Subject:

Year. Month. Date. ( )

ML _ 5th session

exponential family :

یک مجموعه توزیع است که میتوان به فرم زیر نوشت:

natural parameter function

$$P_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp\left(t(\theta)^T s(y)\right)$$

↳ sufficient statistic

for most of distribution $s(y)=y$

☐ Generalized linear models :

هدف این است که با نوشتن تابعی از X به Y برسیم، تا الان تنها دو تابع را برای Y در نظر گرفتیم:

$Y|x,\theta \sim N(\mu, \sigma^2)$

$Y|x,\theta \sim \text{Bernouli}(\theta)$

Normal distribution :

$f(x) = E(Y|x;\theta) = \mu = \beta^T x$

مثل این است که بگیریم تابع g در اینجا identity است یعنی $g(z)=z$

Bernouli distribution:

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

g: response function

اگر Y از توزیع نرمال پیروی کند که g تابع همانی است و اگر Y توزیع برنولی داشته باشد همین گفته شده است.

g^{-1}: link function : $g^{-1}(\theta) = \beta^T x \xrightarrow{\text{برای logistic Reg.}} \log\frac{\theta}{1-\theta} = \beta^T x$

Sunwood

Subject:

Year. Month. Date. ( )

کاری که glm میکنه این است که اگر بدانیم $Y|x;\theta$ از یک توزیع exponential family است آنگاه glm

این امکان را برای ما فراهم میکند که اولاً بتوانیم g را پیدا کنیم و دوماً بتوانیم β را تخمین بزنیم.

$$Y = f(x) = g(\boxed{\beta^T x})$$

میتواند خطی یا غیرخطی باشد

$\beta^T x \rightarrow$ linear predictor

مثال) یک مدل خطی تعمیم یافته برای متغیر پاسخ (Y) که برابر است با تعداد read هایی که به یک gene

hit شده و متغیر x_1 برابر treatment/control و متغیر x_2 برابر gender باشد؟

برای این مثال $Y|x$ به $Possion(\lambda)$ در نظر بگیریم خوب است.

○ سه عنصر اصلی GLM :

① $Y|x;\theta$ از یک توزیع exponential family پیروی کند یعنی بتوانیم بنویسیم:

suppose $s(y) = y$

$$P(y|x;\theta) = \frac{1}{Z(\theta)} h(y) \exp(t(\theta)^T s(y))$$

depends on β & x

② فرض میکنیم $t(\theta) = \beta^T x$ (design choice)

③ $f(x) = E(Y|x;\beta) = g(\beta^T x)$

* حال چند مثال را مورد بررسی قرار میدهیم:

Subject:

Year. Month. Date. ( )

example : Bernouli distribution

$Y|x;B \sim \text{Bernouli}(\theta)$

① $f(x) = E(Y|x,B) = \theta$ *

② $t(\theta) = \log \theta/_{1-\theta} = B^T X$

$\theta/_{1-\theta} = e^{B^T X} \longrightarrow (1-\theta)e^{B^T X} = \theta \Rightarrow \theta = \frac{1}{1+e^{-B^T X}}$

$\xrightarrow{*} f(x) = \theta = \frac{1}{1+e^{-B^T X}} \Rightarrow g(z) = \frac{1}{1+e^{-z}}$

example: normal distribution

$Y|x;B \sim N(\mu, 1)$

① $t(\mu) = \mu = B^T X$

② $f(x) = E(Y|x,B) = \mu$

$\Rightarrow f(x) = B^T X \Rightarrow g(z) = z$

example: poisson distribution

$Y|x;B \sim \text{Possion}(\lambda)$ $P(Y|x;B) = \frac{\lambda^y e^{-\lambda}}{y!}$

HW → exponential Family of poisson

1) $t(\lambda) = \log \lambda = B^T X$

2) $F(x) = E(Y|x;\lambda) = \lambda$

$\log \lambda = B^T X \Rightarrow f(x) = \lambda = e^{B^T X}$

$g(z) = e^z$

Sunwood

Subject:

Year. Month. Date. ( )

هدف تعیین β ها است.

$$\Rightarrow L(\beta) = \prod_{i=1}^{n} P(y[i] \mid x[i], \beta)$$

$$= \prod_{i=1}^{n} \frac{\lambda^{y[i]} e^{-\lambda}}{y[i]!} \xrightarrow{\lambda = e^{\beta^T x}} \prod_{i=1}^{n} \frac{\exp(\beta^T x[i])^{y[i]} \exp(-\exp(\beta^T x[i]))}{y[i]!}$$

حالا که فرم likelihood را داریم اگر $\nabla L(\beta)$ و $\nabla^2 L(\beta)$ را محاسبه کنیم میتوانیم از روش نیوتن استفاده کنیم و تعیین پارامترها را انجام دهیم. (در تمرین حل شده)

□□ soft max regression :

اگر متغیر Y بتواند چند کلاس بگیرد (multi class) یعنی Y دارای توزیع multinomial است که این توزیع هم جزء exponential family است و سوال برهمین اساسی که گفتیم classification را انجام داد.

○ معرفی توزیع نرمال چند متغیره :

univarient : $X \sim N(\mu, \sigma^2) \Rightarrow P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

multivariate : $P(X; \mu, \Sigma) = \frac{1}{(2\pi)^{P/2} |\Sigma|^{1/2}} \times \exp\left(\underbrace{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}_{\text{Quadratic in } X}\right)$

$\mu \in \mathbb{R}^P$ $\qquad E(X) = \mu$

$\Sigma \in \mathbb{R}^{P \times P}$

$x = (x_1, \ldots, x_P)$ $\qquad \mathrm{cov}(X) = E_{z \sim P}\left[(z - E(x))(z - E(x))^T\right] = \Sigma$

← متقارن و نیمه معین مثبت.

شکل‌های مختلف توزیع برای مقادیر مختلف μ و Σ را در اسلایدها نشان دارند.

ضریب کمتر از یک یعنی پراکندگی کمتر است و قله توزیع بالاتر میره. صفحه 3-5 و part4 کتاب cs229

Discriminative versus Generative models

Discriminative: مدلی است که بر اساس تئوری احتمال شرطی $P(Y|X)$ تخمین زده میشه

مثل logestic، KNN، SVM. در این مدل‌ها یک فرم functional برای

$P(Y|x)$ در نظر میگیریم و سپس پارامترها را تخمین میزنیم مثلاً $P(Y|x,B)=\frac{1}{1+e^{-B^Tx}}$ که در لجستیک

داشتیم.

Generative: مدل statistical را برای $P(X,Y)$ طرح میکنیم:

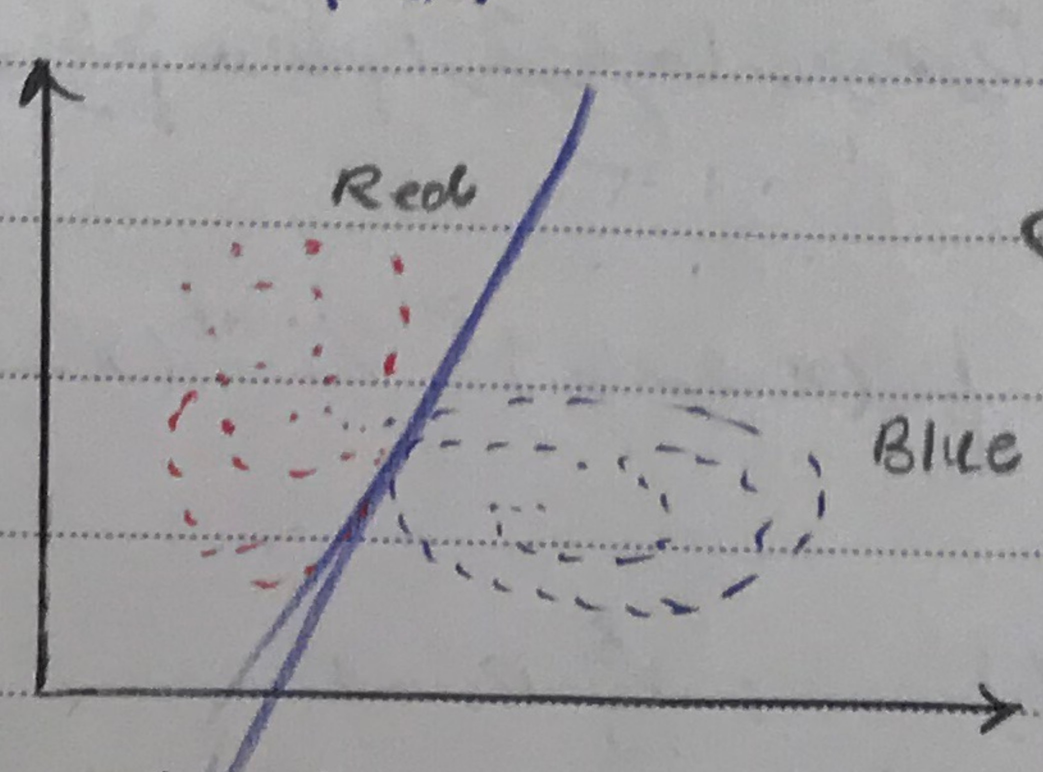
- $P(X,Y) = P(X)\,P(Y|X)$
- $P(X,Y) = P(Y)\,P(X|Y)$

از data تخمین زده میشه

مثلا Bayes classifier هدف محاسبه $P(Y|X)$ است که با استفاده از قانون بیز داریم:

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

مثل: LDA، HMM، NB



مدل‌های Discriminative هدف این است که احتمال مرزی را با شرط دیتا محاسبه کنیم $P(Y=B|X)$ / $=R$ و طبق این احتمال مرزی‌ها را پیدا می‌کنند

اما مدل‌های Generative هدفشان از توزیع‌های $P(Y=B), P(Y=R)$، $P(X|Y=R)$ و $P(X|Y=B)$

لذا محاسبه می‌کنند و از روی این $P(Y|X)$ را با قاعده بیز محاسبه می‌کنیم.

مزیت مدل‌های Generative این است که مکانیسم تولید دیتا را بدست میاوریم.

Linear Discriminant analysis :

LDA یک مدل generative است و دیتایی که داریم هم شکل قبل است و ایده این است که توزیع x را برای هر کلاس مدل کنیم در واقع هدف محاسبه دو توزیع زیر است:

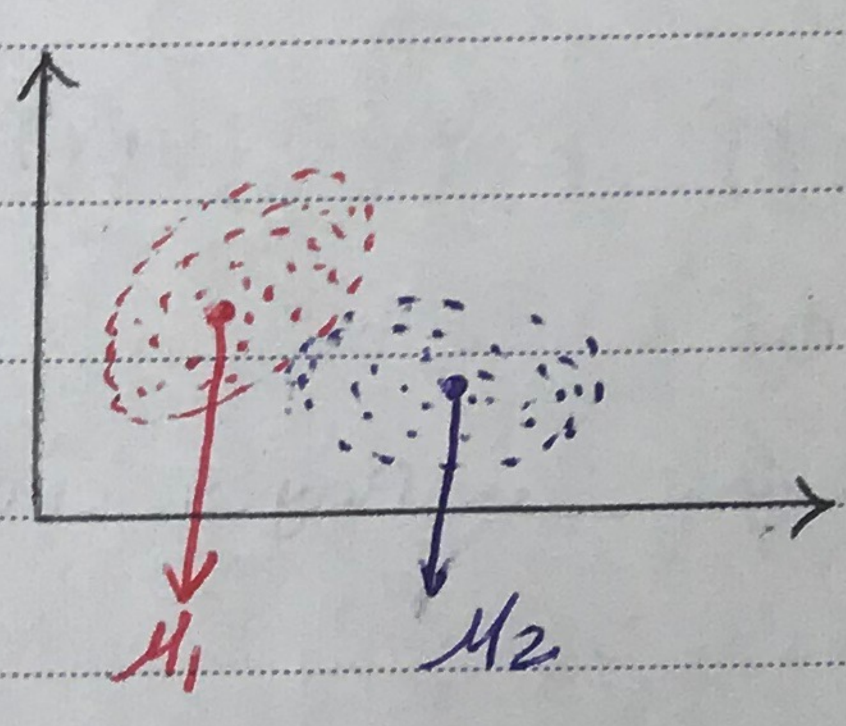
$P(x|Y=k)$ & $P(Y)$

$x \in \mathbb{R}^p$

$D = \{(x[1],y[1]), (x[2],y[2]), \ldots, (x[n],y[n])\}$ $\quad y \in \{1,2,\ldots,k\}$

$$P(Y=k|x) = \frac{P(x|Y=k)\,P(Y=k)}{P(x)}$$

در LDA فرض بر این است که این عبارت دارای توزیع نرمال است



$k=2$

$x|Y=1 \sim N(\mu_1, \Sigma)$

$x|Y=2 \sim N(\mu_2, \Sigma)$

بین کلاس ها مشترک است

حال فرض می‌کنیم هر کلاس از توزیع ها به چه صورت هستند:

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

$P(Y=k) = \pi_k$ (multinomial) ; $\sum_{i=1}^{k} \pi_i = 1$

بنابراین تا اینجای کار مدل را معرفی کردیم و گفتیم پارامترها Σ و μها و πها را باید تعیین بزنیم از دیتا و

در نتیجه میتوانیم احتمال هر دیتایی را محاسبه بکنیم.

$$G(x) = \underset{k}{\operatorname{argmax}}\ P(Y=k \mid x) = \underset{k}{\operatorname{argmax}}\ \log P(Y=k \mid x)$$

$$= \underset{k}{\operatorname{argmax}}\ \log\left(\frac{P(x \mid Y=k)\,P(Y=k)}{P(x)}\right) = \underset{k}{\operatorname{argmax}}\ \left(\log P(x \mid Y=k) + \log P(Y=k)\right)$$

که مستقل از k است بنابراین کنار میگذاریم

$$\propto \underset{k}{\operatorname{argmax}}\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \log \pi_k\right)$$

$$\Downarrow$$

$$= -\frac{1}{2}\left(x^T\Sigma^{-1}x - \underbrace{x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x}_{\text{عدد است در ترانهاده}} + \mu_k^T\Sigma^{-1}\mu_k\right)$$

عدد ساده خود عدد است

$$= -\frac{1}{2}\underbrace{x^T\Sigma^{-1}x}_{\text{به k ارتباطی ندارد}} + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$$

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

پس در نهایت به منحنی رسیدیم که در x خطی است.

حال باید به این سوال پاسخ دهیم که پارامترها را چگونه تعیین بزنیم:

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

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

$$2 \xrightarrow{\log} l(\theta) = \sum_{i=1}^{n} \left(\log P(x[i] \mid y[i]) + \log P(y[i])\right)$$

در حالت ساده فرض می‌کنیم تنها دو کلاس داریم:

$$l(\theta) = \sum_{y[i]=1} \log P(x[i] \mid y[i]=1) + \sum_{y[i]=2} \log P(x[i] \mid y[i]=2)$$

$$+ \sum_{i=1}^{n} \log P(y[i])$$

بنابراین با استفاده از MLE به تعیین $\hat{\mu}_k$ ، $\hat{\Sigma}$ و $\hat{\pi}_k$ می‌پردازیم.

$$\hat{\mu}_k , \hat{\Sigma} , \hat{\pi}_k = \operatorname{argmax} \, l(\theta)$$

تعداد اعضای کلاس k : $n_k = \sum_{i=1}^{n} I(y[i]=k)$

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

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

$$\hat{\pi}_k = \frac{n_k}{n}$$

Quadratic discriminant analysis (QDA)

* تفاوتی که این روش با LDA این است برای هر کلاس در این متد ماتریس‌های کوواریانس جدا داریم:

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

این تابع در x، Quadratic است و در واقع desicien boundary در x، quadratic است

و تفاوتی که با LDA دارد در این است که تعداد پارامترهای بیشتری دارد.

⁘ Regularized QDA

$$\hat{\Sigma}_k(\alpha) = \underset{\underset{CV}{\swarrow}}{\alpha}\,\hat{\Sigma}_k + (1-\alpha)\,\hat{\Sigma} \rightarrow \text{the Same as LDA}$$

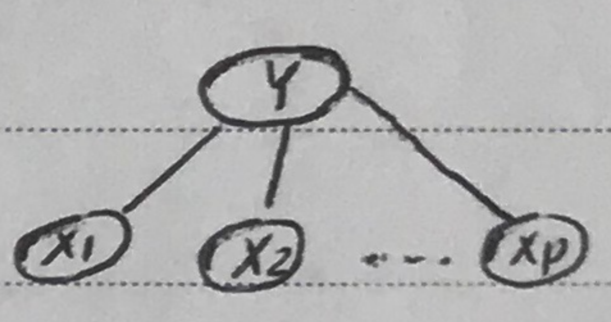
Naive Bayes classifier:

این مدل هم generative است و $P(X,Y) = P(X|Y)P(Y)$ را می‌نویسیم

در این مدل فرضی که داریم این است که x ها به شرط دانستن Y از هم مستقل هستند:

$$P(X|Y) = P(x_1|Y)\,P(x_2|Y)\dots P(x_p|Y)$$

↳ conditional independent



$$P(Y|X) = \frac{P(Y,X)}{P(X)} = \frac{P(x_1|Y)\,P(x_2|Y)\dots P(x_p|Y)\,P(Y)}{P(X)}$$

هدف تعیین پارامترهای مدل است؛ این پارامترهای مدنظر $P(Y)$ ها و $P(x_i|Y)$ ها است:

$$L(\theta) = \prod_{i=1}^{n} P(y[i], x[i];\theta) = \prod_{i=1}^{n} P(x[i]|y[i])\,P(y[i])$$

$$= \prod_{i=1}^{n} P(x_1[i]|y[i]) \dots P(x_p[i]|y[i])\,P(y[i])$$

Subject:

Year. Month. Date. ( )

$$\Rightarrow \log(L(\theta)) = \sum_{i=1}^{n} \left[\log(x_1[i] \mid y[i]) + \dots + \log(x_p[i] \mid y[i]) + \log P(y[i]) \right]$$

Example : X , Y ~ Bernouli

(A) Age , (S) Smoke , (L) Lung cancer

A	S	L
۱	۰	۱
۱	۰	۰
۰	۱	۰
۰	۱	۱

$P(L \mid A, S) = ?$

$$P(L) \begin{cases} P(L=1) = \theta = ۳/۷ \\ P(L=0) = (1-\theta) = ۴/۷ \end{cases}$$

$P(A \mid L) \rightarrow P(A=1 \mid L=1) = \frac{۲}{۳}$

$P(A=1 \mid L=0) = \frac{۲}{۴}$

$P(S \mid L) \rightarrow P(S=1 \mid L=1) = \frac{۲}{۳}$

$P(S=1 \mid L=0) = \frac{۱}{۲}$

حال سوال احتمال A = ۱ و S = ۱ را بخواهیم:

$$P(L=1 \mid S=1, A=1) \propto \overset{2/۳}{P(A=1 \mid L=1)} \, \overset{3/۷}{P(S=1 \mid L=1)} \, \overset{3/۷}{P(L=1)} = \frac{4}{۴۹}$$

$$P(L=0 \mid S=1, A=1) \propto \overset{۲/۴}{P(A=1 \mid L=0)} \, \overset{۱/۴}{P(S=1 \mid L=0)} \, \overset{۴/۷}{P(L=0)} = ۲/۱۸$$

Sunwood