Lecture 5 LDA, Bayesian decision theory

Note Title 9/26/2017

$X \xrightarrow{f} Y$ estimate f or estimate $P(Y|X)$ or $P(X,Y)$

$$P(y=c|x,\theta) = P(X|y=c,\theta)\,P(y=c|\theta)$$

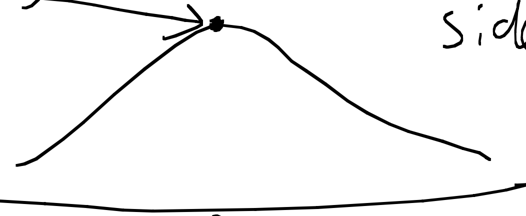
$$D=\{(x_i,y_i\} \sim N(\mu,\Sigma)$$

MLE , MAP $\leftarrow P(\theta|D)$ in Bayseian setting

$P(D|\theta)$ or $\log P(D|\theta)$

$$= \frac{P(D|\theta)\,P(\theta)}{N}$$

side



$$\theta = (\mu,\Sigma)$$

$$P(y=c|x,\theta) = \frac{P(x|y=c,\theta)\,P(y=c|\theta)}{\sum_{c'} P(x|y=c',\theta)\,P(y=c'|\theta)}$$

$= P(X,y=c)$ joint

classifier

generative classifier

discriminative classifier will have a direct function

Back to modelling class conditional densities.

$\hookrightarrow P(X|Y=c)$

Naive Bayes classifier $\rightarrow \prod_{i=1}^{D} P(X_i|Y=c)$

$P(X|Y=c) = \sim N(\mu_c, \Sigma_c)$ = Gaussian discriminant analysis

otherwise

$\Sigma = \Sigma_c$

QDA Quadratic discriminant

LDA linear discriminant analysis

Back to naive Bayes

(2) $\Sigma = \Sigma_c = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_d^2 \end{bmatrix}$

$$= \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}}$$

look into

$$\boxed{\exp\left(-\frac{1}{2}(X-\mu)^T \Sigma (X-\mu)\right)}$$

$$\simeq \exp\left(\frac{1}{-2}\left(\sigma_1^2 (x_1-\mu_1)^2 + \sigma_2^2 (x_2-\mu_2)^2 + \cdots \sigma_D^2 (x_d-\mu_d)^2\right)\right)$$

$$= \frac{1}{\left((2\pi)^{1/2}\right)^D} \frac{1}{(\sigma_1^2)^{1/2} \cdots (\sigma_D^2)^{1/2}}$$

$$\exp\left(-\frac{1}{2}\sigma_1^2 (x_1-\mu_1)^2\right) \exp\left(-\frac{1}{2}\sigma_2^2 (x_2-\mu_2)^2\right) \cdots \exp\left(-\frac{1}{2}\sigma_D^2 (x_D-\mu_D)^2\right)$$

Summary $= N(\mu_1, \sigma_1) \cdots\cdots N(\mu_D, \sigma_D)$

look into this to find decision boundary

$$P(y=1 \mid X) = P(y=2 \mid X)$$

$$\frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(X-\mu_1)^T \Sigma_1 (X-\mu_1)\right) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(X-\mu_2)^T \Sigma_2 (X-\mu_2)\right)$$

4.2.4 MLE for discriminant analysis

$$\log p(D|\theta) \quad (\text{IID})$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{I}(y_i = c) \log \pi_c + \sum_{c=1}^{C} \Big[\sum_{i: y_i = c} \log \mathcal{N}(x \mid \mu_c, \Sigma_c) \Big)$$

$$\hat{\pi}_c = N_c / N$$

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{y_i = c} x_i = \frac{1}{N_c} \sum_{i=1}^{N} \mathbb{I}(y_i = c) x_i$$

$$\underset{D \times D}{\hat{\Sigma}_c} = \frac{1}{N_c} \sum_{y_i = c} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T$$

Side $\hat{\Sigma}_c$ need to be invertible

i.e it need to be full rank.

$= D$

$N_c < D$. $\Rightarrow$ Can't invert

$N_c \gg D$ $\Rightarrow$ high chances of being invertible

Imagenet $2 \cdot 4 \times 10^6$ million image

Even in big data era when we cluster based on class label we ~~of~~ have less sample

1000 classes

So 1000 images / class

each image

$$X_{300\times 300\times 3} \in \mathbb{R}^{270000 = D}$$

$$N_c = 1000 \ll D\ (270000)$$

What to DO ?

- Duplicate sample with perturbation
- Dimentionality reduction (PCA)
- use diagonal matrix (D parameter per class)
- $\underset{\text{Full}}{\Sigma_c} = \Sigma \quad \forall c$

or $\Sigma_c = \Sigma$ and diagonal

— Use Full covariance but impose prior

- determinant of matrix = product of eigenvalues

side — For a matrix sum of eigenvalue = sum of diagonal entries = trace of matric

Chapter 5

5.1 Bayesian Decision theory

$$X \sim N(\mu, \Sigma)$$

choose some action $a \in \mathcal{A}$ (action space)

How well we did?

define a Loss $L(y_{true}, a)$

and we want to minimize the loss

$$L(y_{true}, a) = \mathbb{I}(y_{true} \neq a)$$ [mis classification loss]

what about regression?

$$L(y_{true}, a) = (y_{true} - a)^2$$

<u>Decision rule</u> $\delta : \mathcal{X} \to \mathcal{A}$

optimal policy

$$\delta(x) = \underset{a \in \mathcal{A}}{\arg\min}\ \mathbb{E}[L(y, a)]$$

In Bayesian approach to decision theory, the optimal action, have $x \in \mathcal{X}$, is defined as the action that minimizes the posterior Expected loss

$$\rho(a|x) = \mathbb{E}_{(y|x)}[L(y, a)] = \sum_{y} L(y, a) P(y|x)$$

Bayes Estimate (Bayes decision rule)

$$\delta(x) = \arg\min_{a \in \mathcal{A}} \rho(a|x)$$

MAP estimate minimize 0-1 loss

$$L(y,a) = L(y,\hat{y}) = \mathbb{I}(y \neq \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y} \end{cases}$$

	$a = \hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	1
$y = 0$	1	0

Then posterior expected loss is

$$\rho(a|x) = 1 \cdot p(a \neq y|x) = 1 - p(a = y|x)$$

$$a = \hat{y} = \arg\max_{a \in Y} p(a = y|x)$$

justifies our earlier choice of MAP

5.7.1.3 Posterior mean minimizes L_2 loss for continuous parameter

$p=2$ $\quad \hat{y} = a = E[a|x]$ $\quad [a-y]^p$

If $p=1$ basically $|a-y|$

then optimal action is posterior median $P(y<a|x) = P(y>a|x) = .5$

Note outlier become more prominent to the power p. Like in image normalization using $\|L\|_p$ or max value may not be the right thing to do. Normalize using 98 or 99 percentile.

5.7.2 The False positive and False negative

Binary setting

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	0	L_{FN}
$y=0$	L_{FP}	0

Posterior Expected loss

$$\rho(\hat{y}=0|x) = L_{FN}\, P(y=1|x)$$

$$\rho(\hat{y}=1|x) = L_{FP}\, P(y=0|x)$$

Hence pick $\hat{y}=1$ if

$$\rho(\hat{y}=0|x) > \rho(\hat{y}=1|x)$$

$$L_{FN}\, P(y=1|x) > L_{FP}\, P(y=0|x)$$

$$\frac{P(y=1|x)}{P(y=0|x)} > \frac{L_{FP}}{L_{FN}} = \Gamma$$

$$L_{FP} = c\, L_{FN}$$

5.7.2.1 ROC curves, AUC.

Suppose we are in supervised binary decision setting $D = \{(x_i, y_i)\}$

we apply Γ and count no of true, positive, false, positive, true negative, False negative

		True 1	0	Total
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = TN + FN$

Total $\left| N_+ = TP + FN \quad \begin{matrix} N_- \\ = FP \\ + TN \end{matrix} \right|$

TPR (sensitivity, recall, hit rate) $= \dfrac{TP}{N_+}$

FPR (False alarm, type I error) $= \dfrac{FP}{N_-}$

ROC plot of TPR vs FPR for various value of τ

(Receiver operating characteristic)

$\tau = 1 \Rightarrow FPR = 0 = TPR = 0$

$\tau = 0 \Rightarrow FPR = FPR = 1$

ROC is summarized using one number called AUC (Area under the curve)