Machine Learning

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

Dr Mehran Safayani

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):

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

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

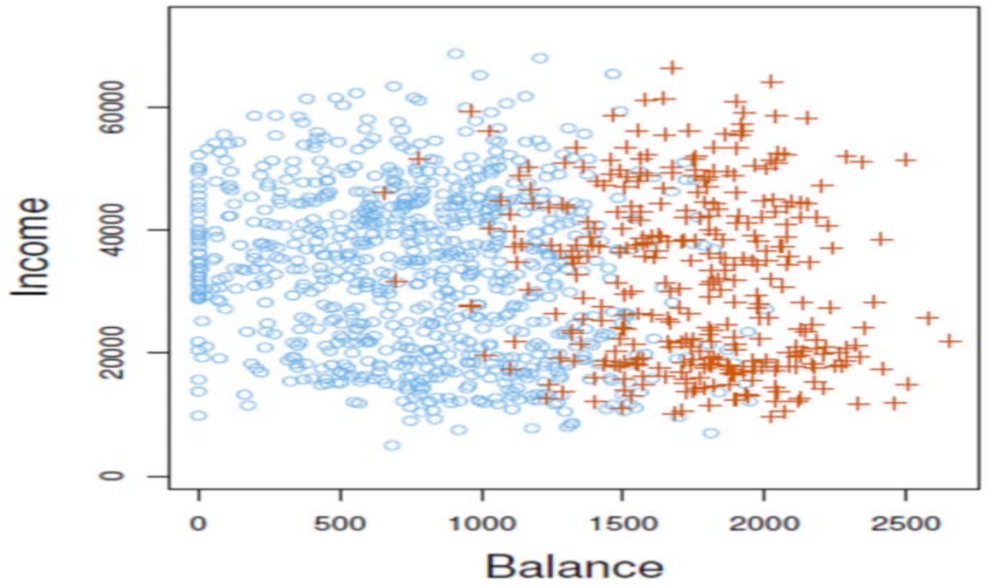
مثال

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.

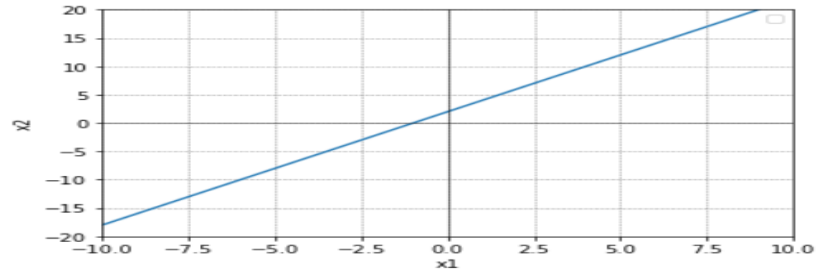
$$y = f\left(\sum \omega_j x_j\right)$$

f تابعی است که ترکیب خطی را به ویژگی ها تبدیل میکند.

$$f(\mathrm{x}) = \begin{cases} 1 & \text{if } \omega^{\mathsf{T}} x > \theta \\ 0 & \text{otherwise} \end{cases}$$

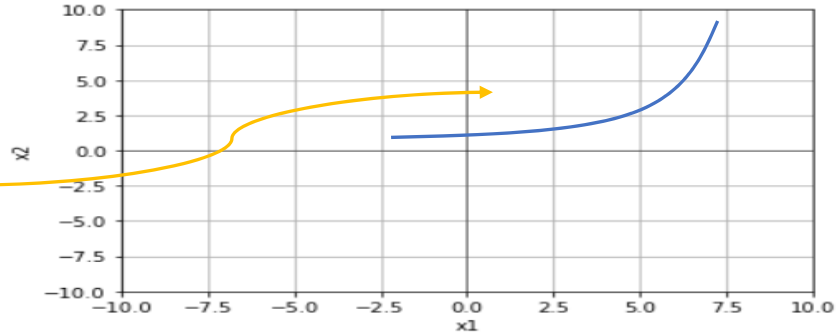
$$w_1 x_1 + w_2 x_2 + b = 0$$
$$w_1 x_1 + w_2 x_2 + b = \theta$$
$$\underbrace{b - \theta}_{b'} = 0$$

$$w_1 x_1 + w_2 x_2 + b = 0$$
$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2}$$



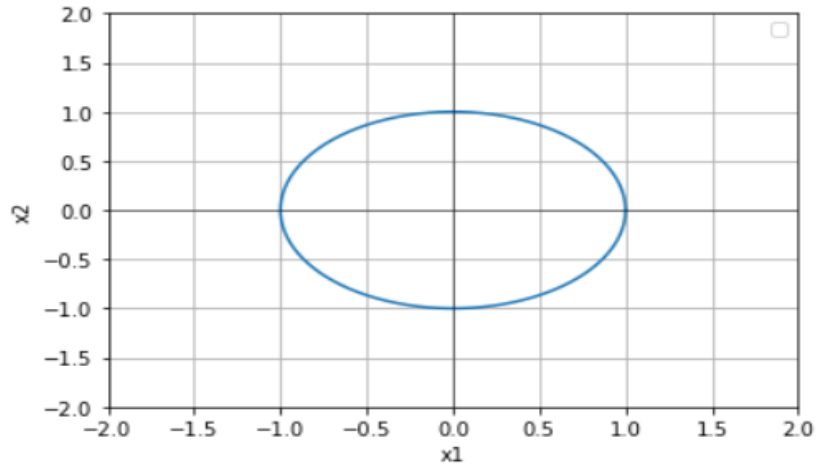
$$w_1 x_1 + w_2 x_2 + w_3 x_1^2 = 0$$

مرز تصمیم



$w_1\, x_1^2 + w_2\, x_2^2 +$ b = 0 , w_1 =1 , w_2 =1 , b= -1

$x_1^2 + x_2^2 = 1$



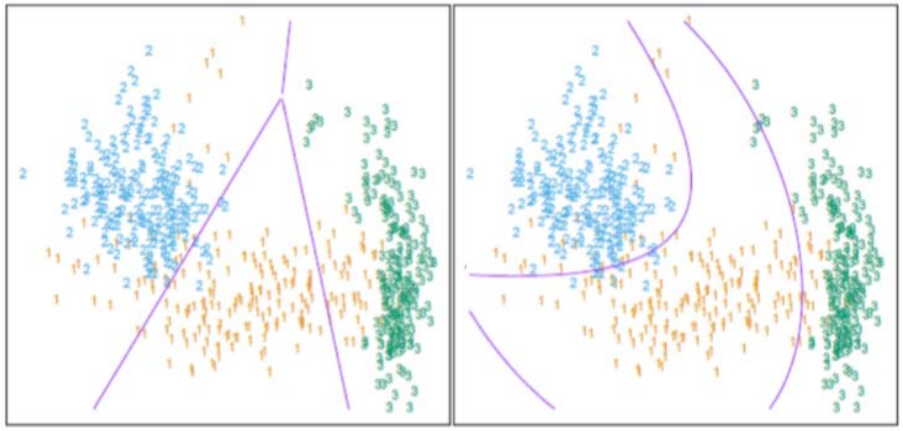
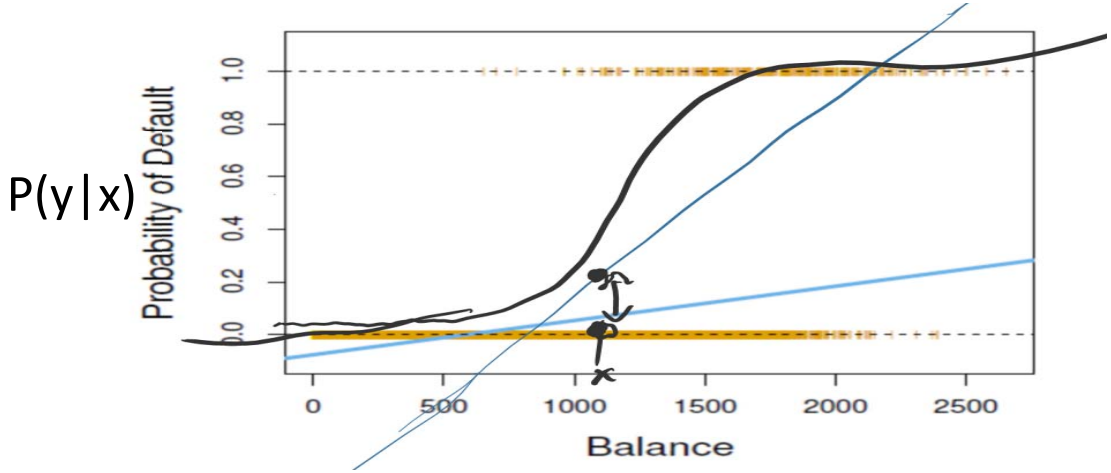


FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Linear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

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

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
- SE یک تابع هزینه مناسب برای هدف دسته بندی نیست
- در مسئله دسته بندی بایستی تعداد نقاط با دسته بندی اشتباه حداقل شود.
- SE فاصله برچسب ها تا پیش بینی را حداقل میکند.
- اگر MSE خیلی کم باشد میتوان تضمین کرد که خطای دسته بندی هم کم است ولی برعکسش صحیح نیست یعنی یک تابع رگرسیون ممکن است خطای زیادی داشته باشد هر چند تعداد داده های با خطای دسته بندی کم باشد.



Nearest Neighbor

دسته بند نزدیکترین همسایگی

- نمونه های نزدیک برچسب یکسانی دارند.
- نزدیکی را با فاصله اقلیدوسی 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^* (برچسب x^*) با قرار میدهیم.

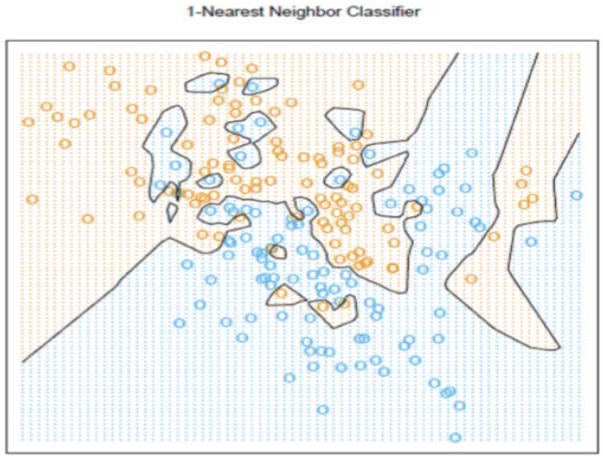
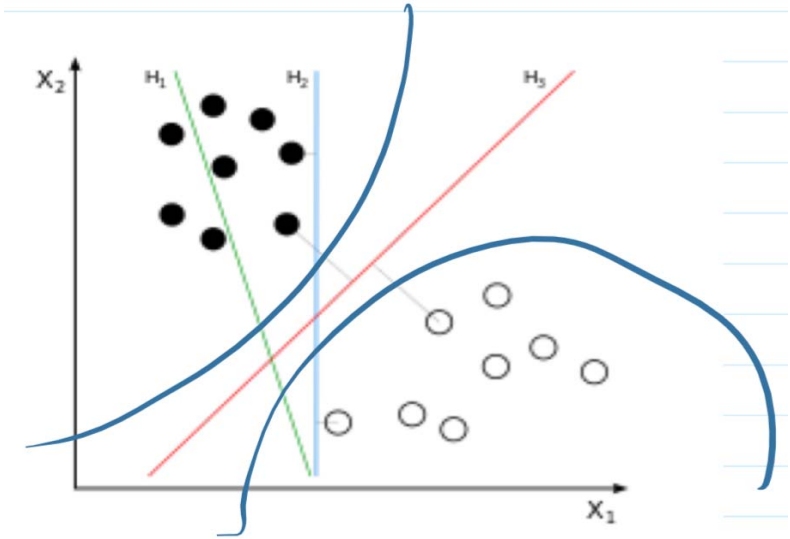


FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$), *and then predicted by* 1*-nearest-neighbor classification.*

مرز های تصمیم خطی

Linear decision boundaries

- فرض کنید که مرز تصمیم خطی است. hyperlinear
- فرض کنید که مسئله تفکیک پذیر خطی است.
- ایده ماشین بردار پشتیبان Support vector machine(svm)



مرز تصمیم غیر خطی

- استفاده از کرنل

- ویژگی های quadratic

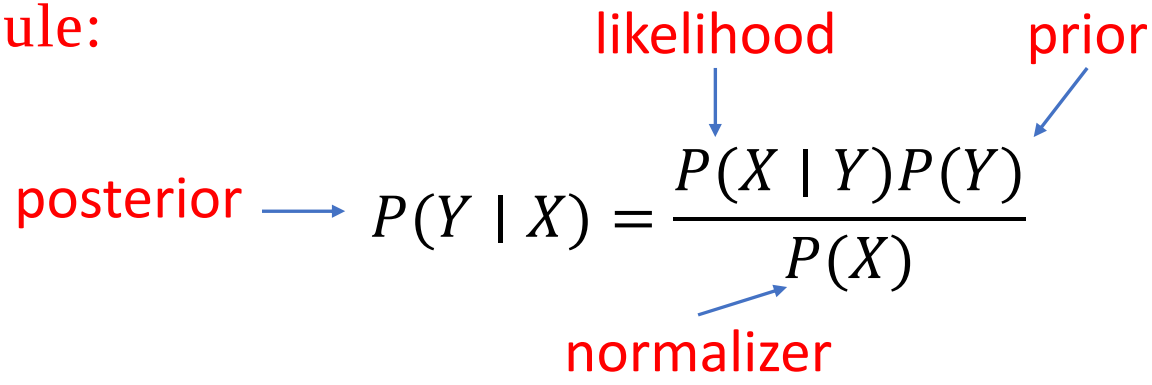
دسته بند بهینه بیز

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:



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:

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

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

تعداد پارامترهای مدل:

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

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

Conditional independence(استقلال شرطی)

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

x به شرط دانستن z از y مستقل است.

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

$T \perp R$ ✗ , $T \perp R \mid L$ ✓

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

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

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

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

چرا مخرج حذف شده است؟؟

IF P(A=0) = P(A=1) ⟹ p(A|B) ∝ p(B|A)

احتمال lightening چقدر است؟؟

P(L| T , R) ∝ P(T, R|L) * P(L) ← یکنواخت

P(T, R|L = 0): P(0,0|L=0) , P(0,1|L=0) , P(1,0|L=0) , P(1,1|L=0)

P(T, R|L = 1): P(0,0|L=1) , P(0,1|L=1) , P(1,0|L=1) , P(1,1|L=1)

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

با فرض استقلال شرطی:

$P(T , R \mid L) = P(T \mid L)\, P(R \mid L)$ $2(2 - 1) + 2(2 - 1) = 4$ params

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؟

اگر فرض استقلال شرطی داشته باشیم:

$K(2^d - 1)$

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 | y)

قانون تصمیم گیری:

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

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

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

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

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

مثال

Text classification / classify email {spam , not spam} / classify news article topic

هر مقاله حداقل از 1000 کلمه تشکیل شده و هر کلمه میتواند 10000 حالت داشته باشد.

$P(x|y)$ بسیار بزرگ است.

تعداد حالات:

$$\propto 2(10000)^{1000} = 2(10)^{4000}$$

With NB:

$$\propto 2^* 10000^*1000 = 2(10)^7$$

- With NB:

- $f_{NB}(x) = \underset{y}{\operatorname{argmax}} P(Y = y) \prod_{y=1}^{\text{lengthdoc}} P(x_j \mid \mathrm{Y} = y)$

تعداد کل منابع

- $P(x_j \mid \mathrm{Y} = y) = \dfrac{\frac{count(x_j = \mathrm{x}, \mathrm{Y} = y)}{N}}{P(Y = y) = \frac{count(\mathrm{Y} = y)}{N}}$

کل منابع در پایگاه داده

$$\frac{\text{count}(X = xj\,, Y = y)}{\text{count}(Y = y)}$$

تعداد منابع در تاپیک Y

Bag of words(BOW)

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