Webmining 2

2022年7月12日 星期二 10:08

- Bayesian Methods.
- Other Classification Approches.
- Accessing Model Performance
- Introduction of Clustering
- K-Means Clustering

BAYESIAN Methods

x = variables
y = target class

$P(y|x) \Rightarrow P(y=C_1|x)$
$P(y=C_2|x)$
$P(y=C_3|x)$

Bayes's therem (贝叶斯公式)

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$\underset{y\in\{C_1,C_2,C_3\}}{\arg\max} P(y|x)$$

$$P(x|y)P(y) = P(y)P(x_1|y)P(x_2|y,x_1)\cdots P(x_n|y,x_1,x_2,\cdots,x_{n-1})$$

Naïve BAYES CLASSIFIERS

先验概率：

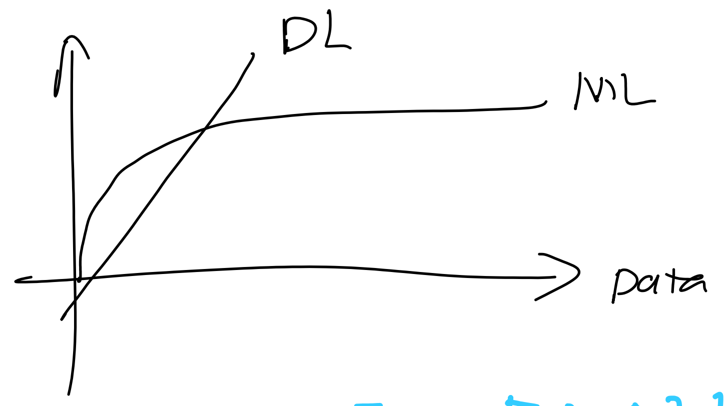
$P(y=\text{Yes}) = \frac{9}{14}$ $P(y=\text{No}) = \frac{5}{14}$

(conditional independence assumption/

(Naïve Bayes)

K-nearest neighbour.

ML: Training & Testing
decision tree
neural network



MODEL EVALUATION

VALIDATION SET Approach.

验证集，测试集 (50% + 50%)

K-FOLD CROSS VALIDATION

↳ K个测试集 → 验证

supervised learning

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} I(y_i=\hat{y}_i)$$

$$I(y_i=\hat{y}_i) = \begin{cases} 1 & \text{if } (y_i=\hat{y}_i) \\ 0 & \text{otherwise} \end{cases}$$

RESAMPLING METHODS

$S_{train} \longrightarrow S_{test}$

LEAVE-ONE-OUT

CROSS-VALIDATION

使用一个测试集，其余有 n-1 个测试集

K-MEANS CLUSTERING

$$\underset{C_1,\ldots,C_K}{\text{minimize}}\left\{\sum_{k=1}^{K} W(C_k)\right\}$$

$$W(C_k) = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$$