوال نشده صبر میکند
  - یادگیری مشتاق : اگر مجموعه‌ای از داده‌های آموزشی داده شد، بر اساس آن‌ها یک مدل دسته‌بندی ایجاد نموده و پس از آن داده‌های تست را با مدل ساخته شده پاسخ میدهد

- **یادگیری تنبل**: زمان کمتر در آموزش اما زمان بیشتری در پیش بینی صرف میکند
- **نمونه این روش‌ها : روش k-نزدیکترین همسایه**

# الگوریتم k – نزدیک‌ترین همسایه

- همه موارد به نقاط در فضای متناظر N-D اشاره دارد.
- تعریف اقلیدسی از نظرفاصله نزدیک ترین همسایه،منطقه(X1,X2)
- تابع هدف می تواند گسسته یا ارزش واقعی باشد.
- برای ارزش گسسته، بازده K-NN شایع ترین و نزدیک ترین مقدار نمونه آزمایش Xq
- نمودار vonoroi: ناشی از سطح تصمیم 1-nn برای نوعی مجموعه نمونه آموزش

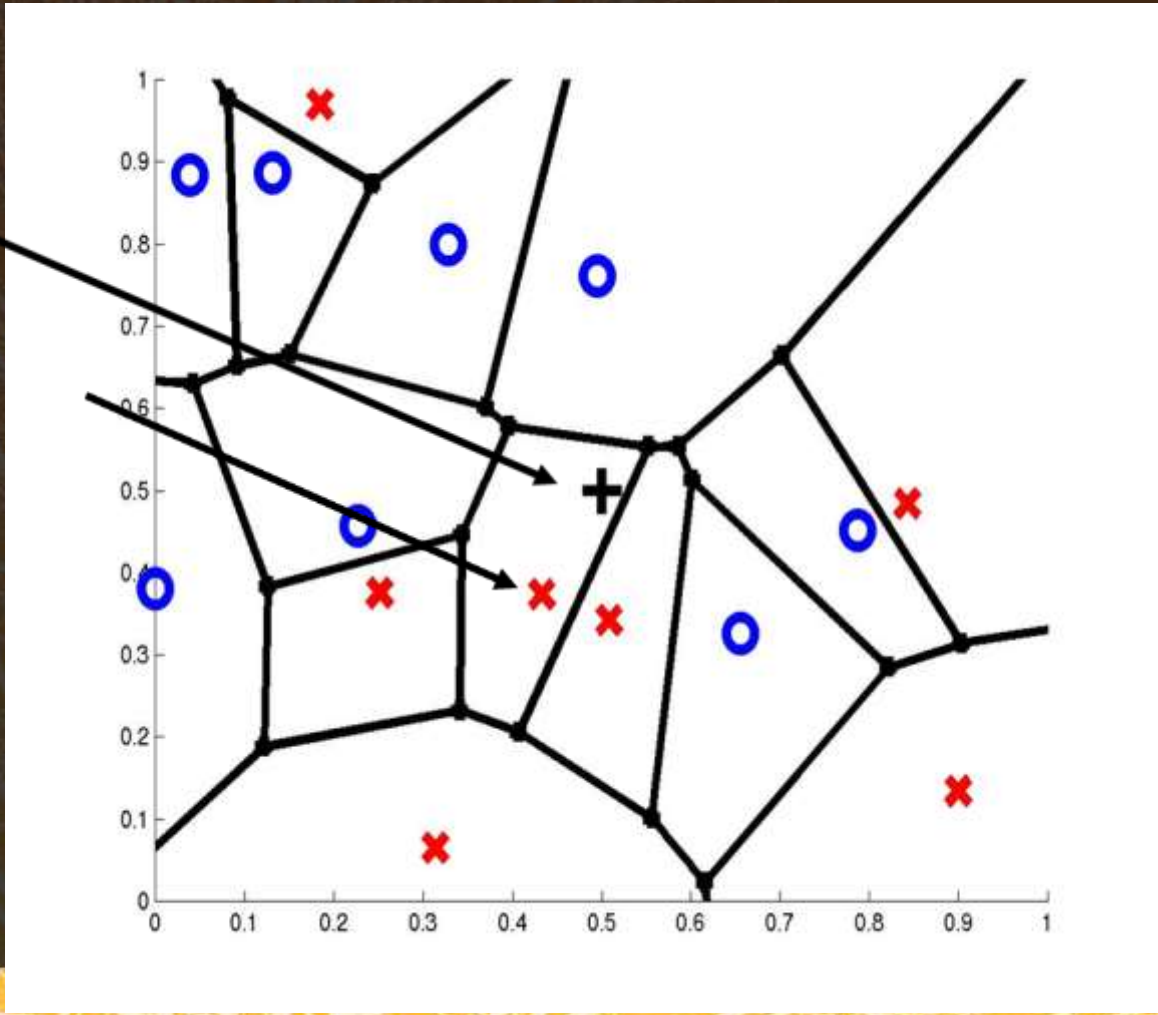


# ۱-نزدیک ترین همسایه

qf نقطه پرس و جو

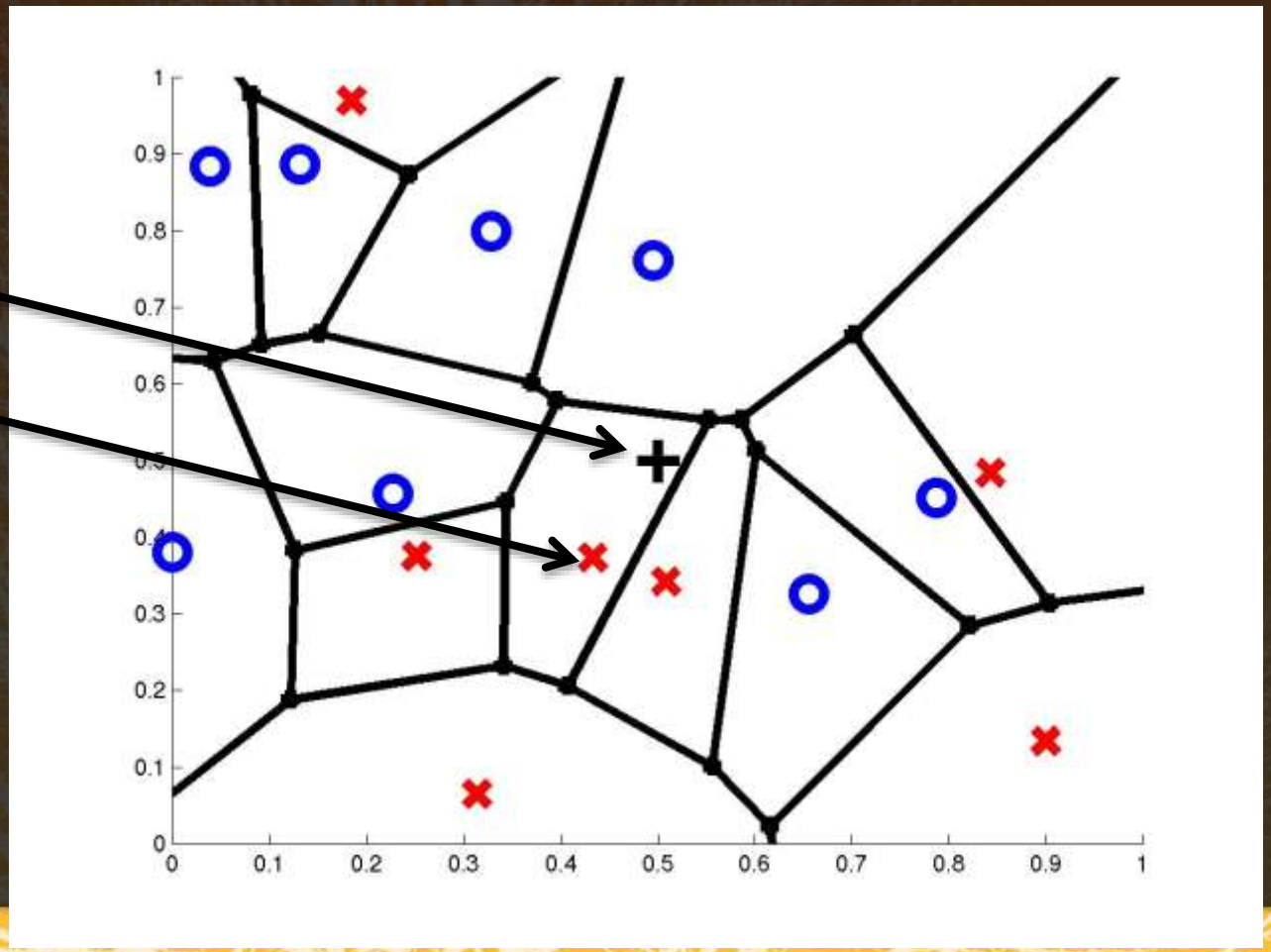
qi 1- نزدیکترین همسایه



# ۱-نزدیکترین همسایگان

$q_f$ نقطه پرس و جو

$q_i$ نزدیکترین همسایه

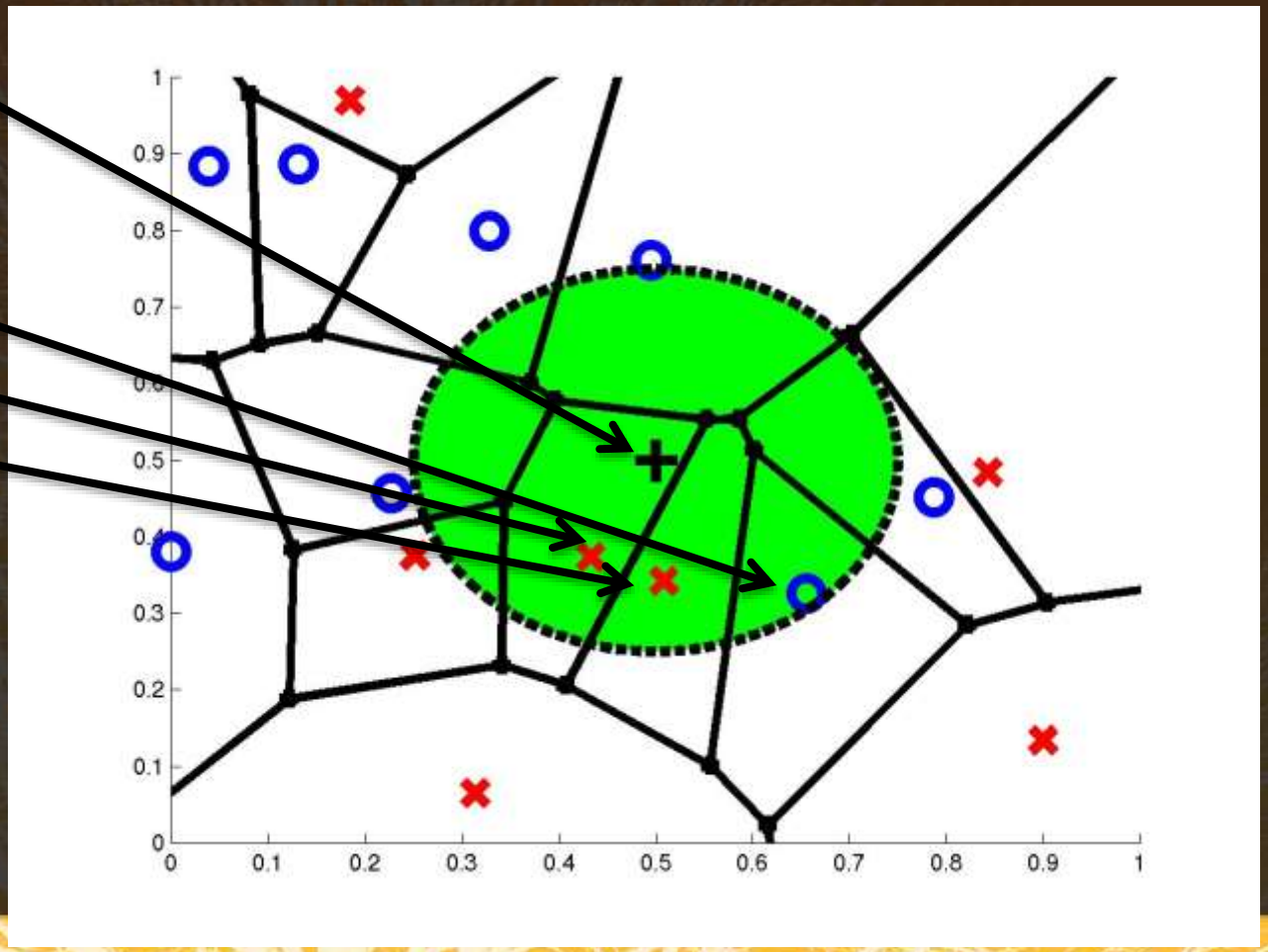


# ۳-نزدیکترین همسایه

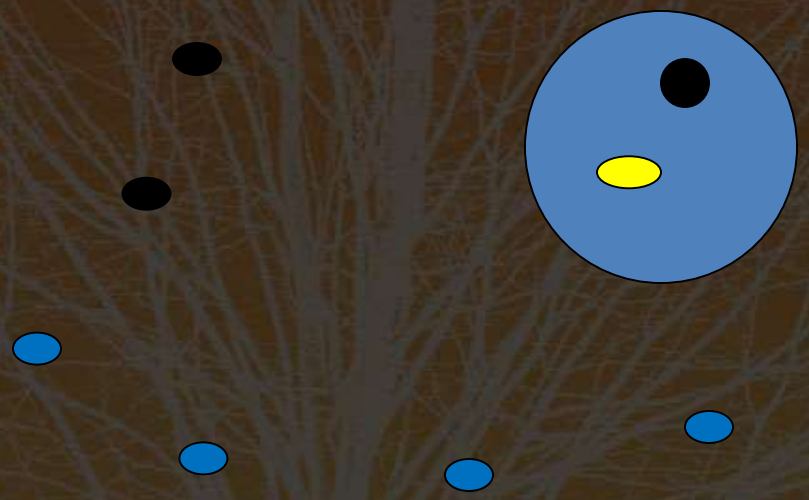
$q_f$ نقطه پرس و جو

$q_i$ ۳-نزدیکترین همسایه

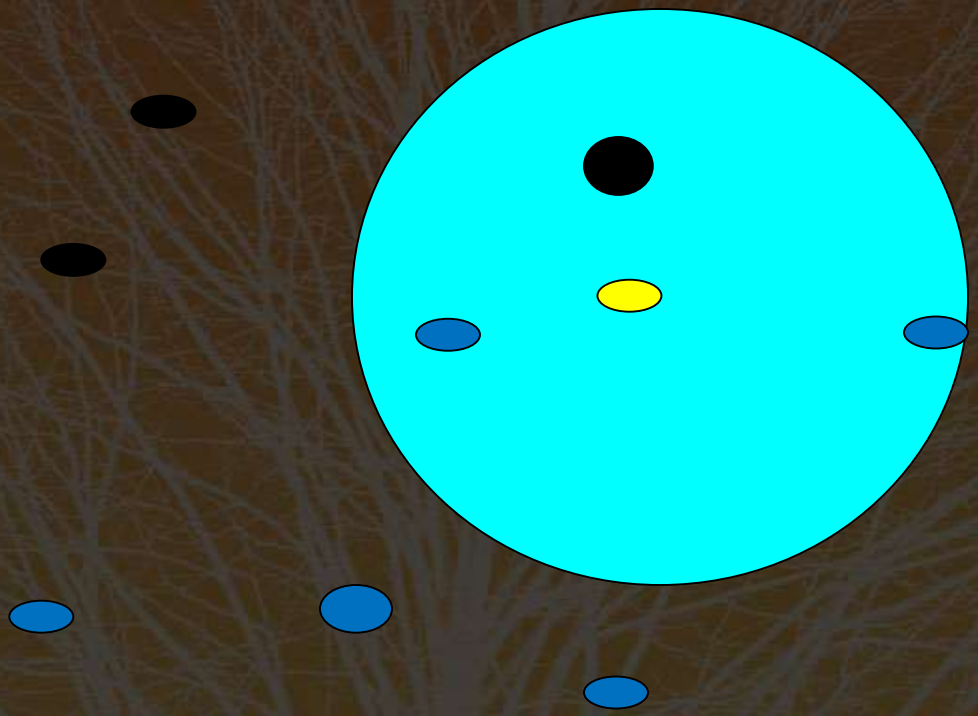
2x,1o



# ۱-نزدیکترین همسایه



# ۳-نزدیکترین همسایه



# یادگیری مبتنی بر نمونه

ایده کلیدی:فقط نمونه‌های آموزش ذخیره می‌شوند

نزدیکترین همسایه: با توجه به <xi,f(xi)>

Xq به عنوان مثال پرس و جو اولین کردهای نزدیک‌ترین آموزش

مثال fxq=fx1

Kـنزدیکترین همسایه:xq را در میان k نزدیکترین همسایگان خود(در صورت گسسته بودن ارزش تابع)

نگاهی به معنی ارزش fاز k نزدیکترین همسایه(در صورت حقیقی)

$$f(x_q)=\Sigma_{i=1}^{k} f(x_i)/k$$

# K_ نزدکترین همسایه آموزش (K-NN)

- اساسی ترین روش مبتنی بر مثال
- نمونه های نقشه برای نقاط در $R^n$

به عنوان مثال , X دلخواه

$$\langle a_1(x), a_2(x), \ldots a_n(x) \rangle$$

نزدیکترین همسایه :فاصله اقلیدسی

فاصله بین $X_i$ و $X_j$

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

# بحث بر روی الگوریتم *k*-NN

- **K-NN پیش بینی ارزش واقعی برای یک تاپل با مقدار ناشناخته**

  بازگرداندن مقادیر میانگین از K نزدیکترین همسایه
- **الگوریتم نزدیکترین همسایه با فاصله وزن‌دار**

  وزن هر یک از همسایگان K با توجه به فاصله آنها به پرس و جو $x_q$

  دادن وزن بیشتر به همسایگان نزدیک
- **مقاوم در برابر داده نویزی با استفاده از میانگین K-نزدیکترین همسایه**

# K?

- **K بزرگ**

حساسیت کمتری نسبت به نویز

برآورد بهتری از احتمال

- **K کوچک**

ساختار مناسب از فضای مشکل ارائه می‌دهد.

هزینه کمتر

- **تعادل باید بین بزرگ و کوچک K باشد.**

# نظایر بر روی الگوریتم k-NEAREST NEIGHBOR

- الگوریتم KNN با فاصله (یال‌های) وزن‌دار، روش استنتاج بسیار موثر برای بسیاری از مشکلات عملی است.
- این روش برای مجموعه‌های به اندازه کافی بزرگ از داده‌های آموزشی موثر است.
- این الگوریتم در مقابل داده آموزشی پر نویز مقاوم است.