Machine Learning

Classification

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

Classification

همانند مسئله رگرسیون است با این تفاوت که y مقدار گسسته می گیرد.

Binary classification (دسته بندی دودویی):

$y \in \{c_1 , c_2\}$, c_i : برچسب کلاس

$y \in \{ -1 , +1\}$, $y \in \{ 0 , +1\}$

هیچ ترتیبی بین دو کلاس وجود ندارد

Multi-class classification (دسته بندی چند کلاسه):

$y \in \{c_0 , c_2 , \dots , c_{k-1}\}$

$y \in \{0 , 1 , 2 , \dots , k-1\}$

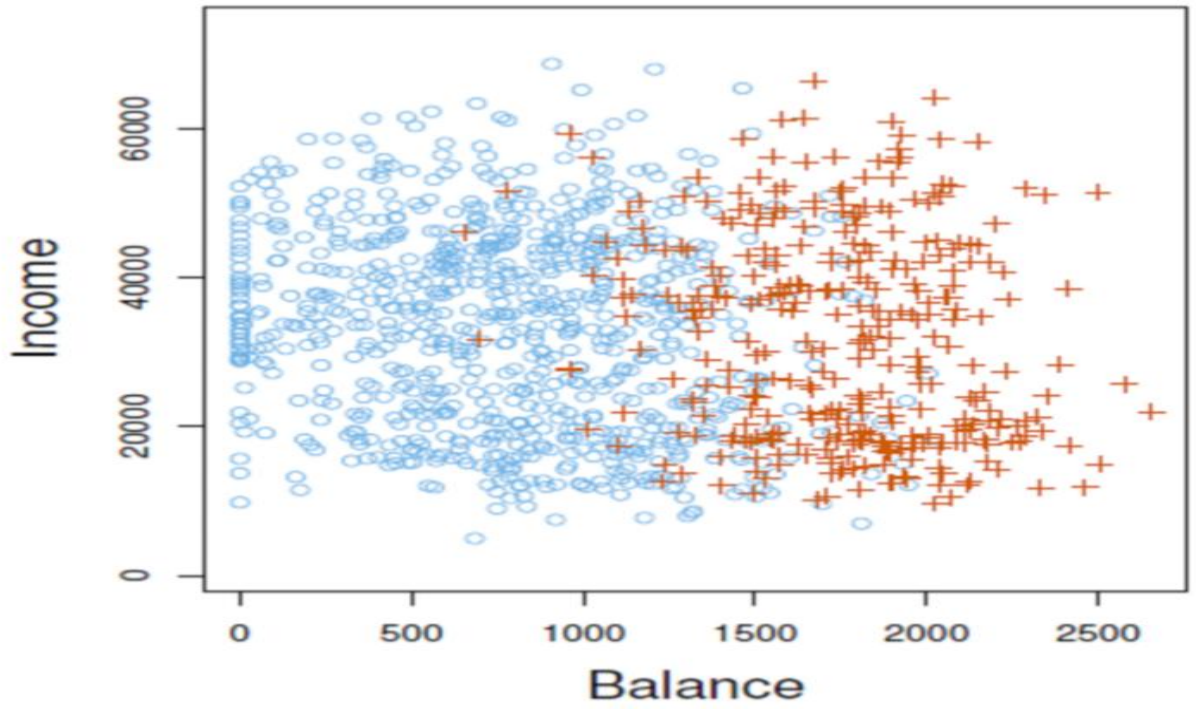
مثال

می خواهیم بدانیم که یک تراکنش کارت اعتباری جعلی است یا اصلی. با استفاده از اطلاعاتی نظیر تراکنش های گذشته و ....

مثالی دیگر:

نارنجی : بدحساب (مبلغ ماهیانه را به موقع پرداخت نمی کند)

آبی: خوش حساب



دسته بند

یک دسته بند فضای ورودی را به ناحیه هایی متعلق به هر کلاس تقسیم میکند. مرز این ناحیه ها را مرز تصمیم (Decision boundry) میگویند.

دسته بند میتواند خطی یا غیر خطی باشد.

دسته بند خطی از ترکیب خطی ویژگی ها استفاده میکند.

$$y = f\left(\sum \theta_j x_j\right)$$

f تابعی است که ترکیب خطی را به خروجی دسته بندی تبدیل میکند.

$$f(\mathrm{x}) = \begin{cases} 1 & \text{if } \theta^{\top} x \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

Decision boundary

$$\theta_1 x_1 + \theta_2 x_2 + b \geq \tau$$ then class 1

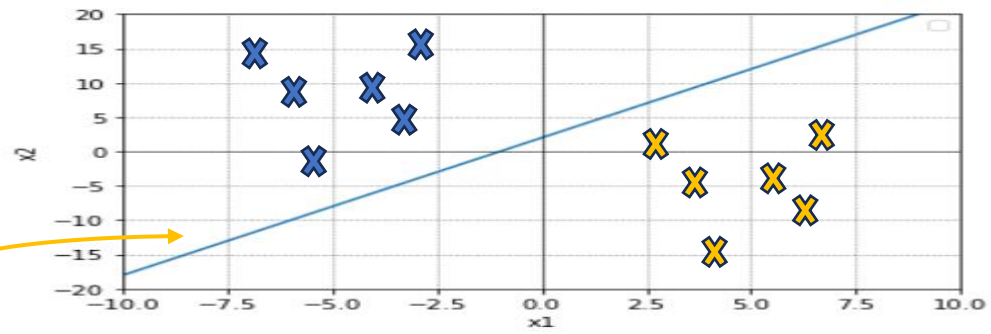
$$\underbrace{b - \tau}_{b'} \geq 0$$

Decision boundary:

$$\theta_1 x_1 + \theta_2 x_2 + b = 0$$

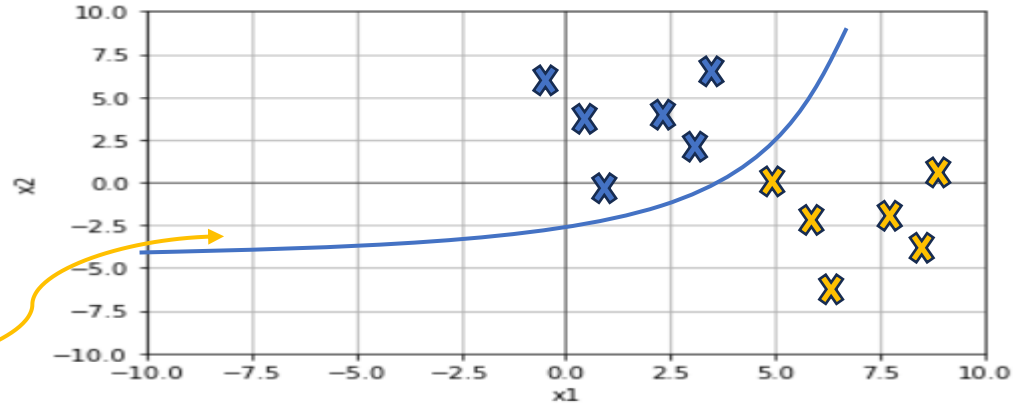
$$x_2 = -\frac{\theta_1}{\theta_2} x_1 - \frac{b}{\theta_2}$$

مرز تصمیم
خطی



$$\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 = 0$$

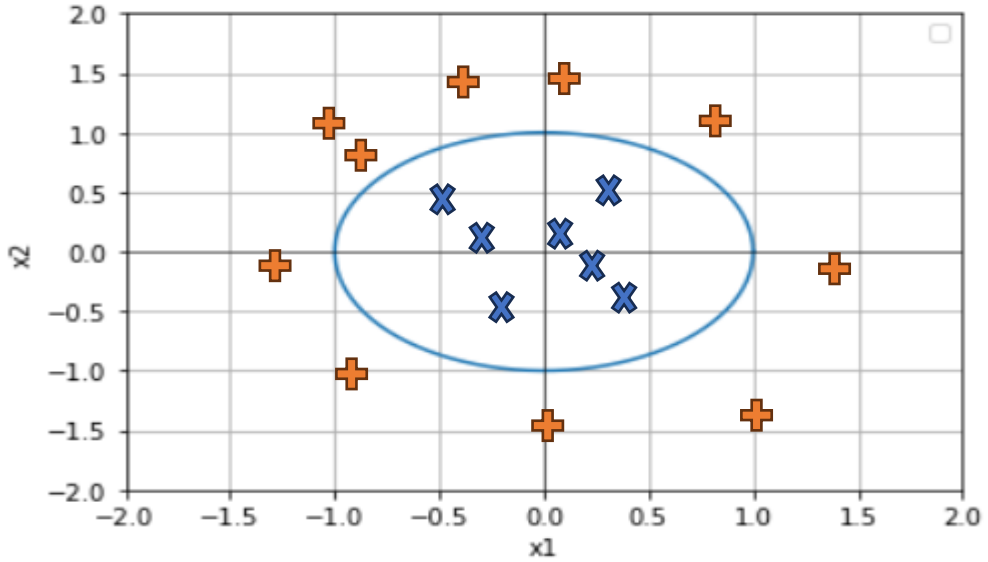
مرز تصمیم
غیرخطی



Decision Boundary

$\theta_1 x_1^2 + \theta_2 x_2^2 +$ b = o , θ_1 =1 , θ_2 =1 , b= -1

$x_1^2 + x_2^2 = 1$



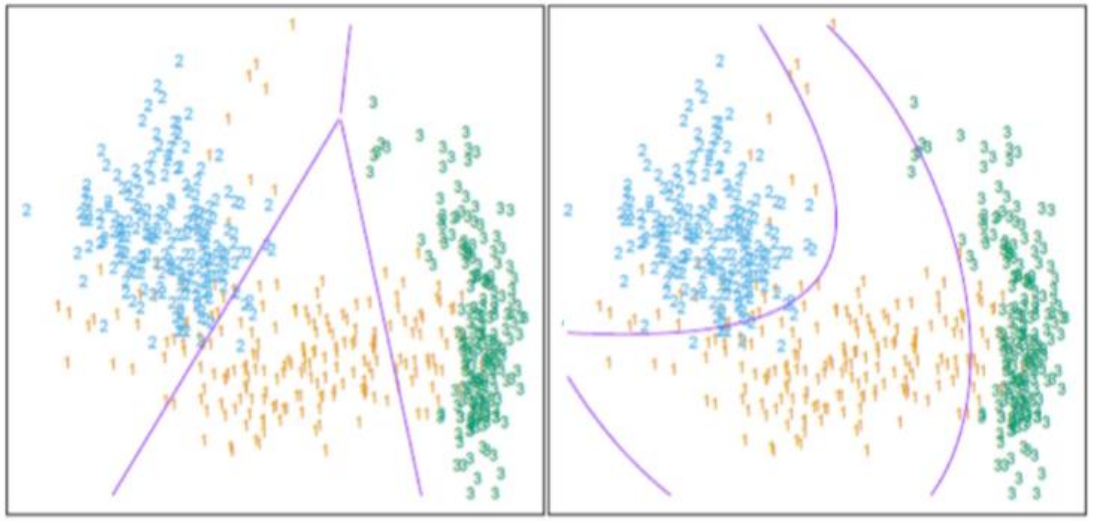


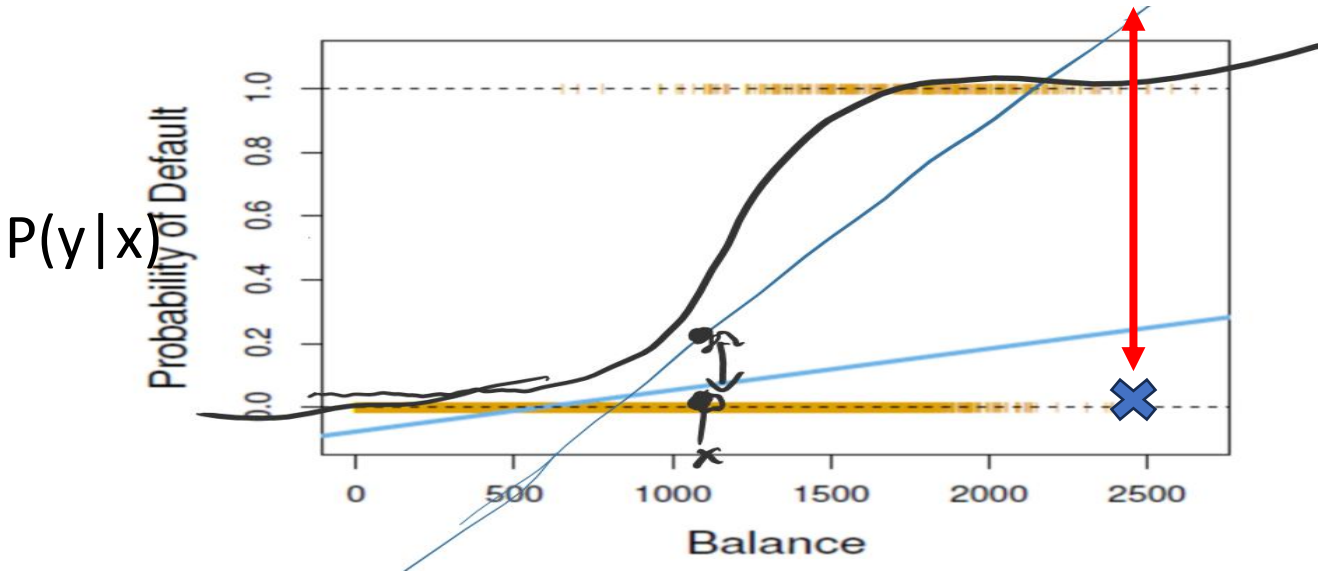
FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Linear inequalities in this space are quadratic inequalities in the original space.*

نکته: مرز تصمیم در دسته بند خطی در فضای ویژگی اولیه (نه فضای ویژگی توسعه یافته) لزوما خطی نیست.

مرز تصمیم در دسته بند خطی در فضای ویژگی توسعه یافته خطی است.

دسته بندی به صورت حالت خاص یک مسئله رگرسیون

- Y=0 , y=1 را میتوان احتمال تعلق به هر یک از کلاس ها در نظر گرفت.
- خروجی مدل میتواند اعداد منفی یا بزرگتر از یک باشد.
- تعداد کمی نقطه میتواند جهت خط را به میزان زیادی تغییر دهد مانند اضافه کردن چند نقطه با بالانس خیلی زیاد



Nearest Neighbor

دسته بند نزدیکترین همسایگی

- نمونه های نزدیک برچسب یکسانی دارند.
- نزدیکی را با فاصله اقلیدسی میتوان اندازه گرفت.
- با داشتن یک محموعه آموزشی s_{train} و یک داده x به دنبال x^* که نزدیکترین نقطه به x است میگردیم و برچسب را برابر $y^*(x^* \text{ برچسب})$ با قرار میدهیم.

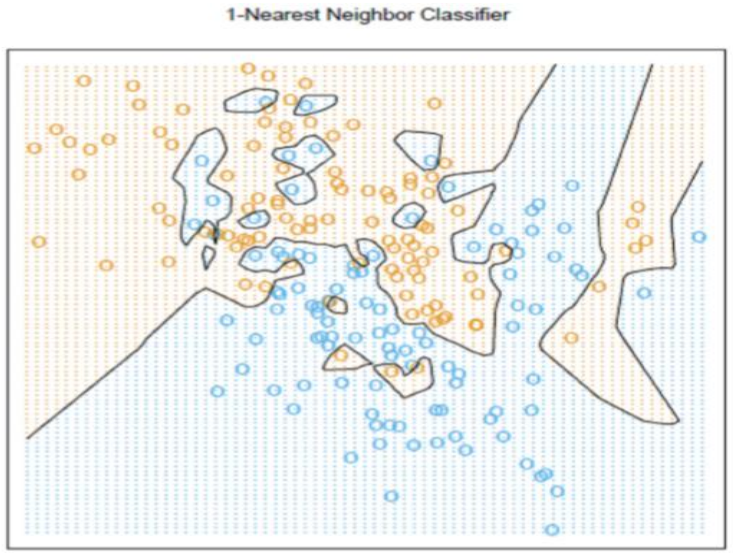


FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE = 0, ORANGE = 1), *and then predicted by* 1*-nearest-neighbor classification.

دسته بند بهینه بیز

فرض کنید x داده و y برچسب است. اگر P(y|x) را به صورت دقیق بدانیم چگونه دسته بندی میکنیم؟؟

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\, P(Y = y \mid X = x)$$

(دسته بند بهینه بیز)

Bayes rule:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

likelihood

prior

posterior

normalizer

چرا یادگیری دسته بند بهینه مشکل است؟؟

Sky	Temp	Humid	Wind	Water	Forecast	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cold	Change	Yes

فرض کنید k کلاس داریم. فرض کنید x شامل d ویژگی باینری است.

prior: $P(Y)$, likelihood: $P(X \mid Y)$

توزیع بر روی مقادیر ویژگی ها:

$$P(X = x \mid Y = y)$$

$$P(S = s, T = w, H = n, W = S, wa = w, F = s \mid E = y)$$

sky	temp	P(sky, temp|enjoy sport=yes)
sunny	warm	0.1
sunny	cold	0.3
rainy	warm	0.5
rainy	cold	0.1

sky	temp	P(sky, temp|enjoy sport=no)
sunny	warm	0.2
sunny	cold	0.1
rainy	warm	0.4
rainy	cold	0.3

تعداد پارامترهای مدل:

$$k(2^d - 1)$$

تعداد زیاد پارامتر که نیاز به تعداد زیاد داده برای تخمین دارد

Conditional independence(استقلال شرطی)

$Recall:\ X \perp\!\!\!\perp Y \quad \Rightarrow p(x,y) = p(x)p(y) \qquad p(x|y) = p(x) \quad p(y|x) = p(y)$

$X \not\perp\!\!\!\perp Y$ X and Y are not independent

$X \perp\!\!\!\perp Y|Z$ X and Y are independent if we know Z

$\forall i,j,k \quad p(X = i \mid Y = j, z = k) = p(x = i \mid z = k)$

مثال:

نمره سواد $\not\perp\!\!\!\perp$ اندازه کفش

نمره سواد $\perp\!\!\!\perp$ اندازه کفش اگر سن فرد را بدانیم

Conditional independence(استقلال شرطی)

$$P(\text{thunder} \mid \text{Rain, Lightening}) = P(\text{thunder} \mid \text{Lightening})$$

$$R \not\perp\!\!\!\perp T \quad , \quad T \perp\!\!\!\perp R \mid L$$

$$P(T, R \mid L) = P(T \mid R, L)P(R \mid L) = \frac{P(T,R,L)}{P(L)} = \frac{P(T,R,L)}{P(R,L)}\frac{P(R,L)}{P(L)} = P(T \mid L)P(R \mid L)$$

بافرض استقلال شرطی

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

اگر بخواهیم میزان بیشترین احتمال A به شرط B را بدست آوریم

$$P(A \mid B) \propto P(B \mid A)P(A)$$

IF P(A=0) = P(A=1) $\quad$ p(A|B) ∝ p(B|A)

احتمال Lightening چقدر است؟؟

$P(L|T,R) \propto P(T,R|L) * P(L)$ ← یکنواخت

Rain	Thunder	P(Rain, Thunder| Lightening=1)
0	0	p(R=0,T=0|L=1)
0	1	p(R=0,T=1|L=1)
1	0	p(R=1,T=0|L=1)
1	1	p(R=1,T=1|L=1)

Rain	Thunder	P(Rain, Thunder| Lightening=0)
0	0	p(R=0,T=0|L=0)
0	1	p(R=0,T=1|L=0)
1	0	p(R=1,T=0|L=0)
1	1	p(R=1,T=1|L=0)

Estimate: P(T, R|L) $2(2^2 - 1) = 6$ params

تعداد پارامترهای مدل در حالت کلی:

$$k(2^d - 1)$$

تعداد زیاد پارامتر که نیاز به تعداد زیاد داده برای تخمین دارد

احتمال Lightening چقدر است؟؟

با فرض استقلال شرطی:

$$P(L \mid T, R) \propto P(T, R \mid L) * P(L)$$ ← یکنواخت

$$P(T, R \mid L) = P(T \mid L)P(R \mid L)$$

Thunder	P(Thunder| Lightening=1)
0	p(R=0|L=1)
1	p(R=1|L=1)

Thunder	P(Thunder| Lightening=0)
0	p(R=0|L=0)
1	p(R=1|L=0)

Rain	P(Rain| Lightening=1)
0	p(R=0|L=1)
1	p(R=1|L=1)

Rain	P(Rain| Lightening=0)
0	p(R=0|L=0)
1	p(R=1|L=0)

Estimate: P(T, R|L) 2(2- 1)+ 2(2- 1)= 4 params

تعداد پارامترهای مدل در حالت کلی:

$$k(2^d - 1)$$

تعداد زیاد پارامتر که نیاز به تعداد زیاد داده برای تخمین دارد

تعداد پارامترهای مدل در حالت استقلال شرطی:

$$k(2-1)d = kd$$

Naive Bayes assumption

ویژگی ها به شرط دانستن کلاس مستقل هستند:

$$P(x_1, x_2 \mid Y) = p(x_1 \mid x_2, Y)P(x_2 \mid Y) = P(x_1 \mid Y)p(x_2 \mid Y)$$

به طور کلی:

$$P(x_1, \dots x_d \mid Y) = \prod_{j=1}^{d} P(x_j \mid Y)$$

حال چند پارامتر خواهیم داشت؟

بدون فرض استقلال شرطی:

$$K(2^d - 1)$$

K: تعداد کلاس
2: تعداد حالت
d: تعداد ویژگی ها

With Naive bayes:
K(d)

کاهش قابل ملاحظه پارامترها. شاید حتی زیادی کم شد.

تعریف دسته بند Naive bayes

Prior: P(y) , d: ویژگی مستقل به شرط کلاس , برای هر x_j : $P(x_j \mid y)$

قانون تصمیم گیری:

$$\hat{y} = f_{NB}(x) = \underset{y}{\operatorname{argmax}} P(y) \prod_J P(x_j \mid y)$$

$$\text{prior: } P(Y = y) = \frac{\text{count}(Y = y)}{N}$$

$$\text{likelihood: } P(X_j = x_j \mid Y = y) = \frac{\text{count}(X = x_j, Y = y)/N}{\text{count}(Y = y)/N}$$

$$P(X \mid Y) = \frac{P(x, y)}{P(y)}$$

یک مشکل ممکن است رخ دهد:

$P(x_1 \mid Y=b) = 0$

آنگاه:

$$P(Y = b \mid X) \quad = P(Y = b \mid X_1)P(Y = b \mid X_2) \cdots P(Y \mid X_d) = 0$$

Smoothing: راه حل

$$P(Y = b) = \frac{\text{count}(Y = b) + \lambda}{N + k\lambda}$$

K: تعداد کلاس
λ: 1

$$P(\text{x1} = a \mid y = b) = \frac{\text{count}(x_1 = a, y = b) + \lambda}{\text{count}(y = b) + A\lambda}$$

A: x1 تعداد حالات ویژگی
λ: 1

Example : Spam classification

- Email 1: "Win a million dollars now!" (Spam)
- Email 2: "Meeting at 10 am" (Not Spam)
- Email 3: "Claim your prize now!" (Spam)
- Email 4: "Schedule change for the meeting" (Not Spam)
- To classify a new email, "Win a prize now!" we calculate the probability of i t being spam or not spam based on the words it contains.

1. Calculate the prior probability for each class (Spam and Not Spam) based o n the training data:
    1. P(Spam) = Number of spam emails / Total number of emails = 2/4 = 0.5
    2. P(Not Spam) = Number of not spam emails / Total number of emails = 2/4 = 0.5

Example: Spam classification

- Email 1: "Win a million dollars now!" (Spam)
- Email 2: "Meeting at 10 am" (Not Spam)
- Email 3: "Claim your prize now!" (Spam)
- Email 4: "Schedule change for the meeting now" (Not Spam)

Win a prize now!

Assume that dictionary size is 10000

Calculate the likelihood for each word in the new email:

1. P("Win" | Spam) = Number of times "Win" appears in spam emails / Total number of words in spam emails
2. P("Win" | Not Spam) = Number of times "Win" appears in not spam emails / Total number of words in not spam em

	P(word|spam)	P(word|not spam)
win	1/11	0/10 ~ 1/10010
a	1/11	0/10~ 1/10010
prize	1/11	0/10~ 1/10010
now	2/11	1/10
!	2/11	0/10~ 1/10010

$$p(spam|win\ a\ prize\ now!) = \left(\frac{1}{11}\right)^3 \left(\frac{2}{11}\right)^2 \frac{1}{2}$$

$$p(not\ spam|win\ a\ prize\ now!) = \left(\frac{1}{10010}\right)^4 \left(\frac{1}{10}\right) \frac{1}{2}$$

داده های پیوسته

- گسسته سازی
- استفاده از توزیع های پیوسته

$$P(x_1, x_2, \dots x_n \mid c_k) \propto P(c_k) \cdot \prod_{i=1}^{n} P(x_i \mid c_k)$$

$$P(x_i \mid c_k) = N(\mu_i, \sigma_i)$$

تخمین به وسیله روش ML