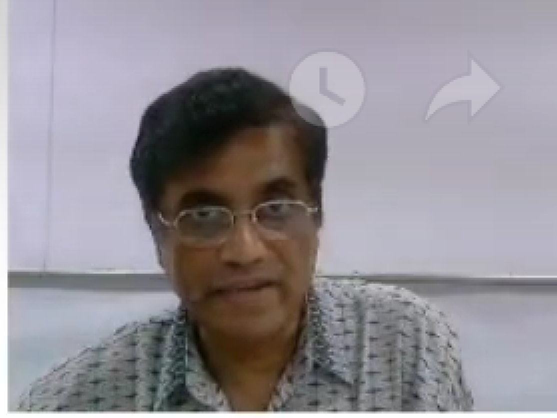
Gaussian Discriminant Analysis Model



Writing the probability density functions:

1) $$P(Y^{(i)}) = \phi^{Y^{(i)}} \cdot (1-\phi)^{(1-Y^{(i)})} .$$

2) $$P(X/Y;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} \cdot |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(X-\mu)^{T}\,\Sigma^{-1}(X-\mu)\right)$$

$|\Sigma|^{\frac{1}{2}}$ is a real number, $(X_j^{(i)}-\mu)$ is $m \times n$ dimensional matrix

$i = 1, 2, \dots m$

$j = 1, 2, 3 \dots n$

$(X-\mu)^{T}$ is $n \times m$ dimensional matrix.

Σ^{-1} is a $n \times n$ dimensional matrix.

$$X = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & X_3^{(1)} & \cdots & X_n^{(1)} \\ X_1^{(2)} & X_2^{(2)} & X_3^{(2)} & \cdots & X_n^{(2)} \\ \vdots & & & & \\ X_1^{(m)} & X_2^{(m)} & X_3^{(m)} & \cdots & X_n^{(m)} \end{bmatrix}_{m \times n}$$

2.00

Gaussian Discriminant Analysis Model

$$L(\mu,\sigma / x^{(i)}) = L(\mu,\sigma / x^{(1)}) \times L(\mu,\sigma / x^{(2)}) \times \cdots\cdots L(\mu,\sigma / x^{m}).$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} \times \cdots\cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(m)}-\mu)^2}{2\sigma^2}} \quad \text{———} (1)$$

For getting maximum likelihood estimate of μ, first take 'log' of both side of (1) $\Rightarrow$

$$\log L(\mu,\sigma / x^{(i)}) = l(\mu,\sigma / x^{(i)}), \text{say} = \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} \times \cdots\cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(m)}-\mu)^2}{2\sigma^2}}\right] \quad \text{———} (2)$$

For maximum likelihood estimate of μ & σ we need to do the following:

$$\frac{\partial l}{\partial \mu} = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{m} x^{(i)}}{m} \quad \text{———} (3)$$

$$\frac{\partial l}{\partial \sigma} = 0 \Rightarrow \sigma^2 = |\Sigma| = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T \quad \text{———} (4)$$

20:17 / 30:47 • Prediction >

Bilingual

2x

Gaussian Discriminant Analysis Model

$$L(\mu,\sigma / x^{(i)}) = L(\mu,\sigma / x^{(1)}) \times L(\mu,\sigma / x^{(2)}) \times \cdots\cdots L(\mu,\sigma / x^{m}).$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} \times \cdots\cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(m)}-\mu)^2}{2\sigma^2}} \quad \text{———} (1)$$

For getting maximum likelihood estimate of μ, first take 'log' of both side of (1) $\Rightarrow$

$$\log L(\mu,\sigma / x^{(i)}) = l(\mu,\sigma / x^{(i)}), \text{say} = \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(1)}-\mu)^2}{2\sigma^2}} \times \cdots\cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(m)}-\mu)^2}{2\sigma^2}}\right] \quad \text{———} (2)$$

For maximum likelihood estimate of μ & σ we need to do the following:

$$\frac{\partial l}{\partial \mu} = 0 \Longrightarrow \mu = \frac{\sum_{i=1}^{m} x^{(i)}}{m} \quad \text{———} (3)$$

$$\frac{\partial l}{\partial \sigma} = 0 \Longrightarrow \sigma^2 = |\Sigma| = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \quad \text{———} (4)$$

Gaussian Discriminant Analysis Model

Now for $Y^{(i)}$ the probability density function of Bernoulli is given by

$$P(Y^{(i)};\Phi) = \Phi^{Y^{(i)}} \cdot (1-\Phi)^{(1-Y^{(i)})}, \text{ where } Y^{(i)} \in \{0,1\}$$

$$= 0 \quad \text{Elsewhere.}$$

How to calculate the maximum likelihood estimate of Φ from given m samples?

As earlier, $L(\Phi) = \prod_{i=1}^{m} \Phi^{Y^{(i)}} \cdot (1-\Phi)^{(1-Y^{(i)})}$ ———— (5)

Taking 'log' of both sides of (5) $\Longrightarrow$

$$\log L(\Phi) = l(\Phi) = \log \prod_{i=1}^{m} \Phi^{Y^{(i)}} \cdot (1-\Phi)^{1-Y^{(i)}}$$

For maximum likelihood estimate of Φ

$$\frac{dl(\Phi)}{d\Phi} = 0 \Longrightarrow \hat{\Phi} = \frac{1}{m}\sum_{i=1}^{m} Y^{(i)} \quad \text{———— (6)}$$

Writing (6), (4) & (3) using 'indicator function' we get all the model parameters required for prediction.

$$\hat{\Phi} = \frac{1}{m}\sum_{i=1}^{m} 1\{Y^{(i)}=1\}; \quad \hat{\mu}_0 = \frac{\sum_{i=1}^{m} 1\{Y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} 1\{Y^{(i)}=0\}}; \quad \hat{\mu}_1 = \frac{\sum_{i=1}^{m} 1\{Y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} 1\{Y^{(i)}=1\}} \quad \text{———— (7)}$$

$$|\hat{\Sigma}_1| = |\hat{\Sigma}_2| = \frac{1}{m}\sum_{i=1}^{m} \left(x^{(i)} - \mu_{Y^{(i)}}\right)\left(x^{(i)} - \mu_{Y^{(i)}}\right)^T$$

1.00

Naïve Bayes Formula

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

$P(Y/X) =$ The posterior Probability of class Y, given X-features/attributes.

$P(Y) =$ Class Prior (or the prior probability of class)

$P(X/Y) =$ Likelihood of a feature to be in a class

$P(X) =$ Prior probability of feature

$X = x_1, x_2, x_3 \cdots x_n$

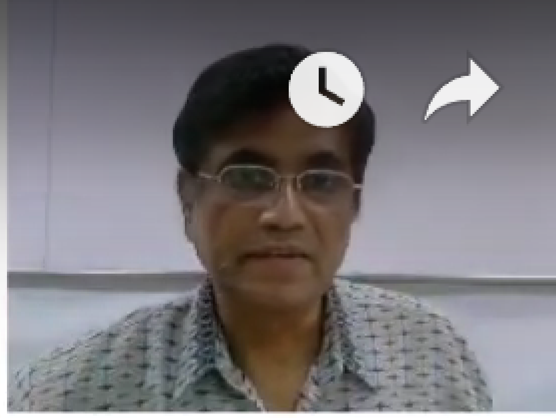
$Y \in \{0, 1\}$

How to calculate $P(X/Y)$ when $X = x_1, x_2, x_3 \cdots x_n$?

$$P(x_1, x_2, \cdots x_n / Y) = P(x_1/Y)\, P(x_2/Y, x_1)\, P(x_3/Y, x_1, x_2) \cdots P(x_n/Y, x_1, x_2 \cdots x_{n-1}).$$

Assuming conditional independence of X

$$\Rightarrow P(x_1/Y) \cdot P(x_2/Y) \cdots P(x_n/Y).$$ (Naive Bayes assumption)

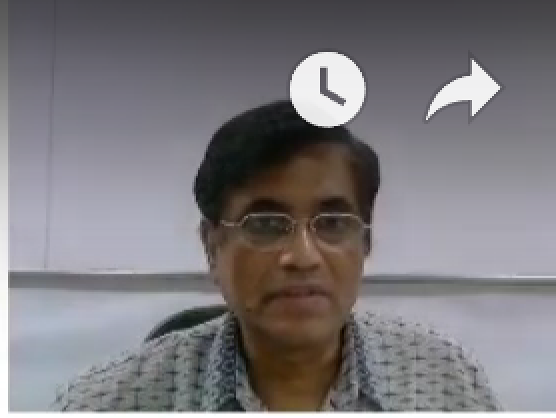


Generative Vs Discriminative Models

- In Discriminative Models-
- it learns from the training samples directly $P(Y/X)$:
- e.g in Linear Regression $P(Y/X;\ w) = W^T X = W_0 X_0 + W_1 X_1 + W_2 X_2 + \cdots + W_n X_n$

in Logistic Regression $P(Y/X;w) = g(W^T X) = g(W_0 X_0 + W_1 X_1 + \cdots + W_n X_n)$

- In case of Generative models-
- it learns $P(X/Y)$ i.e the probability distribution of features given a class
- & class prior $P(Y)$ directly from the data.
- Once this is done then using Bayes rule we compute the posterior distribution on Y given X: $P(Y/X)$



Prediction from GDA model

Once you calculate all the model parameters from the samples, it is assumed the model is trained.

Now , given a X for predicting in which class it belongs, you need to calculate:

$P(Y/X)$ from Bayes rule :

$$P(Y/X) = \frac{P(X/Y) \cdot P(Y)}{P(X)} \rightarrow P(X/Y=1)\,P(Y=1) + P(X/Y=0) \cdot P(Y=0)$$

For Prediction, we use 'arg max' function and don't need to calculate $P(X)$

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

$$= \arg\max_{Y} P(X/Y)\,P(Y).$$