
بسم الله الرحمن الرحیم

داده کاوی

دکتر حسام عمرانپور

دانشگاه صنعتی نوشیروانی بابل

مراجع: Data mining concepts and techniques,Jiawei Han,Michelin Cumber and Jian Pei

داده کاوی مهدی اسماعیلی

تنظیم : فاطمه حیدری

حدیثه پورعلی

فصل 3

دسته بندی

دسته بندی (classification)

ورودی: داده به همراه برچسب

#	طول	ارتفاع	نوع
1	7	4	اتوبوس
2	6,5	4,5	اتوبوس
3	7,5	4,5	اتوبوس
4	9	4,5	اتوبوس
5	3	1,5	پراید
6	2,5	1,7	پراید
7	2	1,6	پراید

	خانه دارد	تعداد فرزندان	اتومبیل دارد	حقوق دریافتی	وام رایس داده است؟
#1	1	2	1	800	بلی
#2	0	1	0	750	بلی
#3	0	2	1	700	بلی
#4	1	0	1	650	خیر
	⋮	⋮	⋮	⋮	⋮

دسته بندی (classification)

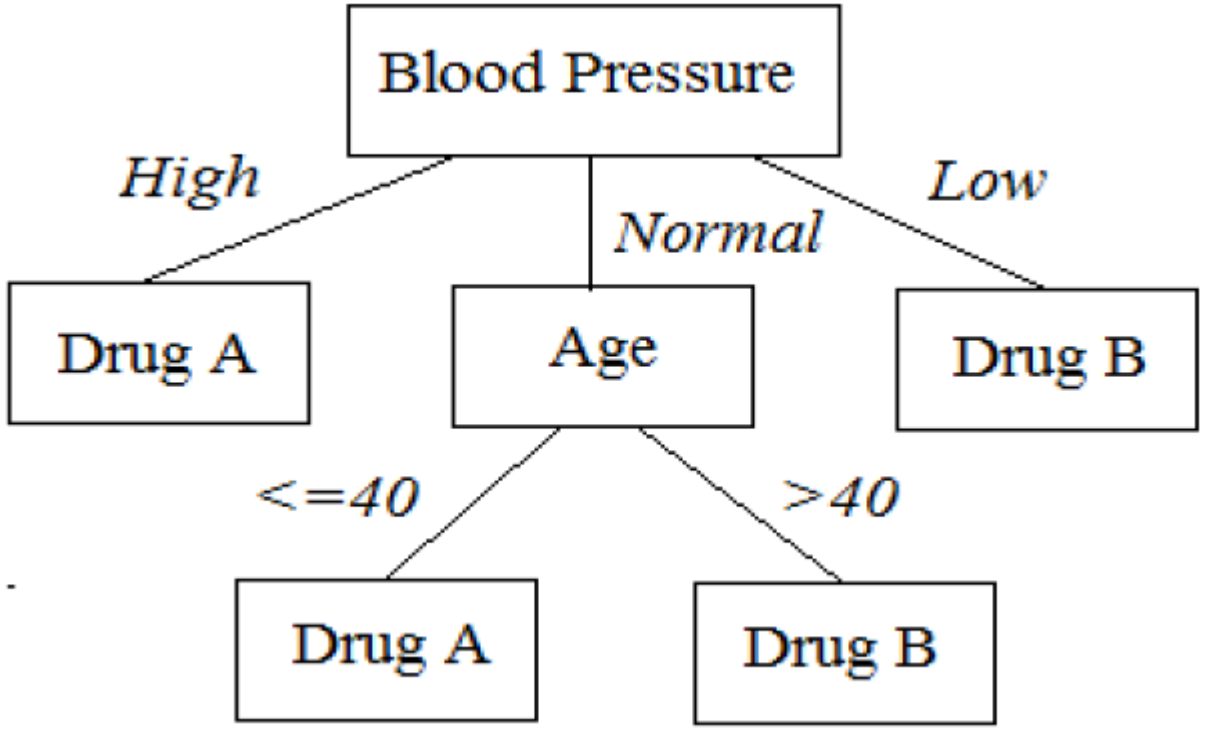
- دسته بندی دو فاز دارد:
- **1. آموزش (train):**
- یک مجموعه داده به همراه برچسب آن ها به مدل داده می شود (یا اینکه داده ها ذخیره می شوند) و براساس این داده ها پارامتر های مدل یادگرفته می شوند.
- **2. آزمایش (test):**
- یک داده بدون برچسب داده می شود و باید مشخص شود در چه کلاسی قرار می گیرد.

روش های دسته بندی

- **1) روش درخت تصمیم decision tree**
- یک درخت است که :
- 1. روی هر گره یک ویژگی ذخیره می شود (داده های category)
- 2. روی هر گره یک ویژگی به همراه یک (یا چند) آستانه ذخیره می شود.

درخت تصمیم

ID	Sex	Age	Blood P.	Drug
1	Male	20	Normal	A
2	Female	73	Normal	B
3	Male	37	High	A
4	Male	33	Low	B
5	Female	48	High	A
6	Male	29	Normal	A
7	Female	52	Normal	B
8	Male	42	Low	B
9	Male	61	Normal	B
10	Female	30	Normal	A
11	Female	26	Low	B
12	Male	54	High	A



درخت تصمیم

- سعی براین است که درخت تصمیم کمترین پیچیدگی را داشته باشد
- در داده کاوی در فاز train با افزایش پیچیدگی مدل، از مکانی به بعد خطای تست افزایش می یابد. در واقع اگر مدل زیاد آموزش ببیند و یا زیاد پیچیده شود، قابلیت تعمیم خود را در داده های test از دست می دهد. به این عمل اصطلاحا بیش برازش داده ها (overfitting) می گویند.

درخت تصمیم

- پیچیدگی محاسباتی در مدل درخت تصمیم می تواند عمق درخت یا تعداد نود ها (ویژگی ها) باشد.

درخت تصمیم

- **دو مسئله:**
- **1. شرط پایان الگوریتم:**
  - همه ی نمونه ها از مجموعه آموزش در یک کلاس باشند.
  - به حداکثر عمق درخت برسیم.
  - تعداد نمونه های موجود در گره از حداقل تعدادی که به عنوان آستانه مشخص می شود کمتر باشد.
  - مقادیر محاسبه شده برای انتخاب ویژگی برای هیچ ویژگی بیشتر از حد آستانه نیست.
- **2. مناسب ترین ویژگی در هر مرحله:**

معیار های انتخاب ویژگی در درخت تصمیم:

1. Information Gain:

$$Information\ Gain_A(D) = Entropy(D) - Entropy_A(D)$$

$$Entropy(D) = -\sum_{i=1}^{c} p_i * \log_2 p_i$$

$$Entropy_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Entropy(D_j)$$

cتعداد کلاس ها :

p_i احتمال وجود داده در کلاس i :

v تعداد اعضا(سطح های) ویژگی A :

D_jمی باشد v_j در آن ها برابر با A که ویژگی D قسمتی از داده های اولیه :

معیار های انتخاب ویژگی در درخت تصمیم:

- مثال از information gain:

ID	Age	Income	Job	Computer
1	Old	Medium	Student	No
2	Middle	High	Teacher	No
3	Old	Low	Teacher	No
4	Young	Medium	Teacher	Yes
5	Young	Low	Teacher	Yes
6	Old	Medium	Student	Yes
7	Middle	Medium	Student	Yes
8	Young	High	Teacher	No
9	Old	High	Student	No
10	Middle	High	Student	No

معیار های انتخاب ویژگی در درخت تصمیم:

- ادامه مثال از information gain:
- از آنجا که از 10 نمونه ی موجود در داده ها 4 نمونه دارای برچسب Yes و 6 نمونه ی دیگر دارای برچسب No هستند پس داریم:

$$Entropy(D) = -\frac{4}{10}Log_2(\frac{4}{10}) - \frac{6}{10}Log_2(\frac{6}{10}) = 0.970$$

- سطوح هر یک از ویژگی سن، درآمد و شغل عبارت است از:

$$Domain(Age) = \{Old, Middle, Young\}$$

$$Domain(Income) = \{High, Medium, Low\}$$

$$Domain(Job) = \{Teacher, Student\}$$

معیار های انتخاب ویژگی در درخت تصمیم:

ادامه مثال از information gain:

سپس باید برای این 3 ویژگی مقدار آنتروپی را محاسبه کنیم:

$$Entropy_{Age}(D) = \frac{4}{10} \times (-\frac{3}{4} Log_2(\frac{3}{4}) - \frac{1}{4} Log_2(\frac{1}{4})) +$$
$$\frac{3}{10} \times (-\frac{2}{3} Log_2(\frac{2}{3}) - \frac{1}{3} Log_2(\frac{1}{3})) +$$
$$\frac{3}{10} \times (-\frac{1}{3} Log_2(\frac{1}{3}) - \frac{2}{3} Log_2(\frac{2}{3})) = 0.875$$

$$Entropy_{Income}(D) = \frac{4}{10} \times (-\frac{4}{4} Log_2(\frac{4}{4}) - \frac{0}{4} Log_2(\frac{0}{4})) +$$
$$\frac{4}{10} \times (-\frac{1}{4} Log_2(\frac{1}{4}) - \frac{3}{4} Log_2(\frac{3}{4})) +$$
$$\frac{2}{10} \times (-\frac{1}{2} Log_2(\frac{1}{2}) - \frac{1}{2} Log_2(\frac{1}{2})) = 0.524$$

$$Entropy_{Job}(D) = \frac{5}{10} \times (-\frac{3}{5} Log_2(\frac{3}{5}) - \frac{2}{5} Log_2(\frac{2}{5})) +$$
$$\frac{5}{10} \times (-\frac{3}{5} Log_2(\frac{3}{5}) - \frac{2}{5} Log_2(\frac{2}{5})) = 0.970$$

معیار های انتخاب ویژگی در درخت تصمیم:

- ادامه مثال از information gain:
- پس از آن مقدار Information Gain برای همه ی ویژگی ها محاسبه می شود:

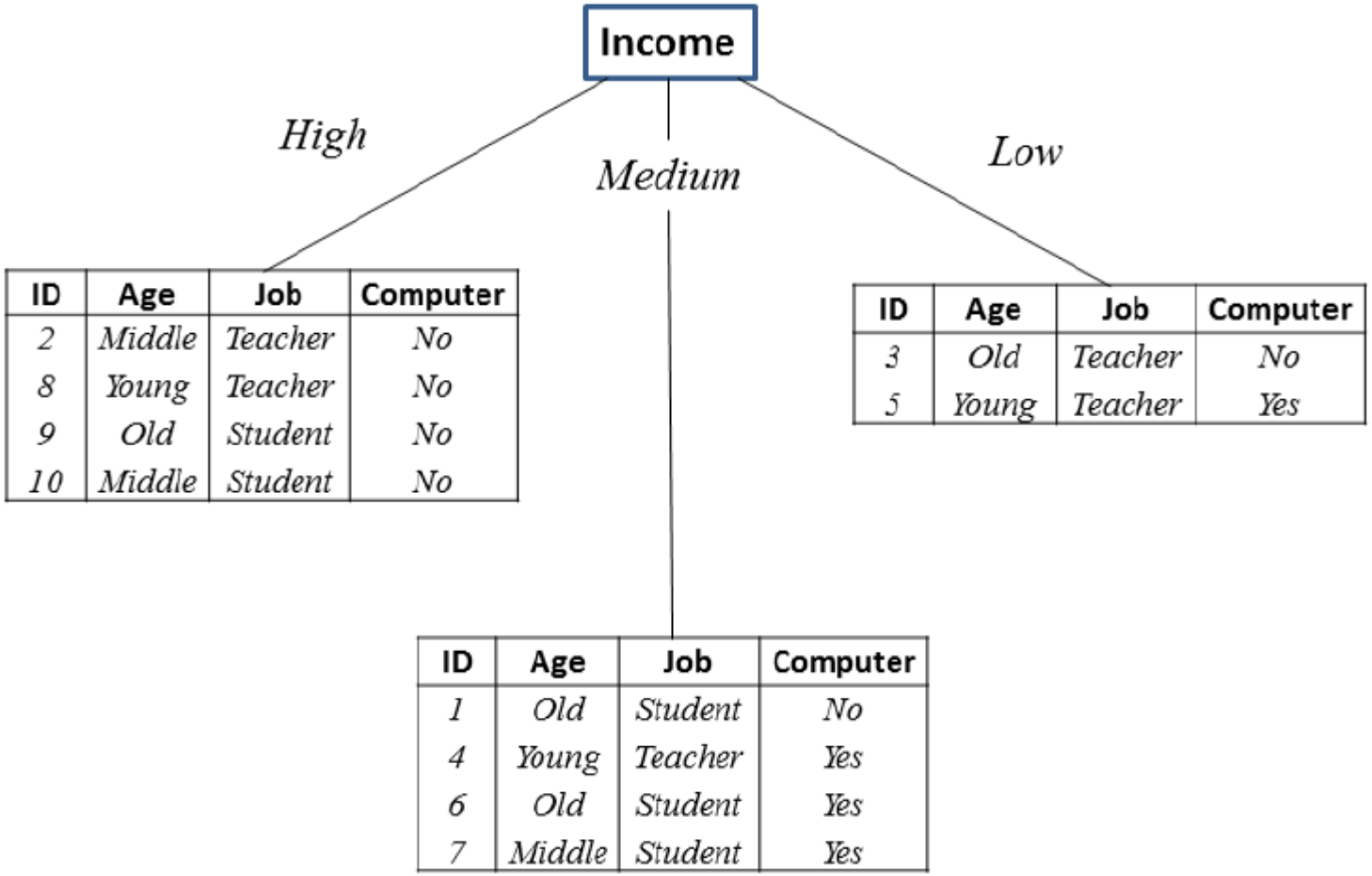
$$InformationGain(Age)=0.970-0.875=0.095$$

$$InformationGain(Income)=0.970-0.524=0.446$$

$$InformationGain(Job)=0.970-0.970=0$$

- ویژگی که دارای بیشترین Information Gain است انتخاب میشود.
- پس ویژگی در آمد که دارای بیشترین مقدار است برای ریشه درخت انتخاب میشود(شکل زیر):

معیار های انتخاب ویژگی در درخت تصمیم:



ID	Age	Job	Computer
2	Middle	Teacher	No
8	Young	Teacher	No
9	Old	Student	No
10	Middle	Student	No

ID	Age	Job	Computer
3	Old	Teacher	No
5	Young	Teacher	Yes

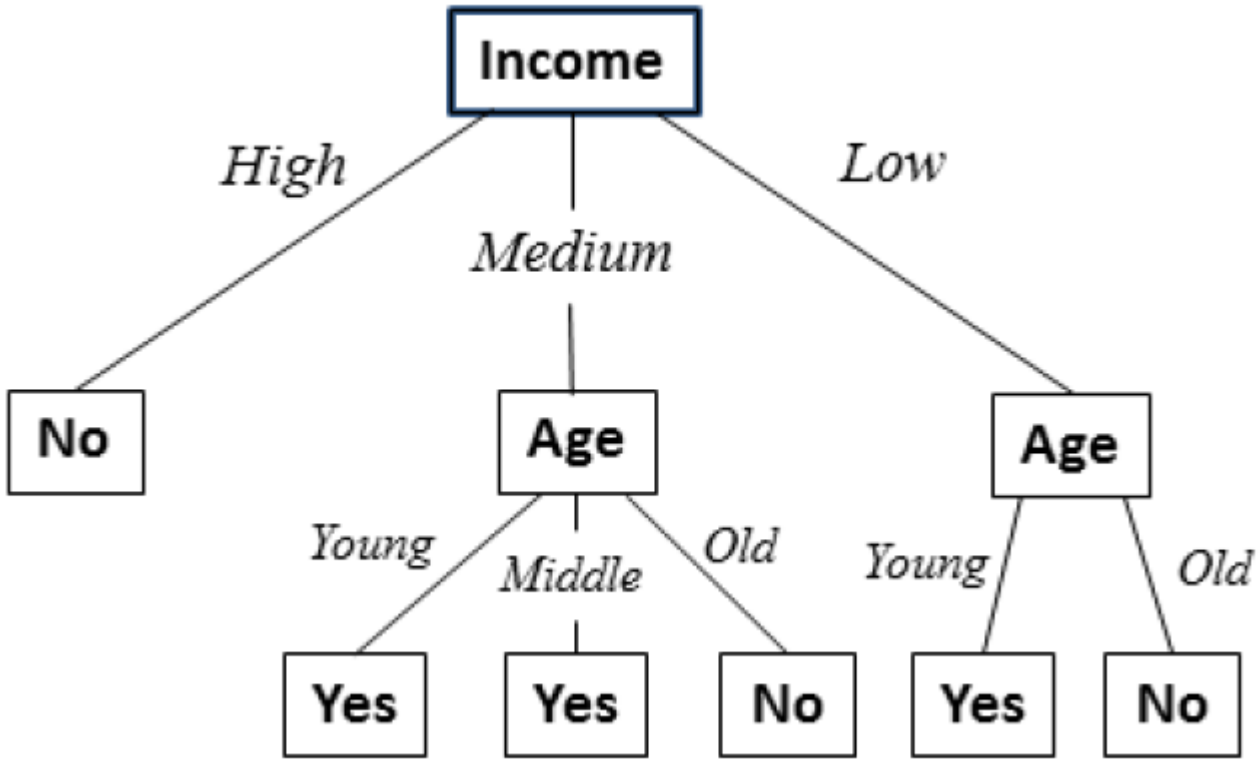
ID	Age	Job	Computer
1	Old	Student	No
4	Young	Teacher	Yes
6	Old	Student	Yes
7	Middle	Student	Yes

معیار های انتخاب ویژگی در درخت تصمیم:

- ادامه مثال از information gain:
- درواقع داده های آموزشی به 3 زیرمجموعه افراز می شوند. برای هر یک از این زیرمجموعه ها همانند قبل بهترین ویژگی جهت انشعاب محاسبه می شود. این بار ویژگی درآمد در محاسبات شرکت نمی کند و از میان دیگر ویژگی ها یکی انتخاب می شود. این کار تا هنگامی که یکی از شروط توقف الگوریتم محقق شود، ادامه می یابد.
- درخت تصمیم نهایی به شکل زیر است:

معیار های انتخاب ویژگی در درخت تصمیم:

- ادامه مثال از information gain:



معیار های انتخاب ویژگی در درخت تصمیم:

2. Gini Index:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$

$$Gini_A(D) = \frac{|D_1|}{|D|} * Gini(D_1) + \frac{|D_2|}{|D|} * Gini(D_2) + .. = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Gini(D_j)$$

- ویژگی با Gini Index کوچکتر انتخاب می شود و یا MAX تغییر درجه ناخالص انتخاب می شود.
- $\Delta Gini(D) = Gini(D) - Gini_A(D)$ = تغییر درجه ناخالص ویژگی
- معمولا v=2 در نظر گرفته می شود.

معیار های انتخاب ویژگی در درخت تصمیم:

- 3. Gain Ratio: نرمال شده ی Information Gain است

$$Gain\ Ratio_A(D) = \frac{Information\ Gain_A(D)}{splite\ info(A)}$$

$$splite\ info(A) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} * \log_2 \frac{|D_j|}{|D|}$$

معیار های انتخاب ویژگی در درخت تصمیم:

- **4. Likehood Ratio:**
- $$Likehood_A(D) = 2 * \ln 2 * |D| * Information\ Gain(A)$$
- **5. DKM: برای مسئله ی دو کلاسه است:**
- $$DKM_A(D) = 2 * \sqrt{\frac{|D_1|}{|D|} * \frac{|D_2|}{|D|}}$$

الگوریتم های درخت تصمیم

- **1. ID3**
- از معیار Iformation Gain برای انتخاب ویژگی استفاده می کند.
- شرط توقف در هر برگ:

1. نمونه های باقی مانده در برگ متعلق به یک کلاس باشند.
2. برگ برای تمام ویژگی ها کوچکتر مساوی صفر Information Gain باشد.

الگوریتم های درخت تصمیم

- **3. C4.5**
- از معیار Gain Ratio برای انتخاب ویژگی استفاده می کند
- توقف هنگامی رخ می دهد که:
  1. تعداد نمونه ها کمتر از مقدار مشخص شده ای باشد.
  2. متعلق به یک کلاس باشند.

ادامه روش های دسته بندی

- **2) روش K: KNN (K Nearest Neighbor) نزدیکترین همسایه**

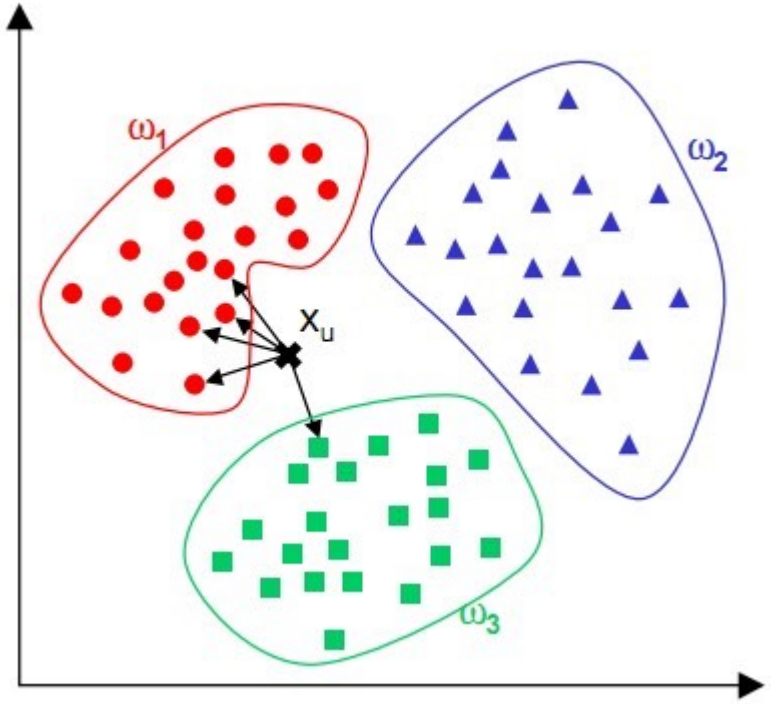
1. نزدیکترین داده به داده تست را از مجموعه آموزش پیدا میکند K
2. سپس رای گیری می کند (براساس برچسب آن ها)
3. در نهایت برچسب خروجی Test مشخص می شود.

- معیار فاصله ی اقلیدسی است:

$$dis_{\text{اقلیدسی}}(x, \acute{x}) = \sqrt{\sum_{i=1}^{d} (x_i - \acute{x}_i)^2}$$

KNN

- چرا به جای k از یک نزدیکترین همسایه استفاده نمیکنیم؟
- اگر تعداد رای ها مساوی شود چی می کنند؟



ادامه روش های دسته بندی

- **3) روش ε :εNNهمسایگی**

- یک عدد از قبل مشخص شده است. برای داده های تست، در شعاع ε تعداد برچسب ها به رای گیری گذاشته می شوند و برچسب داده داده های تست مشخص می شود.

- **سوال:** اگر در شعاع مشخص شده در داده های train هیچ داده ای موجود نباشد یا تعداد رای ها مساوی باشد چه اتفاقی می افتد؟

- **پاسخ:** شعاع را کم کم بزرگ مي کنيم

ادامه روش های دسته بندی

- **4) روش Naive Bayes (بیز):**
- در این روش با استفاده از احتمالات، احتمال رخداد نمونه را در همه ي كلاس ها بدست مي آوريم سپس آن كلاسي كه بيشترين احتمال رخداد را براي اين نمونه داشت به عنوان كلاس مورد نظر براي اين نمونه ي تست در نظر گرفته مي شود.
- فرض براين است كه تاثير مقدار ويژگي روي برچسب كلاس، مستقل از مقادير ديگر ويژگي ها است (استقلال شرطي ويژگي ها Conditional Independence)

Naive Bayes

- $d = < a_1, \ldots, a_n >$
- $$p(c = c_i \mid d) = p(c = c_i \mid A_1 = a_1,\ A_2 = a_2, \ldots,\ A_n = a_n) = \frac{p(A_1 = a_1, \ldots,\ A_n = a_n \mid c = c_i) * p(c = c_i)}{p(A_1 = a_1, \ldots, A_n = a_n)} = \frac{p(A_1 = a_1, \ldots,\ A_n = a_n \mid c = c_i) * p(c = c_i)}{\sum_{k=1}^{m} p(A_1 = a_1, \ldots,\ A_n = a_n \mid c = c_k) * p(c = c_k)}$$
- با توجه به اينكه مخرج كسر بالا براي همه ي كلاس ها روي اين نمونه يک عدد مي شود مي توان آن را در نظر نگرفت.

Naive Bayes

- $p(c = c_i \mid d) = p(c = c_i) * p(A_1 = a_1 \mid A_2 = a_2, \ldots, A_n = a_n,\ c = c_i) * \mathrm{p}(A_2 = a_2, \ldots, A_n = a_n \mid c = c_i)$
- $p(c = c_i \mid d) = p(c = c_i) * p(A_1 = a_1 \mid A_2 = a_2, \ldots, A_n = a_n,\ c = c_i) * p(A_2 = a_2 \mid A_3 = a_3, \ldots, A_n = a_n, c = c_i) * p(A_3 = a_3, \ldots, A_n = a_n \mid c = c_i) =$

  $$p(c = c_i) * p(A_1 = a_1 \mid c = c_i) * p(A_2 = a_2 \mid c = c_i) * \ldots * p(A_n = a_n \mid c = c_i)$$
- $$\boldsymbol{p(c = c_i \mid d) = p(c = c_i) * \prod_{j=1}^{n} p(A_j = a_j \mid c = c_i)}$$

Naive Bayes

- زماني كه ممكن است $p\left(A_j = a_j \middle| c = c_i\right) = 0$ شود مي توان به جاي رابطه زير :

$$p\left(A_j = a_j \middle| c = c_i\right) = \frac{n_{ji}}{n_i}$$

- از رابطه زير استفاده كرد: (λ مي تواند 1 يا $\frac{1}{2}$ باشد)

$$p\left(A_j = a_j \middle| c = c_i\right) = \frac{n_{ji} + \lambda}{n_i + \lambda * d_j}$$

- d_j تعداد اعضا دامنه ويژگي : A_j
- n_{ji} : تعداد نمونه هايي كه در آن A_j برابر a_j استو در كلاس c_i هستند.
- n_i : تعداد داده ها با برچسب c_i

Naive Bayes

ID	A	B	C
1	a	d	y
2	a	e	y
3	b	f	y
4	c	e	y
5	b	f	y
6	b	f	n
7	b	e	n
8	c	d	n
9	c	f	n
10	a	d	n

- مثال: برچسب نمونه تست data=<a,f>

$P(C=y)=5/10=1/2 \qquad P(C=n)=5/10=1/2$

$P(A=a|C=y)=2/5 \quad P(A=b|C=y)=2/5 \quad P(A=c|C=y)=2/5$

$P(A=a|C=n)=1/5 \quad P(A=b|C=n)=2/5 \quad P(A=c|C=n)=2/5$

$P(B=d|C=y)=1/5 \quad P(B=e|C=y)=2/5 \quad P(B=f|C=y)=2/5$

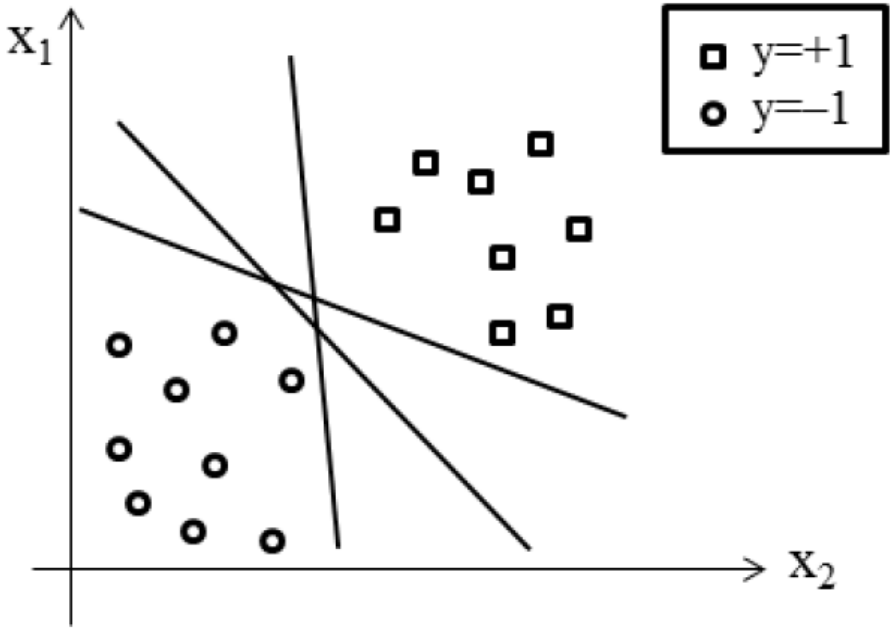
$P(B=d|C=n)=2/5 \quad P(B=e|C=n)=1/5 \quad P(B=f|C=n)=2/5$

$$P(C=y)\times\prod_{j=1}^{2}P(A_j=a_j\,|\,C=y)=\frac{1}{2}\times\frac{2}{5}\times\frac{2}{5}=\frac{4}{50}=0.08$$

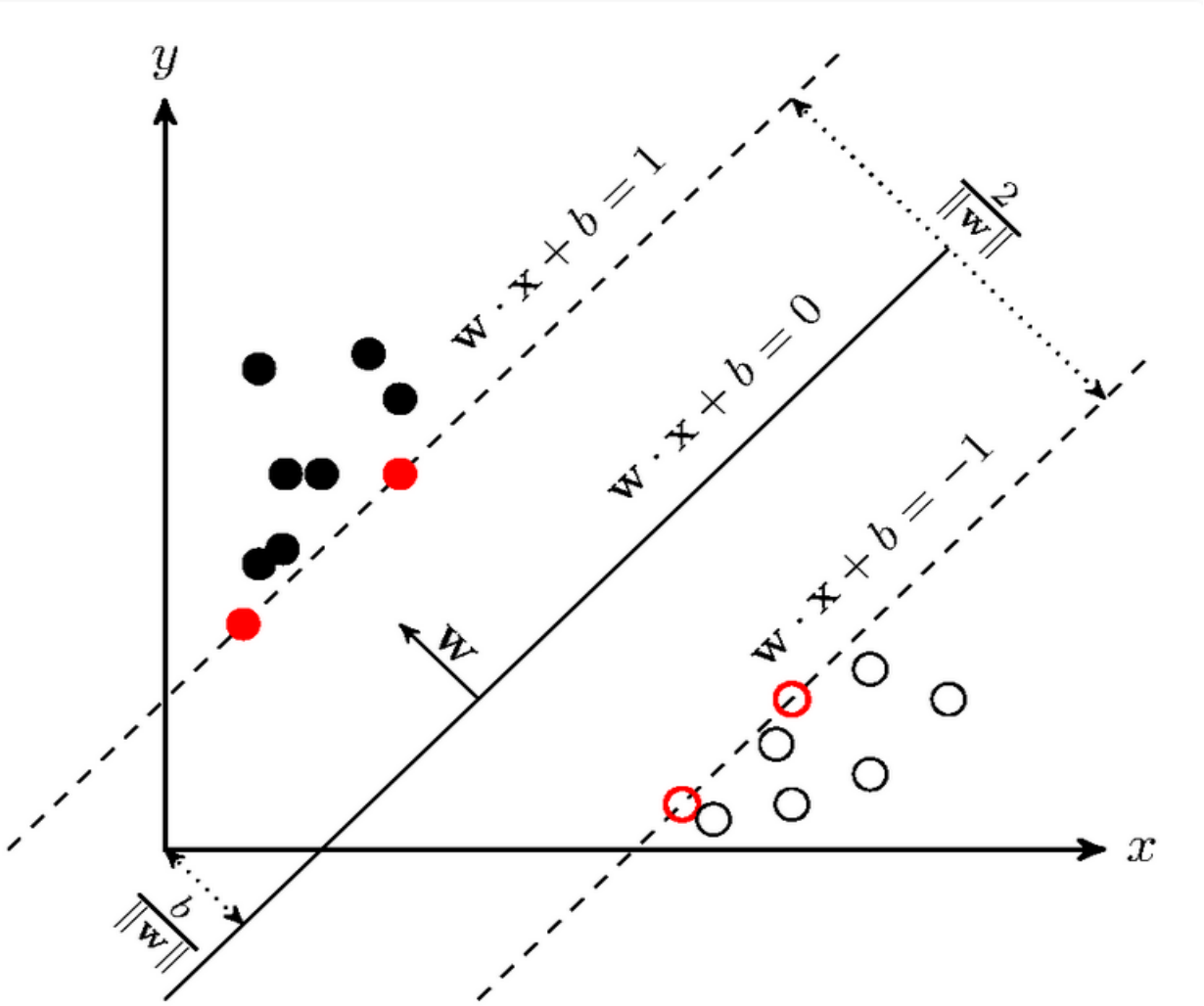
$$P(C=n)\times\prod_{j=1}^{2}P(A_j=a_j\,|\,C=n)=\frac{1}{2}\times\frac{1}{5}\times\frac{2}{5}=\frac{2}{50}=0.04$$

ادامه روش های دسته بندی

- **5) روش دسته بندي ماشين بردار پشتيبان (SVM):**
- داده ها را با یک خط به دو قسمت تقسیم کن (دنیا ساده است).
- اين روش براي صورت مسئله هاي دو كلاسه تعريف شده است ولي براي صورت مسئله هاي چند كلاسه گسترش داده مي شود.



SVM (Support Vector Machine)



SVM

- $\|w\| = \sqrt{w_1^2 + w_2^2 + \ldots + w_d^2}$
- اگر d ، x بعدي باشد معادله خط جداکننده به معادله ابرصفحه (hyper plane) تبديل مي شود.
- $WX + b = 0 \rightarrow w_1x_1 + w_2x_2 + \ldots + w_dx_d + b = 0$
- خط جداساز توسط داده هاي آموزشي پيدا مي شود سپس در مرحله ي تست:

$wx + b \geq 0 \rightarrow \quad y = 1$

- $wx + b \leq 0 \rightarrow \quad y = -1$

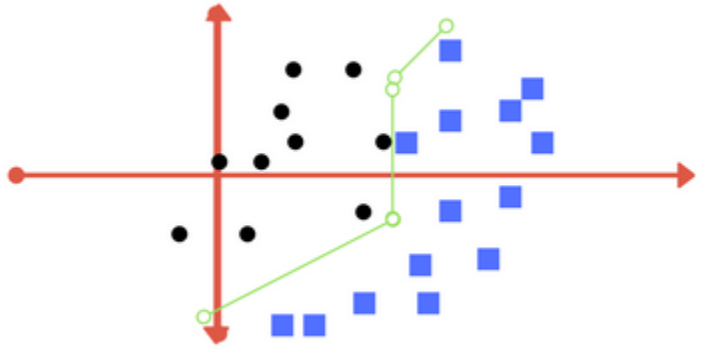
SVM

برچسب =X $sign(\sum_{i=1}^{s} y_i \alpha_i X_i X^T + b)$

- y_i برچسب بردار پشتيبان :
- X_i بردار پشتيبان :
- α_i مقدار مناسب آن پيدا train عدد ثابتي نيست كه بايد در مرحله ي شود.

SVM

- در بعضي مواقع داده ها به صورت خطي از هم جدا پذير نيستند در اين مواقع با اعمال توابع كرنل روي داده ها آن ها را از فضاي d بعدي به فضاي n بعدي مي بريم كه تفكيک پذيري با خط ممكن شود.



SVM

- کرنل هايي که بيشتر مورد استفاده قرار مي گيرند:

$$polynomial \rightarrow k\left(\vec{x_i}, \vec{x_j}\right) = (\vec{x_i} * \vec{x_j} + 1)^d$$

$$Gaussian - Radial\ Basis\ Function(RBF) \rightarrow k\left(\vec{x_i}, \vec{x_j}\right) = \exp\left(-\frac{\left\|\vec{x_i} - \vec{x_j}\right\|^2}{2\delta^2}\right)$$

$$hyperbolic\ tangant \rightarrow k\left(\vec{x_i}, \vec{x_j}\right) = \tanh(\alpha \vec{x_i} \vec{x_j} + c)$$

$$linear \rightarrow k\left(\vec{x_i}, \vec{x_j}\right) = \vec{x_i} \,.\, \vec{x_j}$$

SVM

- $Gram\ Matrix(i,j) = kernel\left(\vec{x_i}, \vec{x_j}\right)$
- در ستون ژام از $Gram\ Matrix$ بيان مي كنيم كه هركدام از داده ها چقدر به داده ي ژام شباهت دارند(مقايسه مي شوند).

SVM

- زماني كه يک دسته بند براي صورت مسئله ي دو كلاسه طراحي شده است چگونه آن را به مسئله ي چند كلاسه گسترش دهيم؟

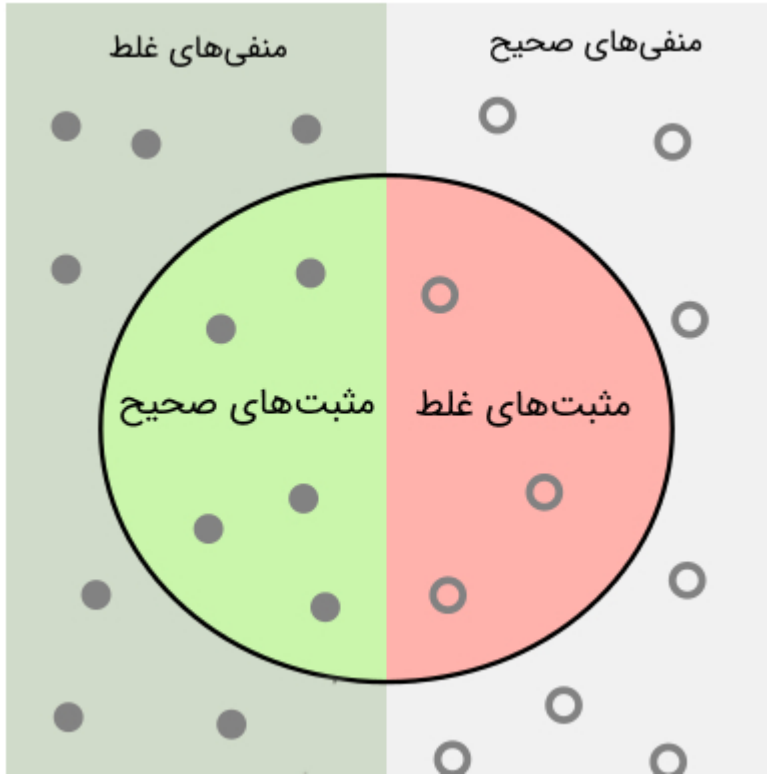
1. One to one : 1و1 ، ... ، 3و1 ، 2و k ، 2 و2 ، ... ، 4و2 ، 3و k ، .....، kو k-1
2. One to other: 1 2,3 در مقابل ,...,k

2 در مقابل 1,3,...,k

و...

معيار ارزيابي روش هاي دسته بندي

- معيار دقت در روش هاي دسته بندي براي مرحله ي test :
- در اين معيار ها برچسب هاي پيش بيني شده براي داده ها با برچسب هاي واقعي مقايسه مي شوند.



معيار ارزيابي روش هاي دسته بندي

- $$precision = \frac{TP}{TP + FP}$$
- $$recall = \frac{TP}{FN + TP}$$
- $$F - Measure = \frac{\alpha * precision * recall}{recall + precision} \qquad \alpha=2$$
- $$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

خوشه بندي clustering

- هدف: كنار هم قرار دادن داده هاي مشابه به صورتي كه داده هاي كه هم خوشه نيستند در خوشه هاي متفاوت قرار بگيرند.
- در واقع الگوريتم هايي كه براي خوشه بندي ارائه مي شوند دو هدف زير را دنبال مي كنند:

1. فاصله ي درون كلاسي كم
2. فاصله ي برون كلاسي زياد

- بعضي مواقع تعداد خوشه ها(براي خوشه بندي)از قبل مشخص است در بعضي مواقع مشخص نيست

clustering

- تعداد خوشه ها مشخص باشد 2 روش وجود دارد:
- **1. k-means:**
- مراحل:

1) برچسب گذاري تصادفي داده ها(اينكه متعلق به كدام خوشه هستند)
2) محاسبه ميانگين و قرار دادن آن به عنوان مركز خوشه
3) برچسب گذاري جديد بر اساس فاصله تا نزديكترين مركز خوشه
4) تكرار 2 و3 تا شرط پايان

- **2. k-medoids**