ML Mid-2 :-

* Bayes Decision Theory :- (Continuous Features).

→ Allowing actions (decisions) other than classification.

→ Loss function ⇒ States cost of each action taken.

Loss function:-

$\lambda(\alpha_i/\omega_j)$ ⇒ loss of taking action α_i when the state of nature is ω_j

$\Omega = \{\omega_1, \omega_2, \ldots \omega_c\}$ ⇒ set of states of nature (or) classes (or) Categories)

$A = \{\alpha_1, \alpha_2, \ldots \alpha_a\}$ ⇒ Set of possible actions.

Given a pattern 'x', function $\alpha(x)$ → action taken.

$\alpha : x \rightarrow \alpha(x)$.

How to find the best action?

$R(\alpha_i/x)$ ⇒ Conditional risk.

⇒ risk of taking action α_i when given pattern is 'x'.

Best action 'α_k' ⇒ $\min\{R(\alpha_1/x), R(\alpha_2/x), \ldots R(\alpha_a/x)\}$.
and min. Risk ⇒ Bayes risk.

$$R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j) \cdot P(\omega_j/x).$$

$$R = \int R(\alpha(x)/x) \cdot P(x)\, dx. \Rightarrow \text{(Should be Minimum)}$$

Two-category classification:-

α_1 → decide 'ω_1'

α_2 → decide 'ω_2'.

$\lambda_{ij} = \lambda(\alpha_i/\omega_j)$

$$R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j) \cdot P(\omega_j/x)$$

$\alpha_1 \Rightarrow R(\alpha_1/x) = \lambda_{11} \cdot P\left(\frac{\omega_1}{x}\right) + \lambda_{12} \cdot P\left(\frac{\omega_2}{x}\right)$

$\alpha_2 \Rightarrow R(\alpha_2/x) = \lambda_{21} \cdot P\left(\frac{\omega_1}{x}\right) + \lambda_{22} P\left(\frac{\omega_2}{x}\right)$

if $R(\alpha_1/x) < R(\alpha_2/x)$

then action α_1 : "decide ω_1" is taken

otherwise action α_2 : "decide ω_2" is taken

Eg:-

$\alpha_1 \rightarrow$ decide 'ω_1' (criminals)

$\alpha_2 \rightarrow$ decide 'ω_2' (innocent)

$$R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j) \cdot P(\omega_j/x)$$

$$\alpha_1 \Rightarrow R(\alpha_1/x) = \lambda_{11} \cdot P(\omega_1/x) + \lambda_{12} \cdot P\left(\frac{\omega_2}{x}\right)$$

$$= 0 + 10 \times \frac{P(\omega_2) \cdot P(x/\omega_2)}{P(x)}$$

$$= 10 \times \frac{0.5 \times 0.6}{0.7}$$

$$= \frac{30}{7}$$

$P(x) = P(\omega_1) \cdot P(x/\omega_1) + P(\omega_2) \cdot P(x/\omega_2)$

$P(x) = 0.5 \times 0.8 + 0.5 \times 0.6$

$= 0.5 \times 1.4$

$= 0.7$

$$\alpha_2 \Rightarrow R(\alpha_2/x) = \lambda_{21} \cdot P\left(\frac{\omega_1}{x}\right) + \lambda_{22} \cdot P\left(\frac{\omega_2}{x}\right)$$

$$= 1 \times \frac{P(\omega_1) \cdot P(x/\omega_1)}{P(x)} + 0$$

$$= \frac{0.5 \times 0.8}{0.7} = \frac{4}{7}$$

$\therefore$ 'α_2' action taken $\Rightarrow$ innocent

* Likelihood ratio :-

Likelihood ratio $\Rightarrow \dfrac{P(x/\omega_1)}{P(x/\omega_2)}$

Eg:-

$R(\alpha_1/x) = \lambda_{11} \cdot P\left(\frac{\omega_1}{x}\right) + \lambda_{12} \cdot P\left(\frac{\omega_2}{x}\right) + \lambda_{13} P\left(\frac{\omega_3}{x}\right) = 0 + 1 \times 0.4 + 2 \times 0.5 = 1.4$

$R(\alpha_2/x) = \lambda_{21} \cdot P\left(\frac{\omega_1}{x}\right) + \lambda_{22} \cdot P\left(\frac{\omega_2}{x}\right) + \lambda_{23} \cdot P\left(\frac{\omega_3}{x}\right) = 1 \times 0.1 + 0 + 2 \times 0.5 = 1.1$

$R(\alpha_3/x) = \lambda_{31} \cdot P\left(\frac{\omega_1}{x}\right) + \lambda_{32} \cdot P\left(\frac{\omega_2}{x}\right) + \lambda_{33} \cdot P\left(\frac{\omega_3}{x}\right) = 3 \times 0.1 + 10 \times 0.4 + 0 = 4.3$

$\therefore$ 'α_2' action is taken.

Reject :-

→ Actions :- assign a class label (or) reject.

→ When misclassification is costly, we can reject to classify.

* Naive Bayes classification :-

(See slides)

* Maximum likelihood parameter estimation :-

→ Here, we assume parameters are unknown but fixed.

Let the parameters we are trying to estimate be

$$\theta = (\mu, \Sigma)^t.$$

D → training set. Contains 'n' samples $x_1, x_2, \ldots, x_n$.

Likelihood of 'θ', $P(D/\theta) = \prod_{i=1}^{n} P(x_i/\theta)$.

Maximum-likelihood estimate ⇒ $\hat{\theta}$ that maximizes $P(D/\theta)$.

log-likelihood function,

$$l(\theta) = \ln P(D/\theta)$$

$$= \ln \prod_{i=1}^{n} P(x_i/\theta)$$

$$= \sum_{i=1}^{n} \ln(P(x_i/\theta))$$

Let parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$. (P-parameter).

∇_θ ⇒ gradient operator.

$$\nabla_\theta = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

t to classify.

be

t.

$\ldots x_n$.

).

imizes $P(D/\theta)$.

)t.

$$\Rightarrow \nabla_\theta l = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_p} \end{bmatrix}$$

For maximum-likelihood estimate, $\nabla_\theta l = 0$.

Estimate the parameters in 'θ'.

* <u>Maximum-likelihood estimation of Gaussian Distribution :-</u>

Let 'D' training set contains 'n' samples.

$$D = \{x_1, x_2, \ldots x_n\}.$$

$$P(D/\theta) = \prod_{i=1}^{n} P(x_i/\theta).$$

Gaussian Distribution, $\theta = (\mu, \sigma^2)$

$$L\left(\frac{x}{\theta}\right) = P\left(\frac{x}{\theta}\right) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Log-likelihood,

$$\ln P\left(\frac{x}{\theta}\right) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right).$$

$$LL(x/\theta) = -\ln\sigma - \frac{1}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Differentiate it with $\frac{\partial}{\partial\mu}$ and $\frac{\partial}{\partial\sigma}$ to maximise ~~estimate~~ 'θ'.

$$\frac{\partial}{\partial\mu}\left(LL(x/\theta)\right) = 0 + 0 - \frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{n}\cancel{2}(x_i-\mu) = 0$$

$$\Rightarrow \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0.$$

$$\sum_{i=1}^{n}(x_i-\mu) = 0.$$

$$\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mu = 0.$$

$$\Rightarrow \boxed{\mu = \frac{\sum_{i=1}^{n}x_i}{n}}$$

$$\frac{\partial}{\partial\sigma}\left(LL(x/\theta)\right) = \sum_{i=1}^{n} -\frac{1}{\sigma} + \sum_{i=1}^{n} \frac{2}{2\sigma^3}(x_i-\mu)^2 = 0.$$

$$\sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i-\mu)^2 = \sum_{i=1}^{n} 1$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = n.$$

$$\Rightarrow \boxed{\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

Univariate :-

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Multivariate :-

$$P(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\left((x-\mu)^t \cdot \Sigma^{-1}(x-\mu)\right)}$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\Sigma = \frac{\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^t}{n}$$

* Regression :-

→ Regression analysis investigates the relationship b/w two (or) more variables in non-deterministic fashion.

→ Linear regression attempts to model the relationship b/w two variables by fitting a linear equation to observed data.

→ Scatter plot helps to determine the strength of relationship b/w two variables

1

$y = b_0 + b_1 x$ ⇒ Linear regression model

$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$ ⇒ Y-intercept.

$$\boxed{b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$ ⇒ slope.

If $b_1 > 0$ ⇒ x (predictor) and y (target) have positive relationship. ($x\uparrow$ $y\uparrow$).

If $b_1 < 0$ ⇒ x (predictor) and y (target) have negative relationship ($x\downarrow$ $y\downarrow$).

$y = b_0 + b_1 x$. ⟶ predicted output.

Sum of squared error ⇒ $\boxed{\sum_{i=1}^{n}(\text{actual output} - \text{predicted output})^2}$

Eg:-

$$\begin{bmatrix} 1 & 6 & 4 & 11 \\ 1 & 8 & 5 & 15 \\ 1 & 12 & 9 & 25 \\ 1 & 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 30 \\ 50 \\ 7 \end{bmatrix}$$

$$\theta = (Z^T . Z)^{-1} . Z^T y$$

$$\cancel{\begin{bmatrix} 1 \\ 6 \\ 4 \\ 11 \end{bmatrix}}$$