Sarang. D. Chandwadkar BE-A-24.

DA. Open Book Test Unit-4

Q1. Probability that John has Swine flu

→ a. As per Bayes theorem

$$P(C|A) = P(A|C) * \frac{P(C)}{P(A)}$$

where,

C = having Swine flu

A = Testing positive for Swine flu

Since,

$$P(A) = P(A \cap C) + P(A \cap \neg C)$$
$$= P(C) * P(A|C) + P(\neg C) * P(A|\neg C)$$

Using given data

$$P(A) = P(C) * P(A|C) + P(\neg C) * P(A|\neg C)$$
$$= 0.0002 * 0.99 + 0.9998 * (0.01)$$

$\therefore P(A) = 0.010196$

$\therefore P(C|A) = P(A|C) * P(C)/P(A)$

$= 0.99 * 0.0002 / 0.010196$

$\therefore P(C|A) = 0.0194$

∴ The probability of John having Swine flu given a positive result is 1.94%

Q1. b) How decision tree selects attributes for splitting?

→ b) 1. When deciding attributes to split on, a decision tree algorithm chooses most informative attribute which is

determined by the attribute with greatest information gain.

2 Information gain of an attribute is defined as difference between base entropy and conditional entropy of attribute

Information gain $_A$ = $H_S - H_{S/A}$

3. Information gain compares degree of purity of the parent node before a split with the degree of purity of child node after split, at each split, an attribute with greatest information gain is considered the most informative attribute

4 Information gain indicates purity of an attribute, the algorithm splits on attribute with largest information gain at each round

Detecting significant split:

- Necessary to measure significance of a split in a decision tree when information gain is small

N_A & N_B → Number of class A and class B in parent node

N_{AL} → Number of class A going to left child node

N_{BL} → Number of class B going to left child node

N_{AR} → Number of class A going to right child node

N_{BR} → Number of class B going to right child node

P_L & P_R → Proportion of data going to left and right node

$$P_L = \frac{N_{AL} + N_{BL}}{N_A + N_B}$$

$$P_R = \frac{N_{AR} + N_{BR}}{N_A + N_B}$$

$$K = \frac{(N'_{AL} - N_{AL})^2}{N''_{AL}} + \frac{(N'_{BL} - N_{BL})^2}{N'_{BL}} + \frac{(N'_{AR} - N_{AR})^2}{N'_{AR}} + \frac{(N'_{BR} - N_{BR})^2}{N'_{BR}}$$

Where,

$$N'_{AL} = N_A \times P_L$$
$$N'_{BL} = N_B \times P_L$$
$$N'_{AR} = N_A \times P_R$$
$$N'_{BR} = N_B \times P_R$$

If K is small, information gain from split is significant
If K is big, it would suggest the information gain from split is significant

Q1. c). which classifier is considered computationally efficient for high dimensional problems?

→ 1. c) ① Naives Bayes classifier should be used

② Naives Bayes assumption of 'conditional independence of each a_i.

- Naives Bayes assigns a classified label corresponds to largest value of $P(C_i/A)$

$$P(C_i/A) = \frac{P(a_1, a_2, \ldots, a_m/C_i) \cdot P(C_i)}{P(a_1, a_2, a_3, \ldots, a_m)}$$

$$i = 1, 2, 3, \ldots, m.$$

- It allows probabilities -

$$P(C_i/A) \propto P(C_i) \prod_{j=1}^{m} P(a_j/C_i) \quad i = 1, 2, 3 \ldots, m$$

to be calculated in straight forward manner which is computationally efficient.

③ The Naives Bayes classifier is simple to implement even without special libraries, the calculations are based on simply counting the occurence of events, making entire classifier efficient to run while handling high-dimensional data.