/ /

ML-9

Generalized linear models (GLM)

+ $X \rightarrow Y$ (mapping $\rightarrow$ classification or regression)
+ we studied $\rightarrow$ linear regression $Y|x,\beta \sim N(\mu, \sigma^2)$
  $\searrow$ logistic regression $Y|x,\beta \sim Bernoulli(\theta)$

درحالت logistic چون جواب صفر و یک بود از تابع sigmoid استفاده کردیم، اما چرا از sigmoid؟

در واقع تابع sigmoid از تابع های Glm به دست آمده است و در قالب تابع Glm قرار می گیرد.

For Normal distribution

$$E(Y|x;\beta) = \mu = g(\underbrace{\beta^T x}_{\text{linear predictor}})$$

فرض کنید روی $\beta^T x$ تابع g را اعمال می کنیم که در حالت توزیع نرمال $g(z)=z$ است

$\downarrow$ تابع identity

For logistic regression

$$Y|x,\beta \sim Bern(\theta)$$

$\Downarrow$ وابسته به x و ضرایب است.

$$E(Y|x;\beta) = \theta = g(\beta^T x) \qquad g(z) = \frac{1}{1+e^{-z}}$$

$\downarrow$ $1\times\theta + 0\times(1-\theta) = \theta$

g: response function → از $\beta^T x$ به میانگین یا average متغیر y می‌رسد.

g^{-1}: link function → میانگین را به $\beta^T x$ ، map می‌کند.

Ex: logistic regression

$$g^{-1}(\theta) = \beta^T x \rightarrow \log \frac{\theta}{1-\theta} = \beta^T x$$

نکته‌ای که وجود دارد این است که glm درصورتی یک framework به جا می‌دهد که $Y|X,\beta$ از distribution exponential family تبعیت کند.

* Glm : given $Y|X,\beta$ is form an exp family, Glm framework assists us:

1) find g
2) estimate β (parameters)

Ex. build a Glm for:

+ response variable (Y): # reads of a gene. → صفر و یک نیست و اعداد integer است.
+ input variables (X): gender, treatment/control group

در اینجا مدل‌های linear و logistic کمک کننده نیست ولی Glm سودمند است.

Y: count data $\Rightarrow$ which distribution $Y|x;\beta$?

بهتر است Negative binomal باشد (یا Beta binomial)

چون پواسن فقط یک متغیر دارد و واریانس و میانگین به هم وابسته‌اند مناسب نیست.

Three elements of Glm:

1) $Y|x;\beta$ follows exp family

$$\boxed{p(y|x,\beta)=\frac{1}{Z(\theta)}h(y)\exp(t(\theta)s(y))} \rightarrow \text{θ به x و β وابسته است.}$$

$\boxed{s(y)=y}$ → فرض می‌کنیم $s(y)=y$ است چون فرمولها را ساده‌تر می‌کند.

for simplicity

$\boxed{t(\theta)}$ → Natural parameter

2) we assume $\underbrace{t(\theta)=\beta^T x}_{\text{design choice}}$ → linear predictor

3) $F(x)=E(Y|x,\beta)=g(\beta^T x)$

سعی به ترکیب شدن به E دارد.

generalized

Ex: Bernoulli distribution

$Y|x,\beta \sim Bern(\theta)$

1) $F(x) = E[Y \mid x, \beta] = \theta \rightarrow$ θ به β و x وابسته است

2) $t(\theta) = \beta^T x = \boxed{\log \frac{\theta}{1-\theta} = \beta^T x} \rightarrow \boxed{\theta = \frac{1}{1+e^{-\beta^T x}}} \rightarrow \boxed{g(z) = \frac{1}{1+e^{-z}}}$

Ex : Normal distribution

$$Y \mid x, \beta \sim N(\mu, 1)$$

1) $f(x) = E[Y \mid x, \beta] = \mu$

2) $t(\mu) = \mu = \beta^T x \longrightarrow \boxed{g(z) = z}$

Ex : poisson distribution

$$\text{poisson}(\lambda) \longrightarrow P(y \mid x, \beta) = \frac{\lambda^{Y} e^{-\lambda}}{y!}$$

1) $f(x) = E[Y \mid x, \beta] = \lambda$

2) $\log \lambda = \beta^T x \longrightarrow \lambda = e^{\beta^T x} \Longrightarrow f(x) = e^{\beta^T x}$

تخمین β برای پواسن : $L(\beta)=\prod_{i=1}^{n} p(y[i]|x[i],\beta) = \prod_{i=1}^{n} \frac{\lambda^{y[i]} e^{-\lambda}}{y[i]!}$

$$= \prod_{i=1}^{n} \frac{(\exp(\beta^T x[i]))^{y[i]} \exp(-\exp(\beta^T x[i]))}{y[i]!} \longrightarrow \text{likelihood}$$ فرم

log بگیرید، gradiant و hessian حساب کنید.

stochastic پواسن رگرسیون کرده و حل کنید ← glm ← family = poisson

- از آنجایی که logistic regression برای مسائل دوکلاسی استفاده می‌شود، تعمیمی از آن به نام Multinomial regression (softmax) برای حالات چندکلاسی استفاده می‌شود و توزیع آن به جای binomial، multinomial است.

- Multivariate Normal distribution

+ also called the multivariate gaussian distribution.

+ univariate $x \sim N(\mu, \sigma^2)$

$$\text{PDF} \rightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

+ multivariate $x \sim N(\mu, \Sigma)$ $\quad \mu \in R^d, \Sigma \in IR^{d\times d}$

Σ ↓ covariance matrix ↓ $\underline{\text{Symmetric, SPD}}$

/ /

$$\text{PDF} \rightarrow P(x;\mu,\Sigma) = \frac{1}{2\pi^{d/2}|\Sigma|^{\frac{1}{2}}} \exp\Big(\overbrace{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)}^{\text{quadratic form}}\Big)$$

$E(x) = \mu$

$\text{cov}(x) = E\big((x - E(x))(x - E(x))^T\big)$

اثبات کنید

$$\boxed{\text{Cov}(x) = E(x \cdot x^T) - E(x)\,E(x)^T}$$ HW

ref: CS229

در شکل زیر، 3 نمودار مربوط به PDF 3 تا توزیع نرمال دو متغیره است. در هر سه $\mu = \begin{bmatrix}0\\0\end{bmatrix}$ است.

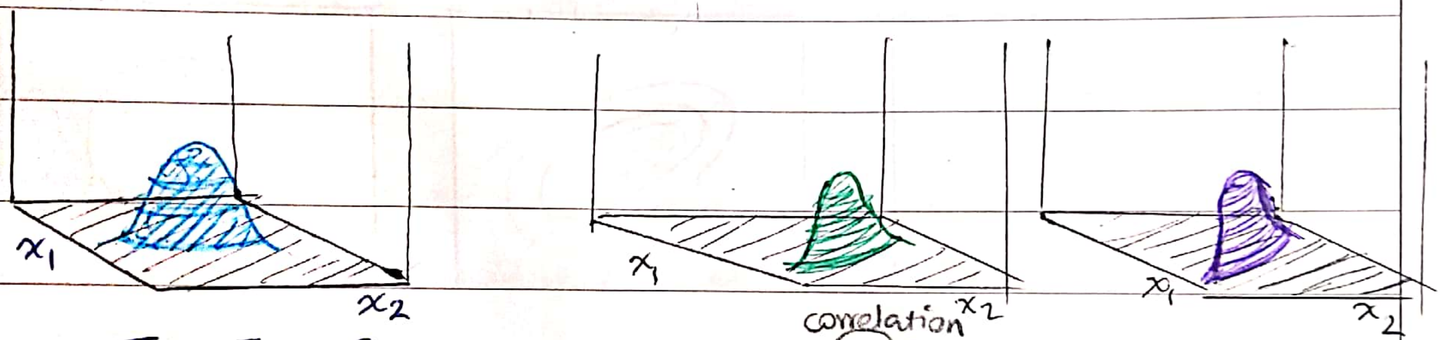
و کوواریانس یک ماتریس قطری است.

$$\Sigma = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$ → عناصر قطری واریانس را نشان می دهند.

⇓

در واقع ارتباط x_1 و x_2 را نشان می دهند اگر صفر باشد یعنی x_1 و x_2 مستقل اند.

اگر + باشد 〃 x_1 و x_2 ارتباط مستقیم دارند.

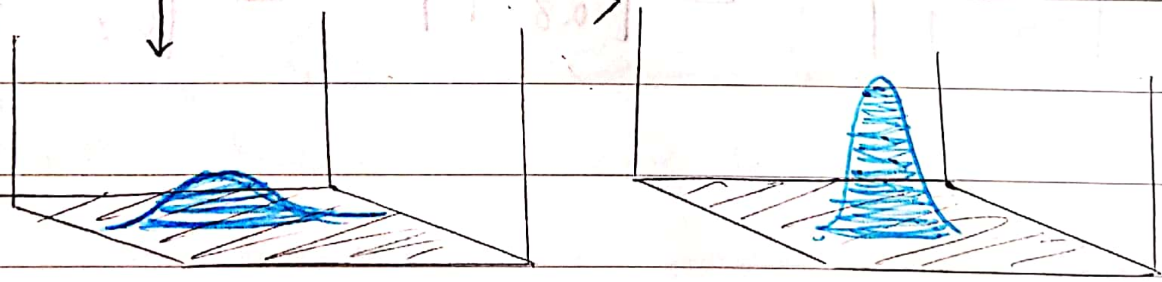


$$\Sigma = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = I$$

correlation

$$\Sigma = \begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}$$

$$\Sigma = \begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix}$$



$$\Sigma = \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix} = 2I$$

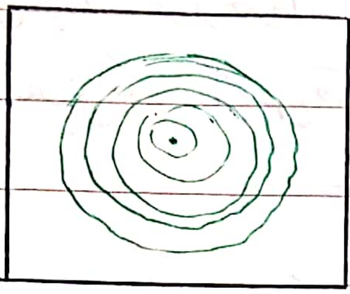
$$\Sigma = \begin{bmatrix}0.6 & 0\\ 0 & 0.6\end{bmatrix} = 0.6I$$

پراکندگی بیشتر شده است.

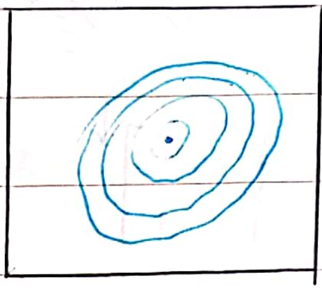
پراکندگی کمتر شده است.

correlation ← لزومی ندارد بین صفر و یک باشد، اما + و منفی بودن آن معنی دار است.

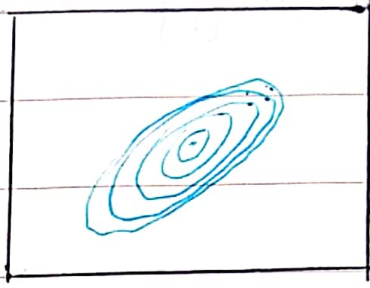
Contour plot :



$$\Sigma = I$$



$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

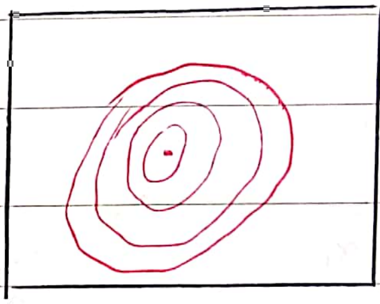


$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

anticorrelation

$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



$Var(x_1)$

$$\Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Symmetric

$Var(x_2)$

Discriminative models

+ based on conditional probability → تاکنون بیشتر اینگونه مدلها را بررسی کرده‌ایم.
  $P(Y|X)$

+ logistic regression, KNN, SVM, ... → a functional form for $P(Y|X)$

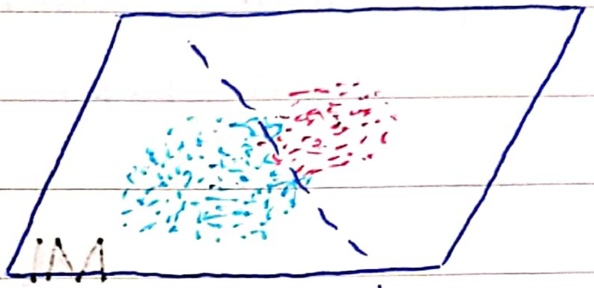
⇒ estimate its parameters

Ex: $P(Y|X) = f(x) = \dfrac{1}{1+e^{-\beta^T x}}$

Scanned with CamScanner

Generative models :

+ a statistical model for $\boxed{P(x,y)}$ joint



در مدلهایی که براساس Discriminative است، هدف این است که یک تمایز ایجاد کنیم.
(مثلاً logistic regression)

مدل generative، هم x و هم y را model کرده است. یعنی هم رنگ و موقعیت x را مشخص می‌کند، و اگر بخواهیم سمپل قرمز جدیدی ایجاد کند، در محدوده‌ی مشخص شده این سمپل مشابه سایر سمپل‌های قرمز تولید می‌شود.

اما در مدلهای Discriminative چنین امکانی وجود نداشت.

Generative models ← مثلاً برای ساخت تصویر جدید چهره‌ی یک انسان به کار می‌رود.

این مدلها join x, y است و اطلاعات x و y هم ذخیره می‌شود ←
1) face detection
2) ساخت تصویر جعلی انسان face

LDA مانند

$$+P(x,y) = \underbrace{P(x|y)}_{\text{estimate from train data}}\ \underbrace{P(y)}_{\text{relative frequency}} \quad \text{or} \quad \boxed{P(Y|X)\,P(X)}$$

این کمتر استفاده می‌شود
چون محاسبه‌ی آن سخت‌تر است

+use Baye's rule to compute $P(Y|X)$.

+example: linear Discriminative analysis]
Nāive bayes]

ML_10

Linear Discriminant Analysis

+ a generative model

+ Data $\{(x[1], y[1]), \ldots, (x[n], y[n])\}$

+ model distribution of x for each class separately.

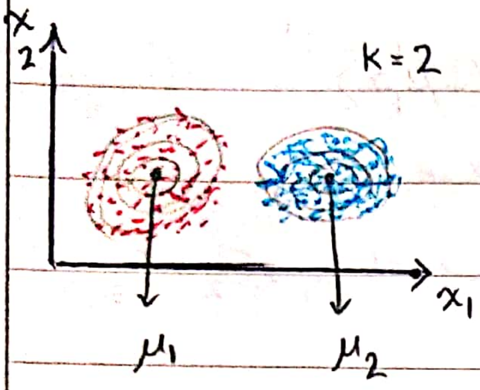
$p(x|Y=k)$ and $p(y)$

برای هر کلاس x از یک توزیع پیروی می کند.

+ then use Baye's rule $P(Y=k|x) = \frac{P(x|Y=k)P(Y=k)}{P(x)}$

به دنبال optimal k هستیم، یعنی به دنبال این هستیم که کدام کلاس محتمل تر است.
در روش logistic regression مستقیما $P(Y|X)$ را مدلسازی می کردیم.

هدف: با این فرض که بدانیم ۲ قرمز است می خواهیم برای کلاس قرمز و آبی دو تا مدل مجزا بسازیم.



* different choices for $x|Y=k$

$\Downarrow$

بستگی دارد x دارای پیوسته باشد یا گسسته.

فرض مدل LDA ← فرض می‌کند توزیع نرمال است.

$$x|Y=k \sim N(\mu_k, \Sigma)$$

$\Downarrow$ shared between our classes

$$P(x|Y=k)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right)$$

$$P(Y=k)=\pi_k \qquad \sum \pi_k=1$$

مثلا $k=\{1,2,3\}$ → توزیع multinomial

یعنی 3 کلاس داریم.

$$G(x)=\underset{k}{\arg\max}\, P(Y=k|x)P(Y=k) \xrightarrow{\text{log می‌گیریم}}$$

x را می‌گیرد و یک کلاس بر می‌گرداند.

$$\rightarrow \log P(x|Y=k)+\log P(Y=k)$$

$$\propto -\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)+\log\pi_k$$

$$=-\frac{1}{2}\left(\cancel{x^T\Sigma^{-1}x}-x^T\Sigma^{-1}\mu_k-\mu_k^T\Sigma^{-1}x+\mu_k^T\Sigma^{-1}\mu_k\right)+\log\pi_k$$

چون k ندارد حذف می‌شود.

$$\propto x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k = \delta_k(x)$$

تابعی از x و K

δ_1 ، δ_2 ، δ_3 را حساب می‌کند، هرکدام خروجی بیشتری داشت.

linear in $x \rightarrow c_0 + c_1 x_1 + c_2 x_2 + \dots + c_d x_d$

How to estimate parameters?

$$\mu_1 \in \mathbb{R}^d \;,\; \mu_2 \in \mathbb{R}^d \;,\; \underset{\frac{d^2-d}{2}+d}{\underset{\uparrow}{\Sigma}} \in \mathbb{R}^{d \times d} \;,\; \pi_1 \;,\; \pi_2$$

با فرض اینکه دو feature
داریم ← x_1 ، x_2

Likelihood function : $L(\theta)$

$$L(\theta) = \prod_{i=1}^{n} p(x[i], y[i] \mid \theta)$$

$$\rightarrow \ell(\theta) = \underbrace{\sum_{i=1}^{n} \log p(x[i] \mid y[i])}_{\text{Gaussian} \rightarrow \mu_k, \Sigma} + \underbrace{\sum_{i=1}^{n} \log p(y[i])}_{\text{multinomial} \rightarrow \pi_k} \qquad \sum \pi_k = 1$$

$$\hat{\mu}_k, \hat{\Sigma}, \hat{\pi}_k = \arg\max \ell(\theta)$$

$$n_k = \sum_{i=1}^{n} I(y[i]=k) \qquad \hat{\pi}_k = \frac{n_k}{n}$$

$$\mu_k = \left(\sum_{i=1}^{n} x[i]\, I(y[i]=k)\right) / n_k$$

$$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{I(y[i]=k)(x[i]-\hat{\mu}_k)(x[i]-\hat{\mu}_k)^T}{n-K}$$

Quadratic discriminant analysis (QDA):

+ separate covariance matrix for each class.

$$\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4, \dots, \Sigma_K$$

وقتی log-likelihood می‌گیریم، $x^T \Sigma^{-1} x$ حذف نمی‌شود.

$$\boxed{\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \log \pi_k}$$

↓ quadratic in x

+ # of parameters is much higher → احتمال overfit هست.

Regularized QDA: → شبیه ridge, LASSO

$$\hat{\Sigma}_k(\alpha) = \alpha \underset{\uparrow \text{QDA}}{\hat{\Sigma}_k} + (1-\alpha) \underset{\uparrow \text{LDA}}{\hat{\Sigma}}$$

α را با cross_Validation تعیین می‌کنیم ← هرچه آلفا به سمت صفر برود ← مدل ساده‌تر

$\Downarrow$

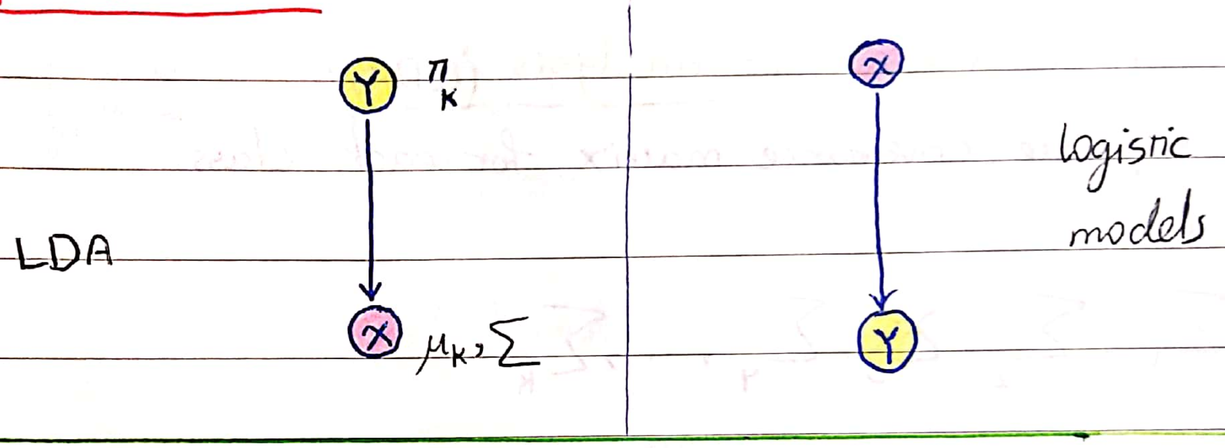
به سمت خطی شدن می‌رود.

(LDA)

$\downarrow$

هرچه α بیشتر ← مدل پیچیده‌تر می‌شود.

Graphical model



Näive Bayes classifier

+ a family of models
+ A Näive assumption $P(X|Y) = P(x_1|y)P(x_2|y) \cdots P(x_p|y)$

$\Downarrow$

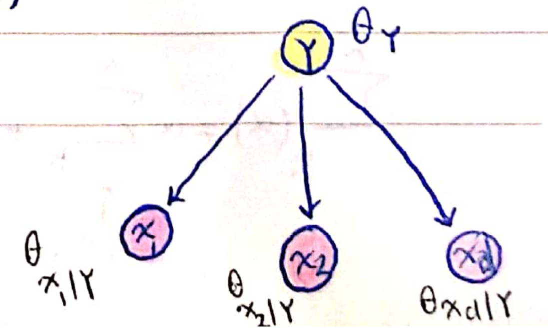
* Conditionally independent given Y

استقلال شرطی

+ Baye's rule $\rightarrow P(y|x) = \dfrac{P(x_1|y)P(x_2|y) \cdots P(x_d|y)P(Y)}{P(x)}$

$$= \frac{P(Y)\prod_{i=1}^{d} P(x_i|y)}{P(x)}$$

$\hat{y} = \underset{y}{\operatorname{argmax}}\ P_{\hat{\theta}}(y|x)$



/ /

Estimation : likelihood

+ parameters $P(Y)$, $P(x_j|y)$

$p(x_j|y)$ توزیعی است که به نوع x بستگی دارد.

اگر x ← binary ← bernoulli

" ← پیوسته ← نرمال یک متغیره

" ← categorial ← multinumial

" ← count data ← پواسن یا Neg Binom

⇓

+ distribution of $p(x_j|y)$ depends on data types.

$$L(\theta)=\prod_{i=1}^{n} P(x[i],y[i]) = \prod_{i=1}^{n} \underbrace{P(x_1[i]|y[i]) \cdots P(x_d[i]|y[i])}_{\text{vector } p(x[i]|y[i])} \times P(y[i])$$

$$\Rightarrow l(\theta)=\sum_{i=1}^{n} \log p(x_1[i]|y[i]) + \sum_{i=1}^{n} \log p(x_2[i]|y[i])$$

$$+ \cdots + \sum_{i=1}^{n} \log p(x_d[i]|y[i]) + \sum_{i=1}^{n} \log p(y[i])$$

پارامترهای feature های اول و دوم و... ارتباطی ندارند.

+ we can estimate parameters independently.

/ /

Ex :

A(age)	Smoking(S)	Lung Cancer(L)
1	0	1
1	0	0
0	1	0
0	1	1
1	1	1
1	0	0
0	0	0

$\arg\max P(L|A,S)$

$A,S \to X$

$L \to Y$

*predict the class for $S=1, A=1$

$P(L=1|S=1,A=1) \propto P(A=1|L=1) * P(S=1|L=1) * P(L=1)$

$P(L=0|S=1,A=1) \propto P(S=1|L=0) * P(A=1|L=0) * P(L=0)$

$$\begin{cases} \theta_L = P(L=1) = \frac{3}{7} , \quad \theta_{A|L=1} = P(A=1|L=1) = \frac{2}{3} \\ \theta_{A|L=0} = P(A=1|L=0) = \frac{1}{2} , \quad \theta_{S|L=1} = P(S=1|L=1) = \frac{2}{3} \\ \theta_{S|L=0} = P(S=1|L=0) = \frac{1}{4} \end{cases}$$

اینجا از MLE استفاده کردیم.

Laplase correction

$$p=\frac{0}{3}=\frac{0+1}{3+2}=\frac{1}{5}$$

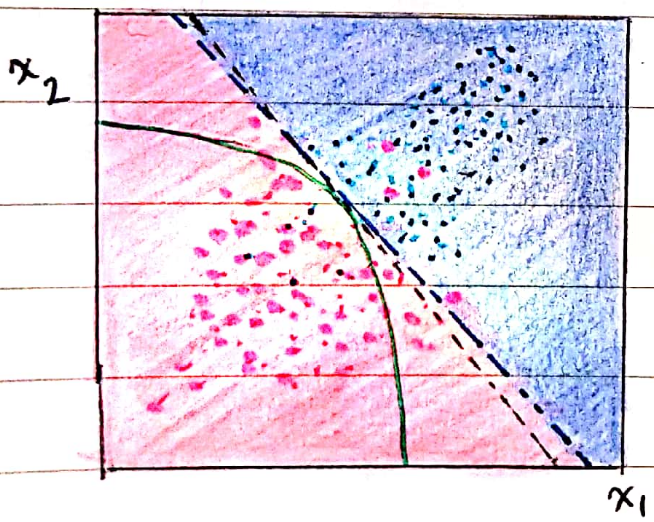
با تصحیح لاپلاس می‌توانیم اثر احتمالات صفر را کم کنیم.

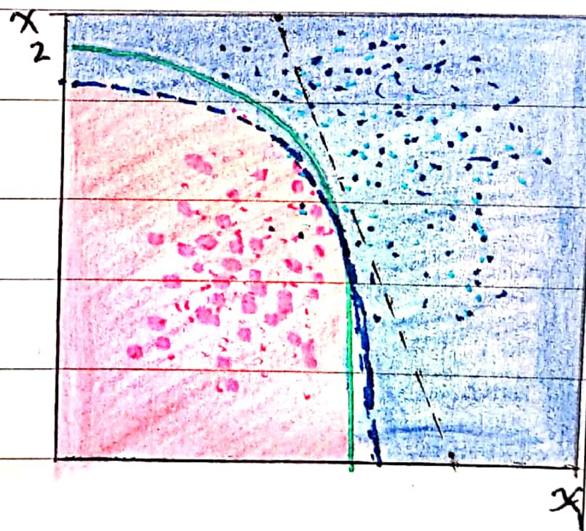
$$\Downarrow$$

$$P(X=1 \mid Y=k)=\frac{1+\sum I(x[i]=1, y[i]=k)}{k+\sum_{i=1}^{n} I(y[i]=k)}$$

pseudocount

مثالی شهودی از LDA و QDA





$\Sigma_1=\Sigma_2$ $\qquad\qquad$ $\Sigma_1\neq\Sigma_2$

— bayes classifier $\qquad$ — QDA $\qquad$ — LDA

همانطور که از شکل مشخص است وقتی که $\Sigma_1\neq\Sigma_2$ باشد (شکل سمت راست)، QDA به bayes classifier نزدیک‌تر و شبیه‌تر است اما در حالتی که $\Sigma_1=\Sigma_2$ (شکل سمت چپ) باشد، LDA کاراتر است و به bayes classifier شبیه‌تر است.