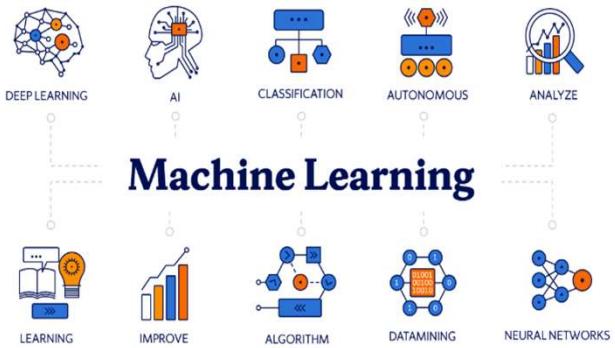
بسمه تعالی

فصل سوم یادگیری ماشین

دسته بندی
Classification



مدرس

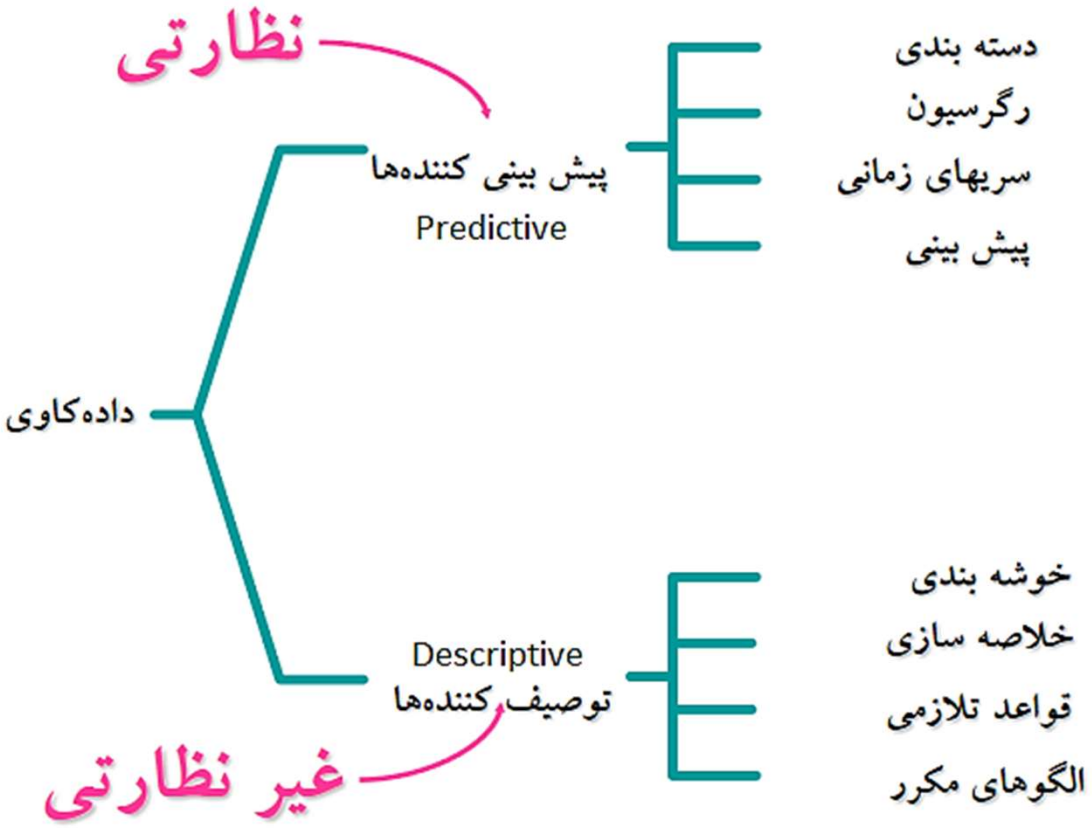
فاطمه دارائی

f_daraei@semnan.ac.ir

https://fdaraei.profile.semnan.ac.ir

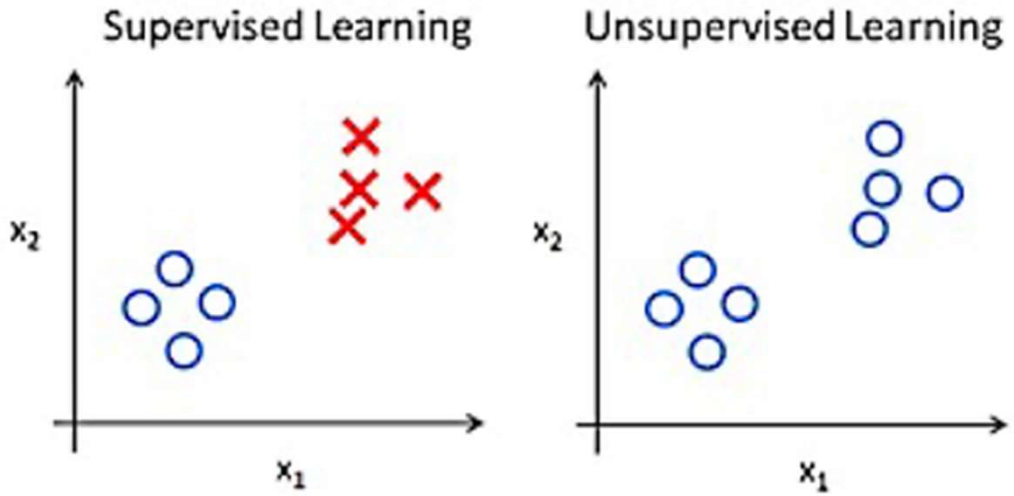
دانشگاه سمنان
Semnan University
پردیس فرزانگان

تکنیک های داده کاوی



یادگیری با نظارت Supervised learning

- این نوع یادگیری، نگاشت داده‌های ورودی به اهداف شناخته شده با توجه به مجموعه‌ای از نمونه‌ها (اغلب توسط انسان‌ها برچسب گذاری شده) است.
- داده های آموزشی همراه با برچسب هستند که کلاس آن داده ها را نشان می‌دهند.
- بیشترین کاربرد و بالاترین دقت در این نوع یادگیری است.



دانشگاه سمنان
Semnan University
پردیس فرزانگان

دسته بندی Classification

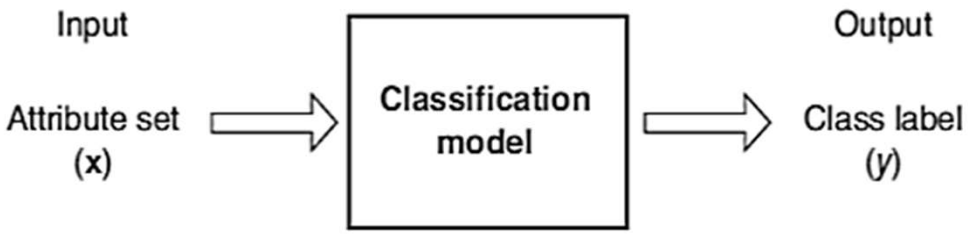
دسته بندی وظیفه یادگیری تابع هدف f است که مجموعه x را به یکی از برچسب های کلاس از پیش تعریف شده y نگاشت می کند.
مجموعه آموزشی شامل رکوردهایی با برچسب های کلاس شناخته شده است و هر رکورد شامل مجموعه‌ای از ویژگی‌ها است که یکی از آن‌ها کلاس است.

وظیفه: یافتن مدلی برای ویژگی کلاس به عنوان تابعی از مقادیر سایر ویژگی‌ها.

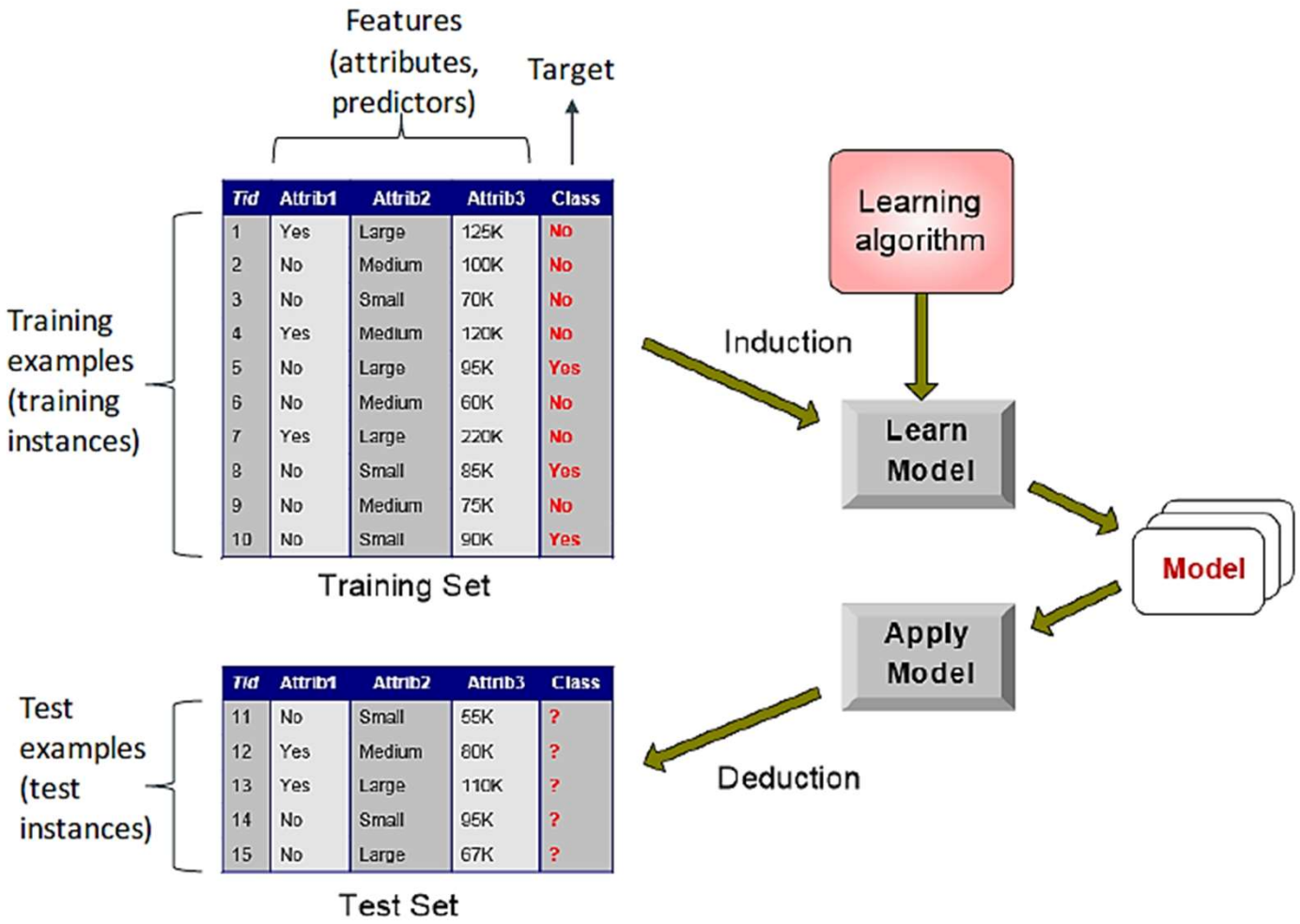
هدف: رکوردهای قبلاً دیده نشده باید با دقت بالا به یک کلاس اختصاص داده شوند.

یک مجموعه تست برای تعیین دقت مدل استفاده می‌شود. معمولاً، مجموعه داده‌های داده شده به مجموعه‌های آموزشی و تست تقسیم می‌شود،

با این تفاوت که مجموعه آموزشی برای ساخت مدل و مجموعه تست برای اعتبارسنجی آن استفاده می‌شود.



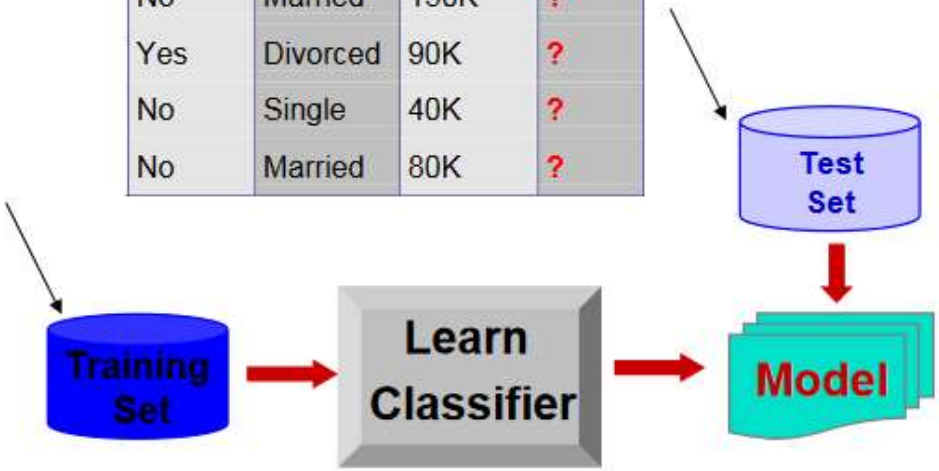
مثال دسته بندی Classification



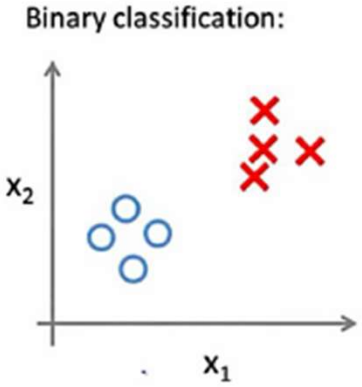
Two class labels (or classes): Yes (1), No (0)

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

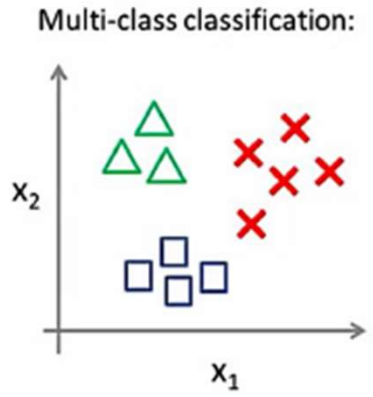


دسته بندی باینری و دسته بندی چند کلاسه



Binary classification:

وظایف طبقه‌بندی با تنها دو کلاس، که معمولاً با {+،−}، {+۱،−۱} یا {مثبت، منفی} نشان داده می‌شوند.
مثال: تشخیص مثبت بودن بیماری، تحلیل احساسات (مثبت/منفی).



Multiclass classification:

وظایف طبقه‌بندی با بیش از دو کلاس.
مثال: شناسایی موضوع ایمیل، تحلیل احساسات (مثبت/خنثی/منفی).

تفاوت رگرسیون و دسته بندی

دسته بندی

پیش‌بینی برچسب‌های دسته‌ای کلاس (گسسته یا نامی).

مثال: وضعیت آب و هوا برای فردا چیست؟ (بارانی، ابری، آفتابی)

طبقه‌بندی یا Classification زمانی استفاده می‌شود که هدف پیش‌بینی یک برچسب دسته‌ای باشد، مانند نوع آب و هوا یا دسته‌بندی ایمیل به اسپم یا ایمیل عادی.

در طبقه‌بندی، خروجی مدل معمولاً از مجموعه‌ای محدود از کلاس‌ها انتخاب می‌شود (مثلاً بین چند دسته).

رگرسیون

مدل‌سازی توابع با مقدارهای پیوسته.

مثال: دمای فردا چقدر خواهد بود؟

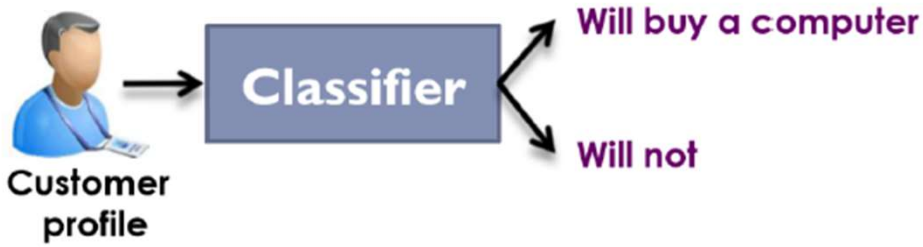
رگرسیون یا Regression برای پیش‌بینی مقادیر عددی پیوسته استفاده می‌شود، مانند دمای هوا یا قیمت خانه.

در رگرسیون، خروجی مدل یک مقدار عددی پیوسته است و می‌تواند هر مقداری در یک دامنه باشد.

دانشگاه سمنان
Semnan University
پردیس فرزانگان

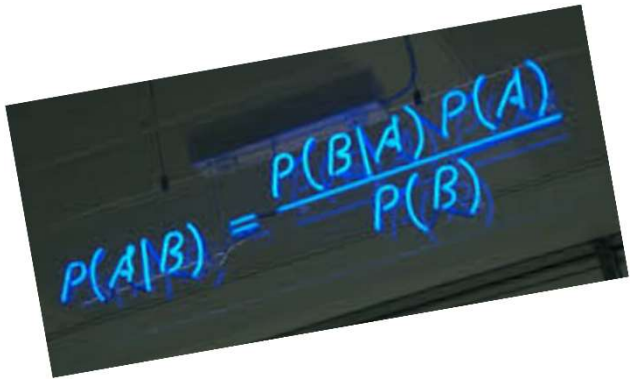
مثال از تفاوت رگرسیون و دسته بندی

یک مدیر بازاریابی دوست دارد بداند آیا یک مشتری خاص محصولی را خواهد خرید یا خیر.



یک مدیر بازاریابی دوست دارد پیش‌بینی کند چه مقدار یک مشتری مشخص در طول یک فروش صرف خواهد کرد.

روش بیزین Bayesian Method



- یک روش احتمالاتی است.
  - پیش‌بینی احتمال کلاس‌ها: از مدل‌های احتمالی برای پیش‌بینی کلاس یا دسته‌بندی استفاده می کند. به‌جای گفتن هر پاسخ دقیق، خروجی به‌صورت احتمال نسبت به هر کلاس ارائه می‌شود.

- دسته‌بندی با Naïve Bayes
  - یک کلاس از الگوریتم‌های دسته‌بندی است که از قضیه بیز برای تخمین احتمال عضویت یک نمونه استفاده می‌کند.
- مبتنی بر قضیه بیز Bayes' theorem
  - مدل با استفاده از قانون بیز کار می‌کند که پاشش احتمال شرطی بین ویژگی‌ها و کلاس هدف را محاسبه می‌کند. دقت و سرعت قابل قبول هنگام استفاده با پایگاه‌های داده بزرگ دارد.
- به‌صورت افزایشی Incremental است. یعنی
  - می‌تواند به‌طور افزایشی به‌روزرسانی شود، بدون نیاز به بازسازی کامل مدل.

Bayes' theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

- اثبات آن ساده است.
- فرض کنید X یک نمونه داده با برچسب کلاس ناشناخته باشد.
- H فرضیه ای می باشد که X به کلاس C متعلق است.
- طبقه‌بندی برای تعیین P(H|X) است (احتمال پسین) (posteriori probability)
- P(H) احتمال پیشین (prior probability): مقدار احتمال اولیه
- مثلاً، X برای خرید کامپیوتر باشد، بدون توجه به سن، درآمد و غیره
- P(X) احتمال مشاهده‌ی نمونه
- P(X|H) احتمالِ صحت/احتمالِ تحقّق، احتمال مشاهده‌ی نمونه X با فرض درست بودن فرض
- مثلاً اگر X خرید کامپیوتر باشد، احتمال اینکه X 31...40 ساله، با درآمد متوسط باشد.

دسته بندی برمبنای قانون بیز

هدف: تخصیص نمونه X به یکی از کلاس‌های Ci با بیشترین احتمال پسین P(Ci|X)

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

چون P(X) مستقل از کلاس است و برای همه کلاس‌ها ثابت است، برای تصمیم‌گیری تنها باید معیار P(X|Ci) P(Ci) را مقایسه کرد.

در نتیجه برای تصمیم‌گیری نهایی، از قاعده قطعیت استفاده می‌شود: کلاس با بیشترین مقدار P(X|Ci) P(Ci) انتخاب می‌شود.

$$\underset{i}{\operatorname{argmax}}\, P(C_i|\boldsymbol{X}) = \underset{i}{\operatorname{argmax}}\, P(\boldsymbol{X}|C_i)P(C_i)$$

دانشگاه سمنان
Semnan University
پردیس فرزانگان

مثال: احتمال اینکه یک فرد سیگاری سرطان داشته باشد

Smoke	Cancer
Y	N
Y	Y
N	N
N	Y

S \ C	Y	N
Y	200	1800
N	100	2200

$$P(\text{cancer}) = \frac{300}{4300}$$

$$P(\text{smoke} \mid \text{cancer}) = \frac{200}{300}$$

$$P(\text{cancer} = \text{yes} \mid \text{smoke} = \text{yes}) = \frac{P(\text{cancer})\,P(\text{smoke} \mid \text{cancer})}{P(\text{smoke})} = \frac{\frac{300}{4300} \times \frac{200}{300}}{a}$$

$$P(\overline{\text{cancer}} \mid \text{smoke}) = 1 - P(\text{cancer} \mid \text{smoke}) = \frac{P(\overline{\text{cancer}})\,P(\text{smoke} \mid \overline{\text{cancer}})}{P(\text{smoke})} = \frac{\frac{4000}{4300} \times \frac{1800}{4000}}{a}$$

If you need to compute the exact probabilities: $a = \frac{3}{43} \times \frac{2}{3} + \frac{40}{43} \times \frac{18}{40}$

مثال

Age	Weight	Blood pressure	Healthy / Patient
25	60	12	Healthy
27	80	14	Patient
...	...	...	...

X: (Age=23, Weight=65, Blood pressure=13)

$$P(\text{healthy} \mid X) = \frac{P(\text{healthy})\, P(X \mid \text{healthy})}{P(X)}, \quad P(\text{patient} \mid X) = \frac{P(\text{patient})\, P(X \mid \text{patient})}{P(X)}$$

$$\frac{P(\text{healthy})\, P(X \mid \text{healthy})}{P(X)} \gtrless \frac{P(\text{patient})\, P(X \mid \text{patient})}{P(X)}$$

$$P(\text{healthy}) = \frac{\text{count of healthy people}}{\text{count of examples}}, \quad P(\text{patient}) = \frac{\text{count of patient people}}{\text{count of examples}} = 1 - P(\text{healthy})$$

$$P(X \mid \text{healthy}) = \frac{\text{count of healthy people with age=23 \& weight=65 \& blood pressure=13}}{\text{count of healthy}}$$

→ Unlikely!!

دسته بند Naïve Bayes

فرض ساده‌سازی: ویژگی‌ها به طور شرطی نسبت به یکدیگر مستقل‌اند.

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

P(xk|Ci) احتمال شرطی است که نشان می‌دهد ویژگی Ak مقدار xk را دارد، با فرض اینکه نمونه به کلاس Ci تعلق دارد.

مثال

RID	Age	Income	Student	Credit	Buy
1	youth	high	No	fair	No
2	youth	high	No	excellent	No
3	middle	high	No	fair	Yes
4	senior	medium	No	fair	Yes
5	senior	low	Yes	fair	Yes
6	senior	low	Yes	excellent	No
7	middle	low	Yes	excellent	Yes
8	youth	medium	No	fair	No
9	youth	low	Yes	fair	Yes
10	senior	medium	Yes	fair	Yes
11	youth	medium	Yes	excellent	Yes
12	middle	medium	No	excellent	Yes
13	middle	high	Yes	fair	Yes
14	senior	medium	No	excellent	No

Prior probabilities	P (buy = Y)	9 / 14 = 0.643
	P (buy = N)	5 / 14 = 0.357
Compute $P(X_k \mid C_i)$ for each class and each attribute		
P (age $\leq$ 30 | buy = Y)		2 / 9 = 0.222
P (age $\leq$ 30 | buy = N)		3 / 5 = 0.600
P (31 $\leq$ age $\leq$ 40 | buy = Y)		4 / 9 = 0.444
P (31 $\leq$ age $\leq$ 40 | buy = N)		0
P (student = Y | buy = Y)		6 / 9 = 0.667
P (student = Y | buy = N)		1 / 5 = 0.200
P (credit = F | buy = Y)		6 / 9 = 0.667
P (credit = F | buy = N)		2 / 5 = 0.400
...		...

Training phase

دانشگاه سمنان
Semnan University
پردیس فرزانگان

ادامه مثال

X = (age <= 30 , income = medium, student = yes, credit_rating = fair)

P(X|buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

P(X|buys_computer = “yes”) x P(buys_computer = “yes”) = 0.044 x 0.643 = 0.028
P(X|buys_computer = “no”) x P(buys_computer = “no”) = 0.019 x 0.357 = 0.007

Therefore, X belongs to class (“buys_computer = yes”)
With what probability?

حل مشکل احتمال صفر

نایو بییز Naïve Bayes برای هر احتمال شرطی باید غیرصفر باشد. در غیر این صورت، احتمال پیش‌بینی شده برابر با صفر خواهد بود.”

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

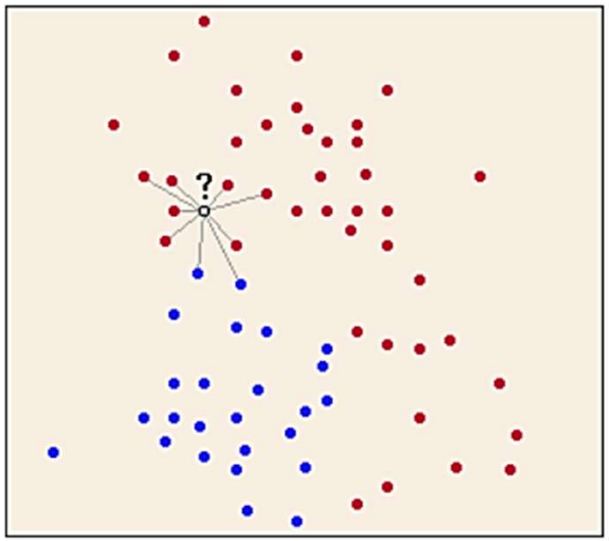
- Example:
  - Consider a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)
- Use Laplacian correction (or add-one smoothing)
  - Adding 1 to each case
    - Prob(income = low) = 1/1003
    - Prob(income = medium) = 991/1003
    - Prob(income = high) = 11/1003

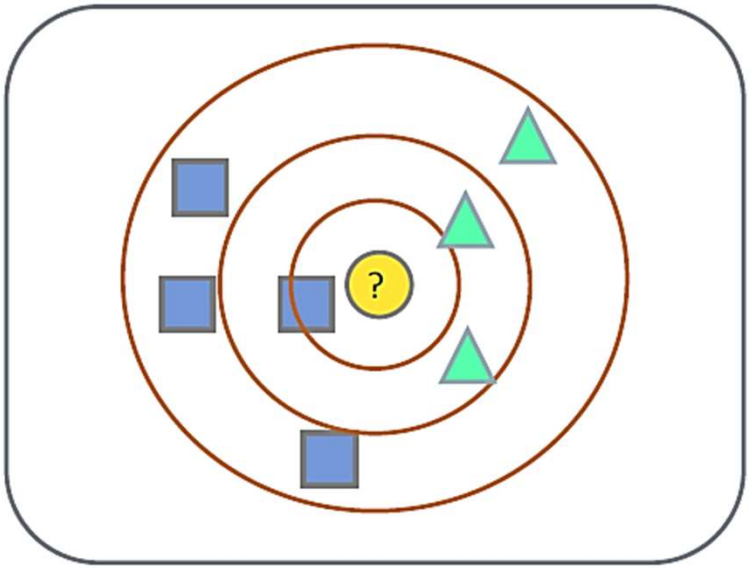
دانشگاه سمنان
Semnan University
پردیس فرزانگان

K نزدیک‌ترین همسایه
K-Nearest Neighbors (KNN)

از ساده‌ترین الگوریتم‌های داده‌کاوی است.
نیازمند سه چیز است:

- فضای ویژگی‌ها (داده‌های آموزشی)
- معیار فاصله
- مقدار k





- k = 1:
  - Belongs to square class
- k = 3:
  - Belongs to triangle class
- k = 7:
  - Belongs to square class

KNN

- قابل استفاده برای مسائل طبقه‌بندی و رگرسیون
- ترکیب برچسب‌های k نزدیک‌ترین همسایه:
- گرفتن رأی اکثریت (میانگین) برچسب‌ها بین همسایه‌ها
- وزن‌دهی: $w = \frac{1}{d} \text{ or } \frac{1}{d^2}$
- مثال: فرض کنید این‌ها سه نزدیک‌ترین همسایه هستند.

Value	5	8	9
Distance	3	2	5

$$\text{prediction} = \frac{\frac{1}{3}\times 5 + \frac{1}{2}\times 8 + \frac{1}{5}\times 9}{\frac{1}{3}+\frac{1}{2}+\frac{1}{5}}$$

انتخاب مقدار K :

- اگر k خیلی کوچک باشد، مدل به نقاط نویزی حساس می‌شود.
- اگر k خیلی بزرگ باشد، همسایگی ممکن است شامل نقاط نامربوط شود.
- مقدار k را فرد انتخاب کنید تا از تساوی آرا جلوگیری شود.

1. دسته‌بندهای نزدیک‌ترین همسایه، یادگیرندگان تنبل هستند:
همه محاسبات تا زمان طبقه‌بندی به تعویق می‌افتد.

2. مرحله آموزش سریع است، اما مرحله اعمال (پیش‌بینی) بسیار کند است.
3. نیاز به تعداد زیادی نمونه آموزشی دارد.
4. روش KNN یک روش یادگیری مبتنی بر نمونه (غیرپارامتری) است.

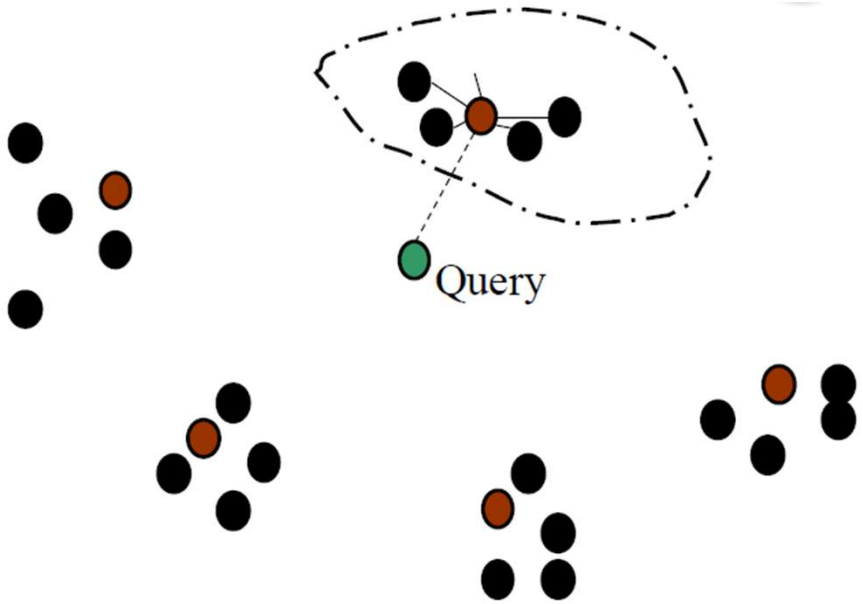
5. نیاز است ویژگی‌ها مقیاس‌بندی شوند تا از غالب شدن یک ویژگی بر معیارهای فاصله جلوگیری شود.

افزایش سرعت KNN

خوشه‌بندی داده‌های آموزشی

- جستجوی نزدیک‌ترین همسایه‌ها به نقطه تست در نزدیک‌ترین خوشه
- مصالحه بین دقت و سرعت:

نزدیک‌ترین **i** خوشه‌ها را در نظر بگیرید.



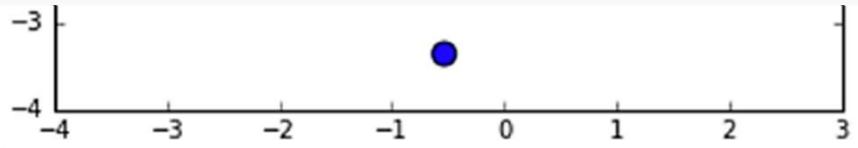
مثال پایتون

```
from sklearn.neighbors import KNeighborsClassifier

clf = KNeighborsClassifier(n_neighbors=1)
clf.fit(X,y)
print('With k=1: ', clf.predict([[-2,-2], [1,1]]))

clf = KNeighborsClassifier(n_neighbors=3)
clf.fit(X,y)
print('With k=3: ', clf.predict([[-2,-2], [1,1]]))
```

```
With k=1:   [1 1]
With k=3:   [0 1]
```



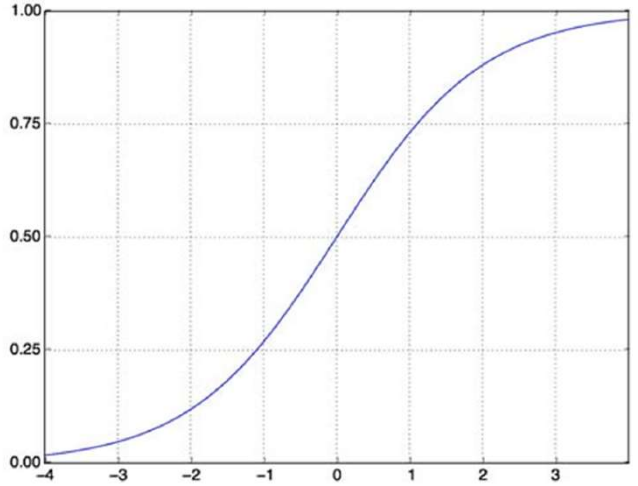
Logistic regression

- احتمال کلاس‌های مختلف را ارائه می‌دهد، به جای اینکه فقط کلاس را پیش‌بینی کند.
- فقط برای ویژگی‌های عددی کار می‌کند.
- ترکیب خطی از ویژگی ها را می سازد.

$$Z = w_0 + w_1X_1 + w_2X_2 + \cdots + w_nX_n$$

- ما تابع سیگموید (لجستیک) را بر روی z اعمال می‌کنیم تا مقداری به‌دست آوریم که بتوان آن را به‌عنوان احتمال تفسیر کرد.

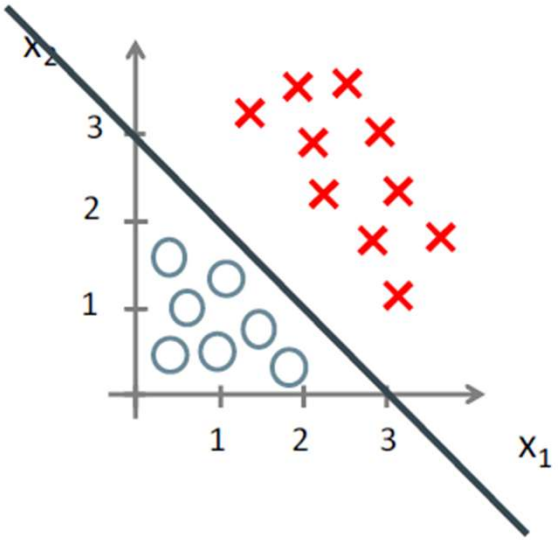
$$\sigma(Z) = \frac{1}{1+e^{-Z}}$$



فرض کنید آستانه عددی مانند ۰.۵ است.

Predict y=1 if $\sigma(z) \geq 0.5$

That is, predict y=1 if $w_0 + w_1 X_1 + w_2 X_2 + \cdots \geq 0$

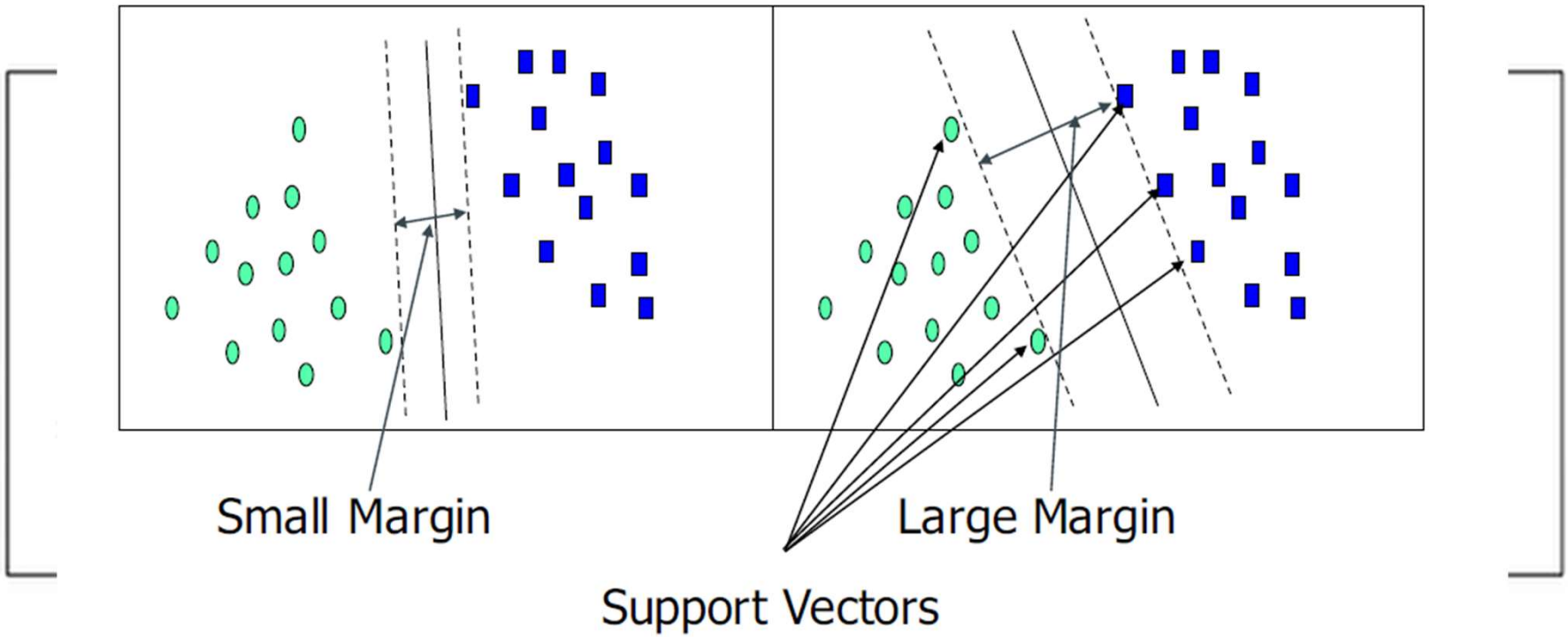


برای مثال پیش بینی می کند که y=1 اگر:

$$-3 + X_1 + X_2 \geq 0$$

یک دسته‌بند خطی است:
دارای یک مرز تصمیم‌گیری خطی است.

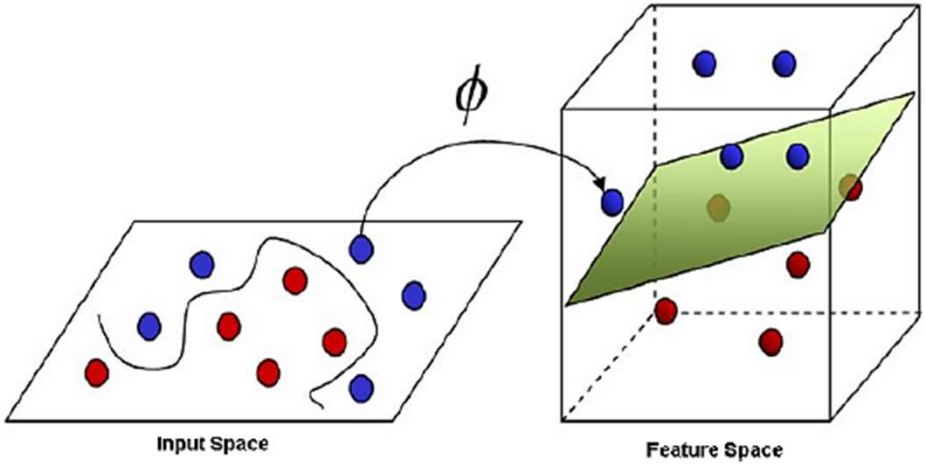
ماشین بردار پشتیبان SVM



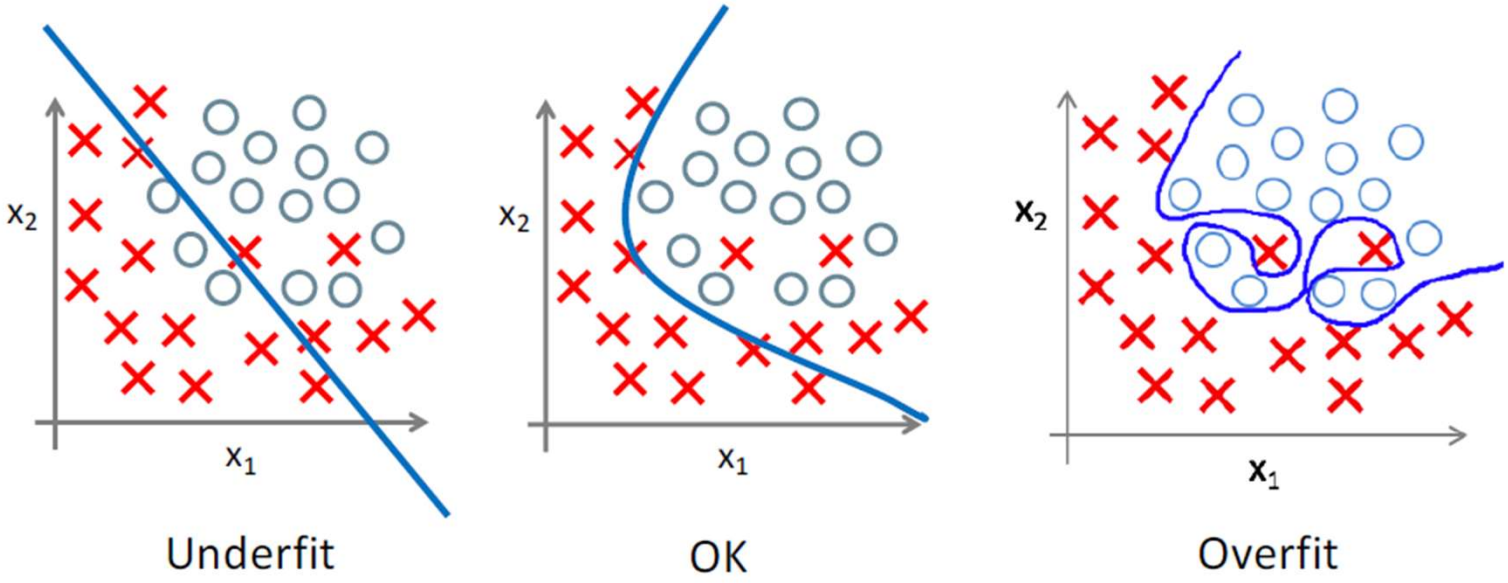
ماشین بردار پشتیبان SVM

با نگاشت غیرخطی مناسب به یک بُعد به اندازه کافی بالا (ترفند کرنل)، داده‌های دو کلاس همیشه می‌توانند با یک اَبَرصَفحه از هم جدا شوند.

در بُعد جدید، SVM به دنبال یافتن مرز تصمیم‌گیری خطی بهینه است.



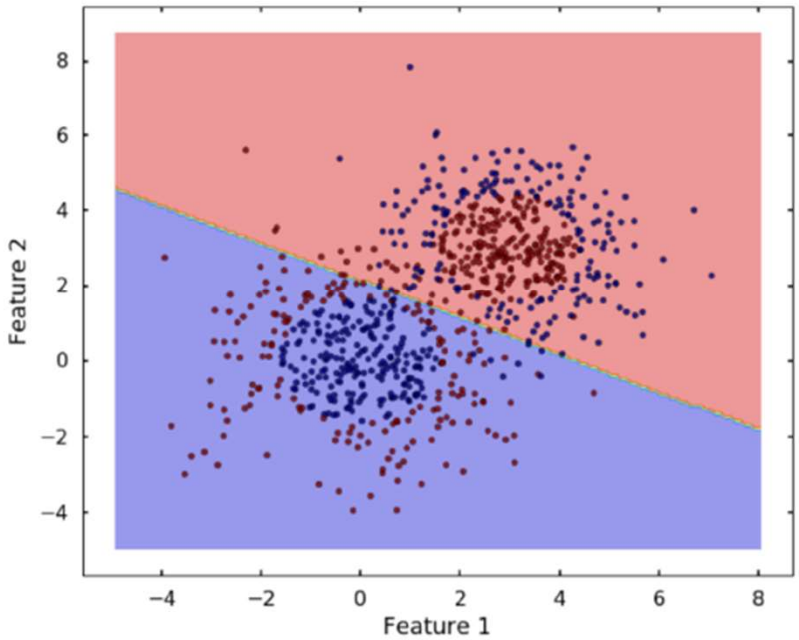
http://thecaffeinedev.com/2-support-vector-machine-learning-math-behind-part2/



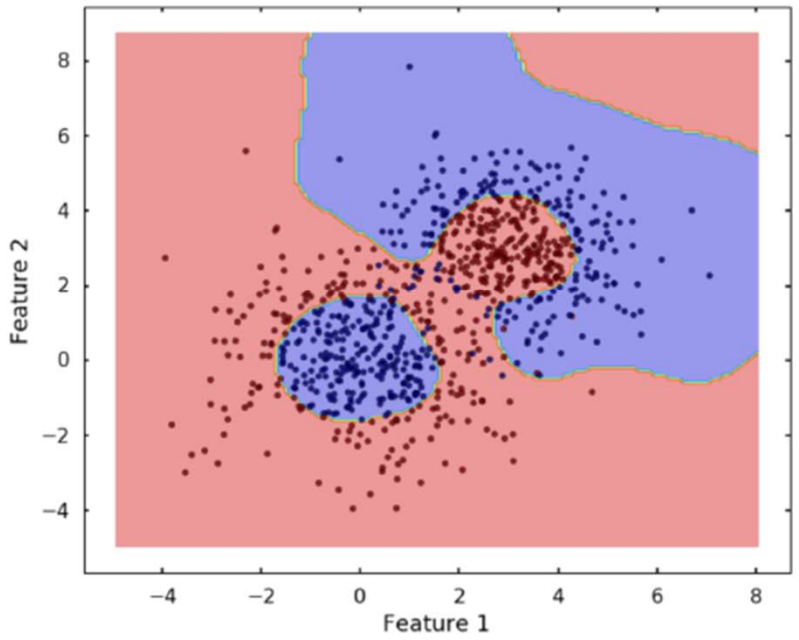
در کتابخانه‌ی sklearn، این مورد توسط پارامتر C در SVM و پارامتر gamma در کرنل rbf کنترل می‌شود.

10

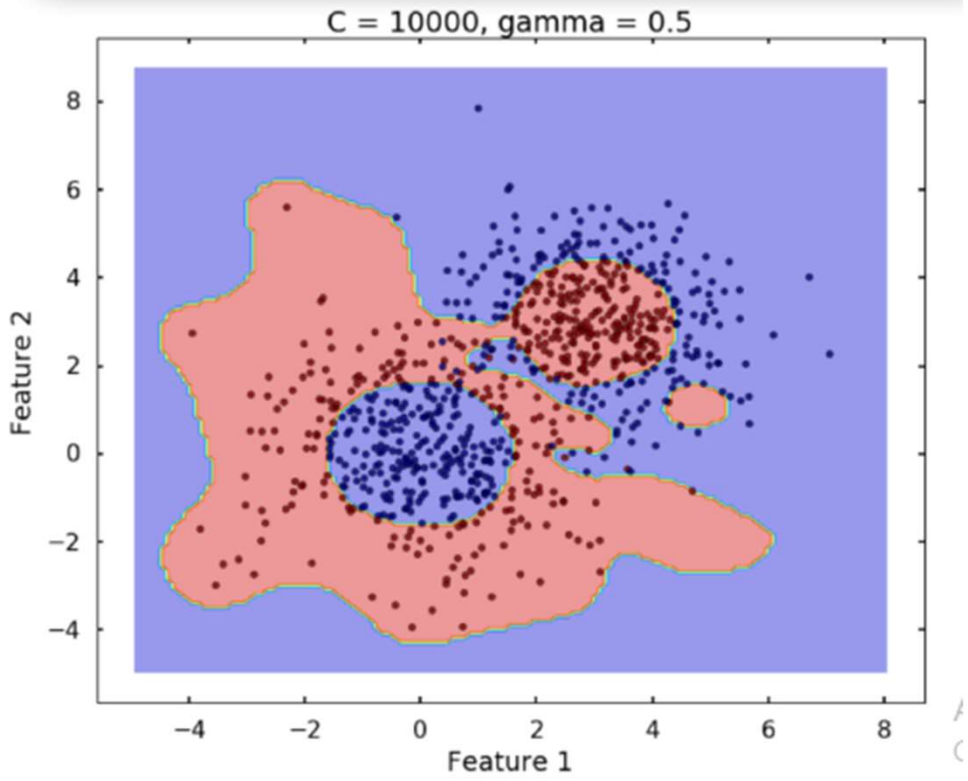
```
clf = svm.SVC(kernel='linear')
clf.fit(X,y)
```




```
clf = svm.SVC(kernel='rbf')
clf.fit(X,y)
```



```
clf = svm.SVC(kernel='rbf', C = 10000, gamma = 0.5)
clf.fit(X,y)
```



مقدار بزرگ برای C هزینه‌ی خطای دسته‌بندی اشتباه را بالا می‌برد.

این باعث می‌شود که الگوریتم برای برازش داده، از مدل انعطاف‌پذیرتری استفاده کند.

اما مقادیر خیلی بزرگ باعث می‌شود که SVM بیشتر مستعد بیش‌برازش overfit داده شود.