180010009

Week-11

LDA (Linear Discriminant Analysis)

Multi-class & binary class

2-class

$$\underbrace{\Sigma_W^{-1} \Sigma_B}_{\text{Rank 1 Matrix}} W = \lambda W$$

$$\Sigma_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

There is only one Eigen vector

$$\underset{N \times N}{\Sigma_W^{-1}} \underset{N \times 1}{(\mu_1 - \mu_2)} \underbrace{\underset{1 \times N}{(\mu_1 - \mu_2)^T} \underset{N \times 1}{W}}_{\text{let say Scalar Value} = C}$$

$$\Sigma_W^{-1} (\mu_1 - \mu_2)\, C = \lambda W$$

$$\Sigma_W^{-1} (\mu_1 - \mu_2) = \frac{\lambda}{C} \omega$$

$$\Sigma_W^{-1} (\mu_1 - \mu_2) \propto W$$

No matter what is N, there is only eigen vector for 2 class Separation.

N = dimension of feature Vector.

Scanned with CamScanner

Page No.
Date

LDA for k-classes

→ In 2 classes, projection is done onto vector w

→ In k-class projection is done onto vector W

$W = [w_1, w_2, w_3 \ldots w_{k-1}]$ where w_i is $N \times 1$

→ Projected means $W^T\mu_1, W^T\mu_2 \ldots W^T\mu_k$.

→ Projected cov $W^T\Sigma_1 W, W^T\Sigma_2 W \ldots$

$$W^T\Sigma_1 W = w_1^T\Sigma_1 w_1 + w_2^T\Sigma_1 w_2 \ldots + w_{k-1}^T\Sigma_1 w_{k-1}$$
$$= Tr(W^T\Sigma_1 W)$$

Similarly $W^T\Sigma_2 W = Tr(W^T\Sigma_2 W)$
$W^T\Sigma_k W = Tr(W^T\Sigma_k W)$

$$Tr(W^T\Sigma_w W) = Tr(W^T\Sigma_1 W) + Tr(W^T\Sigma_2 W) \ldots Tr(W^T\Sigma_k W)$$

Σ_w is within class Covariance

Objective minimize $Tr(W^T\Sigma_w W)$

$\Sigma_{B1} = (\mu_1 - \mu_2)(\mu_1^T - \mu_2^T)$ is b/w class cov.

$$F = W^T(\mu_1 - \mu_2)(\mu_1^T - \mu_2^T)W$$

Scanned with CamScanner

$E = Tr(W^T \Sigma_{B1} W)$

Goal is to maximize $W^T \Sigma_{B1} W$

$$\Sigma_B = \Sigma_{B1} + \Sigma_{B2} - + \Sigma_{BK}$$

$$\max \frac{Tr(W^T \Sigma_B W)}{Tr(W^T \Sigma_W W)}$$

$$L(W, \lambda) = (W^T \Sigma_B W) - \lambda(W^T \Sigma_W W - 1)$$

$$\frac{\partial L}{\partial W} = 2 \Sigma_B W - 2\lambda \Sigma_W W = 0$$

$$\Sigma_B W = \lambda \Sigma_W W$$

$$\underbrace{\Sigma_W^{-1} \Sigma_B W}_{\text{is } k-1 \text{ rank matrix}} = \lambda W$$

Thus have $k-1$ eigen ~~vec~~ vector

$$\begin{matrix} \Sigma_B = & \Sigma_{B12}, & \Sigma_{B13} & \cdots, & \Sigma_{B1k} \\ & \vdots & \ddots & & \vdots \\ & \Sigma_{Bk1}, & \Sigma_{Bk2} & \cdots & \Sigma_{B(k-1)k} \end{matrix}$$

With class

$$\Sigma_W = \Sigma_{B11}, \Sigma_{B22}, \Sigma_{B33} \cdots, \Sigma_{Bkk}.$$

we know Σ_W
we know Σ_T (Total variance

$$\Sigma_T = \Sigma_W + \Sigma_B$$

$$\Sigma_B = \Sigma_T - \Sigma_W$$

$$\Sigma_T = \frac{1}{M} \sum_{i=1}^{M} (x_i - \mu)(x_i - \mu)^T$$

Pair-wise classifier

~~1 v/s rest~~ type

Thus. for k-classes, then $k(k-1)/2$ pairwise classifier

For test eg.

→ Classification result from each classifier

→ Final decision based on best classification among all the classifier

LDA one v_s rest

→ k classifier in total

→ Final decision based on best classifier among all the classifier

Bayesian Approach

Probability : Frequency count v/s Bayesian Approach

Conditional Probability

$P(A/B)$ = ~~Rd~~ Prob of happing A when B happened

$$P(X/y)\,P(Y) = P(Y/X)\,P(X)$$

$$P(X/Y) = \frac{P(Y/X) \cdot P(X)}{P(Y)}$$

core of Bayesian Approach.

objective freq count → # favourable / # Total

Subjective Bayesian: compute prob. in terms of other prob values

Baysesian learning : ~~for~~ approximation

Eg. Regression

Bayesian Classification
- Bayes Optional Classifier
- Naive Bayes
- Gibbs Sampling

Page No.
Date

Baysian Network

→ tree-like classifier using Bayesian Approach

~~$P(Y/X) = P(X/Y)$~~

Bayesian Classification

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

X :- represnt feature vector
Y :- Class label

→ $P(Y/X)$ is a posteriori probability

Given X probability it represent Y

→ $P(X/Y)$ is likelihood prob or Class Condition. how well model represents a given set of feature vector X.

for given class label ~~how~~ what is prob X represent Y

→ prior prob :- a prior knowledge about model parameter Y
$P(Y)$

→ An evidence prob evidence ~~of~~ about X
$P(X)$

	Discriminative	Generative
$P(Y/X)$	Training	Testing
$P(X/Y)$	Testing	Training.

For Generative Classifier

Training $P(X|Y)$
Testing $P(Y|X)$

$$P(X) = \Sigma_{Y_i} P(X/Y_i)\, P(Y_i)$$

Ideally

$$P(Y) = P(X) = 1/N \quad N = \text{numclass}$$

$$P(X/Y) = P(Y|X)$$

Same is exploited in Bayesian Approach

Freq count employs maximum likelihood approach for model parameter estimation

$$\hat{Y} = \text{argmax}_Y P(X/Y)$$

Bayesian method employs posterior Approach for model parameter estimation

$$\hat{Y} = \text{arg}_Y P(Y/X)$$

Bayesian learning

learning hypothesis function by exploiting bayesian concept is Bayesian learning

$y :\rightarrow$ replaced by h
$x \rightarrow D$

$$p(h/D) = \frac{P(D/h) \cdot p(h)}{p(D)}$$

how to find h? (In Bayesian learning framework)

$h_{MAP} = max_{h \in H}\, p(h/D)$

Using Bayes theorem $h_{MAP} = max_{h \in H} \frac{P(D/h)\, p(h)}{p(D)}$

p(D) is independant of hypothesis

MAP : Maximum a posteriori based hypothesis

$h_{MAP} = p(D/h) \cdot p(h)$

If all hypothesis are equally likely i.e. prior prob p(h) are equal

we get $h_{ML} = max_{h \in H}\, p(D/h)$

This is called likelihood hypothesis also called Maximum likelihood (ML)

Page No.
Date

Finding hypothesis by Bayesian learning

Given function $d_i = f(x_i) + \epsilon_i$

d_i is from Normal distribution $N(f(x_i), \sigma^2$

let h represent hypothesis function represent f

Can be estimated using Bayesian learning

$h_{ML} = \arg\max_h p(D/h)$

$$= \arg\max_h \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$= \log\left(\arg\max_h \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}\right)$$

$$= \arg\min_h \sum_{i=1}^{M} \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$

Same as least square error equation (Not prob. Appro

Assumption: ~~we have~~ becoz of Generative process we have hypothesis functi

Sample space of hypothesis is given.

hypothesis by non-probablistic Approach is same as what we got in Bayes learning.

~~hgp~~ Bayes Optimal Classifier

h_{MAP} may not be Optimal Classifier for a given test data

binary Classification task (+ or -)

for eg. There are 3 hypothesis h_1, h_2, h_3

MAP h_1 has $p = 0.4$ +
h_2, h_3 has $p = 0.3$ - total $0.6 > 0.4$

Then Bayes Optimal Classifier is useful

Given by $\arg\max_{v_j} = \sum_{v_j \in V} P(v_j / h_i) \, P(h_i / D)$

→ It is called optimal bcoz, no other classifier with same set of hypothesis and prior can outperform this on average.

→ hypothesis space is big & can't be applied in exhaustive way.

→ Gibbs Sampling that randomly chooses an hypothesis and finds class Error by this is within allowable limits.

Scanned with CamScanner