$P(x_1, \ldots x_d | y)$, $x_j \in \{1, 2, \ldots m\}$

parameters $\equiv m^d \times k$

ex:- Text classificat^n. $X \equiv x_1 \ldots x_d$

x_j = whether doc. contains j^{th} word in vocab or not.

$y \in$ {topics $\in$ {sports, science ....}} ($k = 500$)

$d = 50{,}000$

$\therefore$ # parameters $= 2^{50{,}000} \times K$

Naive Bayes assumption

$$P(x_1, \ldots x_d | y) = \prod_{j=1}^{d} P(x_j | y)$$

ie conditioned on the class variable, the attributes are independant of each other. This is quite reasonable and holds approximately in many real applications.

~~$P(x_1, \ldots x_d) = \prod_{i=1}^{d} P(x_j)$~~

Doc classificat^n. using NB

$$P(x_1, \ldots x_d | y) = \prod_{j=1}^{d} P(x_j | y)$$

For each subject or class, a prob. of occurence of each word in the vocab.

$\theta_{kj} = P(x_j | y = k) \Rightarrow$ Learned during training

$\therefore$ # of parameters $= d \times K$

$X[0,1,0,0,0,1,0,1]$ $d=8$

$P(X|y=k) = (1-\theta_{k1})\theta_{k2}(1-\theta_{k3})\cdots\cdots(1-\theta_{k7})\theta_{k8}$

$\therefore P(y=k|X) \propto P(y=k) \times P(X|y=k)$

$P(X|y=k) = \prod_{j=1}^{d} (1-\theta_{kj})^{1-x_j}\theta_{kj}^{x_j}$

Then normalise it

Gaussian Discriminant Analysis (GDA)

Only works for real valued attrib^n.

$P(x_1, \cdots x_d | y) \sim N(\mu_y, \Sigma_{d\times d})$

$x_1, \cdots x_d \in R^d$ $\quad \mu_y \in R^d$

Multidimensional gaussian

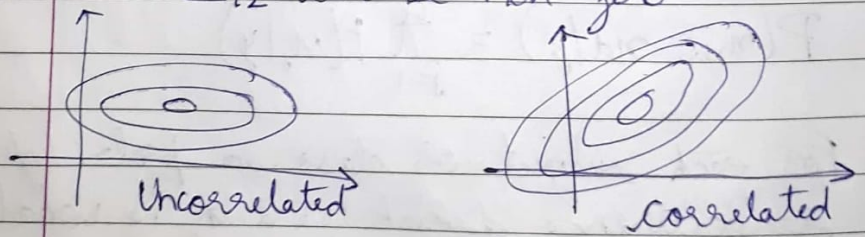
$[x_1, x_2] \sim N\left([0.5, 1], \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \Sigma\right)$

When x_1 & x_2 are un-correlated

σ_1, σ_2: Std. dev. of attrib. 1 & 2.

When x_1 & x_2 are correlated,

Σ_{12} will be non-zero



$\therefore \Sigma_{jk}$: covariance b^n. attributes x_j & x_k

Σ_{jj} : Variance of attribute x_j

$x_1, \ldots x_d \sim N(\underset{d\times 1}{\mu}, \Sigma_{d\times d})$

$\therefore P(x_1, \ldots x_d) = \frac{1}{\sqrt{2\pi|\Sigma|}^{d/2}} e^{\frac{-(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}}$

Suppose $d=2$,

$\log P(x_1, x_2) = \frac{-1}{2}[x_1\ x_2]\begin{bmatrix}\sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} + \ldots\ldots$

$= \frac{-1}{2}[x_1\ x_2]\begin{bmatrix}p_{11} & p_{12} \\ p_{21} & p_{22}\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix}$

$= \frac{-1}{2}[p_{11}x_1^2 + p_{21}x_2x_1 + p_{12}x_1x_2 + p_{22}x_2^2]$

$P(x_1, \ldots x_d | Y=k) \sim N(\mu_y, \Sigma_y)$

(μ_y: d parameters; Σ_y: d^2 parameters; K classes)

Parameter estimation

For each class k,

$$\hat{\mu}_k \equiv \frac{\sum_{\substack{i=1 \\ y_i=k}}^{N} x^i}{\sum_{i=1}^{N}[[y_i == k]]? = N_k}$$

(N_k ← No. of eg. in k^{th} class.)

$$\Sigma_k = \frac{\sum_{i=1}^{N}(x^i - \mu_k)(x^i - \mu_k)^T}{N_k - 1}$$

(for unbiased)

Page No.:
Date:

Alternately,

σ_j^2 = Variance of j^{th} attributes

σ_{jl} = co-variance of attributes j & l.

$$\Sigma_k = \begin{bmatrix} \sigma_1^2 & \sigma_{jl} \\ \sigma_{lj} & \sigma_2^2 \end{bmatrix}$$ ← Symmetric matrix

Linear Discriminant Analysis (LDA)

Assume Σ_k is the same for all K classes.

$\Sigma_1 = \Sigma_2 = \dots = \Sigma_K$

$P(x_1 \dots x_d | y = K) = N(M_K, \Sigma)$

8/08

The estimate of the shared covariance is

$$\Sigma = \frac{\sum_{i=1}^{N} (X^i - M_{y_i})^T (X^i - M_{y_i})}{N-1}$$

Decision Boundary

$k = 1, 2$ $M_1, \Sigma_1, M_2, \Sigma_2$, $P(y=1) \equiv \pi_1$

$P(y=2) \equiv \pi_2$

$\pi_1 + \pi_2 = 1$

Under what X is,

$P(Y=1|X) > P(y=2|X)$?

$$\therefore \log P(y=1|x) > \log P(y=2|x)$$

$$\therefore \log P(y=1)\,P(x|y=1) > \log P(y=2)\,P(x|y=2)$$

$$\therefore \log \pi_1 + \log \frac{1}{\sqrt{2\pi}\,|\Sigma_1|^{d/2}} e^{-(x-M_1)^T \Sigma_1^{-1}(x-M_1)}$$

$$\log \pi_1 - \log |\Sigma_1|^{d/2} - \frac{1}{2}(x-M_1)^T \Sigma_1^{-1}(x-M_1) > \log \pi_2 - \log |\Sigma_2|^{d/2} - \frac{1}{2}(x-M_2)^T \Sigma_2^{-1}(x-M_2)$$

→ const. & indep. of x.

$$\therefore (x-M_2)^T \Sigma_2^{-1}(x-M_2) - (x-M_1)^T \Sigma_1^{-1}(x-M_1) > \text{const}$$

$$\therefore -x\Sigma_1^{-1}x + 2M_1^T\Sigma_1^{-1}x - M_1^T\Sigma_1^{-1}M_1 + x\Sigma_2^{-1}x - 2M_2^T\Sigma_2^{-1}x + M_2^T\Sigma_2^{-1}M_2 > \text{const}$$

(const.)

$$\therefore -x^T[\Sigma_1^{-1} - \Sigma_2^{-1}]x + 2(M_1^T\Sigma_1^{-1} - M_2^T\Sigma_2^{-1})x > \text{const}$$

When $\Sigma_1 = \Sigma_2$; the quadratic term vanishes.
The 2nd term,

$$2(M_1^T - M_2^T)\Sigma^{-1}x \equiv a_1x_1 + a_2x_2 + \dots + a_dx_d$$

↑ linear in x.

Decision boundary

$a_1 x_1 + a_2 x_2 + \dots a_d x_d > \text{const}^n$

When $\Sigma_1 \neq \Sigma_2$

$$x^T(\Sigma_1^{-1} - \Sigma_2^{-1})x \equiv \sum_{l=1,j=1}^{d,d} (p_{ej1} - p_{ej2}) x_e x_j$$

+ linear terms

quadratic decision boundary

Boundary for naive Bayes

$$\sum_{j=1}^{d} \left(\frac{1}{\sigma_{j1}^2} - \frac{1}{\sigma_{j2}^2} \right) x_j^2 + \text{linear terms}$$

The max likelihood training method

Identify the form of the distrib^n that characterizes

Let θ = the parameters of distrib^n

The max. likelihood estimates of θ are the ones which maximize the prob. of the training data.

Typically assume that training samples are i.i.d.

$$D \equiv \{(x^1, y_1), (x_2^2, y_2) \ldots (x^N, y_N)\}$$

Generative $P(x, y|\theta)$

log. likelihood of D given $P(x, y|\theta)$

$$\equiv LL(\theta, D) = \sum_{i=1}^{N} \log P(x_i^i, y_i|\theta)$$

During training, find θ for which $LL(\theta, D)$ is maximized.

$$\hat{\theta}^{ML} = \arg\max_{\theta} LL(\theta, D)$$

Apply ML training for LDA

$P(y=1) = \pi_1$, $P(y=2) = \pi_2$, μ_1, μ_2, Σ

$$\theta \equiv \{\pi_1, \pi_2, \underset{d\times 1}{\mu_1}, \underset{d\times 1}{\mu_2}, \underset{d\times d}{\Sigma}\}$$

$$|\theta| = d^2 + 2d + 2$$

$$LL(\theta, D) = \sum_{i=1}^{N} \log P(x^i, y_i|\theta)$$

$$= \sum_{i=1}^{N} \log P(y_i) P(x^i|y_i, \theta)$$

$$= \sum_{i=1, y_i=1}^{N} \log \pi_1 + \log \frac{1}{\sqrt{2\pi}|\Sigma|^{d/2}} e^{-\frac{1}{2}(x_i^i - \mu_1)^T \Sigma^{-1}}$$

$$+ \sum_{i=1, y_i=2}^{N} \log \pi_2 + \log \frac{1}{\sqrt{2\pi}|\Sigma|^{d/2}} e^{-\frac{1}{2}(x_i^i - \mu_2)^T \Sigma^{-1}}$$

$$\therefore N_1 \log \pi_1 + \sum_{\substack{i=1 \\ y_i=1}}^{N} \left(-\frac{1}{2}\right)(x^i - \mu_1)^T \Sigma^{-1} (x^i - \mu_1)$$

$$+ \sum_{\substack{i=1 \\ y_i=2}}^{N} \left(-\frac{1}{2}\right)(x^i - \mu_2)^T \Sigma^{-1} (x^i - \mu_2)$$

$$+ N_2 \log \pi_2 - \frac{N d}{2} \log |\Sigma| + \text{const}^n.$$

$$\therefore \max_{\pi_1, \pi_2, \mu_1, \mu_2, \Sigma} LL(\pi_1, \pi_2, \mu_1, \mu_2, \Sigma) \quad \text{s.t. } \pi_1 + \pi_2 = 1$$

∇. with each parameter $\delta = 0$ if LL is concave.