Data mining midterm

20222421 JAEYOON SIM

$\boxed{\#1}$

(a) False any ~~kinds~~

(b) True

(c) True

(d) False ~~variety~~ → veracity

(e) False ~~complex~~

(f) False entire ~~population~~ → partitions

(g) True

(h) False ~~supervised~~ → unsupervised

(i) True

(j) True

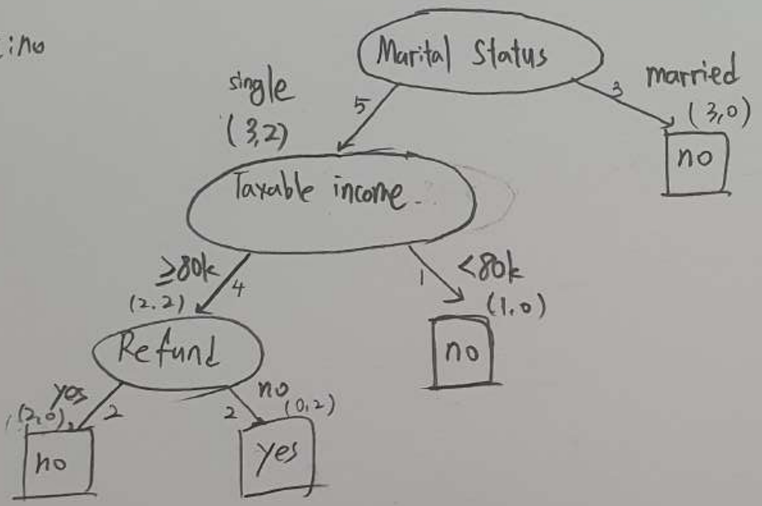
20222421 JAEYOON SJM

#2

index	Refund	Marital Status	Taxable Income	Cheat
1	yes	single	125K	no
2	no	married	100k	no
3	no	single	70k	no
4	yes	married	120k	no
5	no	single	95k	yes
6	no	married	60k	no
7	yes	single	220k	no
8	no	single	85k	yes

(a) Marital status → Taxable income (80K) → Refund

default: no



(b)

index 1 : yes

index 2 : no.

20222421 JAEYOON SIM

#3

$$N(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

(a) $L(\mu, \sigma \mid x) = \prod_{i=1}^{n} N(x_i \mid \mu, \sigma)$

$$= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right)$$

$$l(\mu, \sigma \mid x) = \ln \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right)$$

$$= \sum_{i=1}^{n} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right)$$

$$= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x_i-\mu)^2 \right)$$

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

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

(b) (i) μ_{mle}

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

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

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

$$\therefore \mu_{mle} = \frac{1}{N} \sum_{i=1}^{n} x_i$$

(ii) σ_{mle}

$$\frac{\partial l}{\partial \sigma^2} = 0 \Rightarrow \left(-\frac{1}{2}\right)\left(-\frac{1}{(\sigma^2)^2}\right) \sum_{i=1}^{n} (x_i-\mu)^2 - \frac{N}{2\sigma^2} = 0$$

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

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

$$\therefore \sigma_{mle}^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i-\mu_{mle})^2$$

$$\sigma_{mle} = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (x_i-\mu_{mle})^2}$$

(c) $\mathrm{Bias}[\hat{\mu}] = \mathbb{E}[\hat{\mu}] - \mu$

In this case, $\mathbb{E}[\hat{\mu}] = \mu$. So $\mathrm{Bias}[\hat{\mu}] = \mathbb{E}[\hat{\mu}] - \mu = \mu - \mu = 0$.

$\therefore$ μ_{mle} is unbiased estimator of μ.

2022421 JAEYOON SIM

#4

(a) Curse of dimensionality

→ When the dimensionality of data increases, then the data becomes increasingly sparse at their space.

(b) Additivity property of length

→ We can see the same ordinal attribute to different view. We can measure real length ① or we can just provide the ranking ②. ① has only order. but ② has order and additivity.

(c) Feature selection vs. Dimension Reduction

→ Feature selection is projecting feature space to lower dimension.

Dimension reduction is one way to reduce dimension and in this case, the models' accuracy can be good but lower comprehensible.

(d) Ridge regression vs. LASSO

→ LASSO is L1 penalty term and ridge regression is L2 penalty term of original error function to make the model robust.

LASSO → $\sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{j=1}^{n} x^{(d)} w_j\right)^2 + \lambda \sum_{j=1}^{n} |w_j|$

Ridge regression → $\sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{j=1}^{n} x^{(d)} w_j\right)^2 + \lambda \sum_{j=1}^{n} w_j^2$

LASSO tends to make the weight zero inherently performing feature selection.

Ridge regression shrinks the weight and this is not biased to concept of feature selection.

(e) Mahalanobis distance

→ Mahalanobis distance is one of dissimilarity measure and it uses the concept of correlation, linear-relation of data. This wants to see the trend of data.

Mahalanobis distance $= \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$ where Σ is covariance matrix

20222421 JAEYOON SIM

$\boxed{\#5}$

X (Income)	Y (owner)	P(X, Y)
High	yes	0.2
Mid	yes	0.3
Low	yes	0.15
High	no	0.1
Mid	no	0.2
Low	no	0.05

Notation : P_t is equal to $P(T=t)$ where $T=\{X, Y\}$, $t=\{High, Mid, low, yes, no\}$.

ex) $P_{High} = P(X=High)$

(a) marginal distribution

$\boxed{X}$ $P_{High} = 0.2+0.1=0.3$

$P_{Mid} = 0.3+0.2=0.5$

$P_{Low} = 0.15+0.05=0.2$

$\boxed{Y}$ $P_{yes} = 0.2+0.3+0.15 = 0.65$

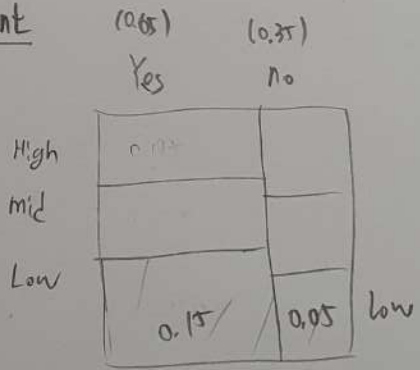
$P_{no} = 0.1+0.2+0.05 = 0.35$

(b) Independence $\iff$ $P(X,Y) = P(X)P(Y)$

$P(X=High, Y=yes) = P(X=High)\ P(Y=yes)$?

$$0.2 \neq \underset{\underset{0.195}{\parallel}}{0.3 \times 0.65}$$

$\therefore$ not independent



(c) $P(Y=Yes \mid X=Low) = \dfrac{P(X=Low \mid Y=Yes)\ P(Y=Yes)}{P(X=Low)}$

$$\boxed{\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}}$$

$$= \frac{0.15}{0.2} \quad \left(= \frac{\frac{0.15}{0.65} \times 0.65}{0.2} \right)$$

$$= \frac{15}{20}$$

$$= \frac{3}{4}$$

$$= 0.75$$