Naïve Bayes Classification

Bayes Theorem

- Hypothesis is that a given data belongs to class C. Without any information about the data, let the probability of any data belonging to C be P(C). This is called priori probability of the hypothesis.
- Given data *X,* posteriori probability of a hypothesis H, that is X is in class C, is written as P(H|X). It is given by the Bayes theorem

Bayes Theorem

- Consider the Bayes theorem

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- P(H|X) is Posterior probability of H conditioned on X.
  - i.e. X is an apple given that we have seen that X is red and round.
  - Eg. Suppose the data samples consists of fruits described by their color and shape. Suppose that X is red and round, and H is the hypothesis that X is an apple.
- P(X|H) is the posterior probability of X conditioned on H.
  - ie. The probability that X is red and round given that we know that it is true that X is an apple.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

P(H) is the prior probability of H.

Eg. The probability that any given data sample is an apple, regardless of how the data sample looks.

P(X) is the prior probability of X.

Eg. It is the probability that a data sample from our set of fruits is red and round.

RID	Age	Income	Student	Credit-rating	Class:buys-computer
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31..40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31..40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Fair	Yes
12	31..40	Medium	No	Excellent	Yes
13	31..40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

- X=(age="<=30", income = "medium", student = "yes", credit_rating="fair")

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- P(buys_computer = "yes") = 9/14 = 0.643
- P(buys_computer = "No") = 5/14 = 0.357
- $P(X/C_i)=P(x_1/c_1,\ x_2/c_2, x_3/c_3...)$
- P(age = "<30"| buys_computer = "yes") = 2/9 = 0.222
- P(age = "<30"| buys_computer = "No") = 3/5 = 0.600

- P(income = “medium”| buys_computer = “yes”) = 4/9 = 0.444
- P(income = “medium”| buys_computer = “no”) = 2/5 = 0.400
- P(student= “yes”| buys_computer = “yes”) = 6/9 = 0.444
- P(student= “yes”| buys_computer = “no”) = 1/5 = 0.200
- P(credit_rating= “fair”| buys_computer = “yes”) = 6/9 = 0.667
- P(credit_rating= “fair”| buys_computer = “no”) 2/5 = 0.400
- P(X| buys_computer = “yes”) = 0.222x0.444x0.444x0.667 = 0.044
- P(X| buys_computer = “no”) = 0.600x0.400x0.200x0.400= 0.019
- P(X| buys_computer = “yes”) P(buys_computer = “yes”) =
- = 0.044 x 0.643
- = 0.028
- P(X| buys_computer = “no”) P(buys_computer = no”) =
- = 0.019 x 0.357
- = 0.007

- The probability of 0.028 is high and hence the record is classified as buys computer.