NAÏVE BAYES

What is Bayesian Classification?

- Bayesian classifiers are statistical classifiers

- For each new sample they provide a probability that the sample belongs to a class (for all classes)

Bayes Classifier

- A probabilistic framework for solving classification problems

- Conditional Probability: is the probability that a random variable will take on a particular value given that the outcome for another random variable is known.

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

- Bayes theorem:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Prediction Based on Bayes' Theorem

- Given training data **X**, *posteriori probability of a hypothesis* H, P(H|**X**), follows the Bayes theorem

Likelihood

Prior

$$P(H\,|\,\mathbf{X}) = \frac{P(\mathbf{X}\,|\,H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}\,|\,H) \times P(H) / P(\mathbf{X})$$

Normalization Constant

$$P(H): \mathit{Independent}\ \text{probabilityof H : prior probability}$$

$$P(X): \mathit{Independent}\ \text{probabilityof X}$$

$$P(X|H): \mathit{Conditional}\ \ \text{probabilityof X given H :likelihood}$$

$$P(H|X): \mathit{Cond.}\ \ \text{probabilityof H given X : posterior probability}$$

Prediction Based on Bayes' Theorem

- Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause) \times P(Cause)}{P(Effect)}$$

- Informally, this can be written as

$$posteriori = \frac{likelihood \times prior}{evidence}$$

- Predicts **X** belongs to C_2 iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|X)$ for all the *k* classes

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what’s the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables

- Given a record with attributes (X_1, X_2,…, X_d)
  - Goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes $P(Y| X_1, X_2,\ldots, X_d)$

Using Bayes Theorem for Classification

- Approach:
  - Compute posterior probability $P(Y \mid X_1, X_2, \ldots, X_d)$ using the Bayes theorem

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

  - *Maximum a-posteriori*: Choose Y that maximizes
    $P(Y \mid X_1, X_2, \ldots, X_d)$

  - Equivalent to choosing value of Y that maximizes
    $P(X_1, X_2, \ldots, X_d \mid Y)\, P(Y)$

- How to estimate $P(X_1, X_2, \ldots, X_d \mid Y)$?

Properties of Bayes Classifiers

- Incrementality: with each training example, the prior and the likelihood can be updated dynamically: flexible and robust to errors.

- Combines prior knowledge and observed data: prior probability of a hypothesis multiplied with probability of the hypothesis given the training data.

- Probabilistic hypotheses: outputs not only a classification, but a probability distribution over all classes.

- Meta classification: the outputs of several classifiers can be combined, e.g., by multiplying the probabilities that all classifiers predict for a given class.

Bayesian Classification Method

- There are two implementation of Bayesian classification method
  - Naïve Bayes Classifier
  - Bayesian Belief Network

Naïve Bayes Classifier

- Assume independence among attributes X_i when class is given:

$$P(X_1, X_2, \ldots, X_d | Y_j) = P(X_1 | Y_j)\, P(X_2 | Y_j) \ldots P(X_d | Y_j)$$

  - Now we can estimate $P(X_i | Y_j)$ for all X_i and Y_j combinations from the training data

  - New point is classified to Y_j if $P(Y_j) \prod P(X_i | Y_j)$ is maximal.

Naïve Bayes Classifier (Dataset)

Binary Categorical Continuous Class

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Class: $P(Y) = N_c/N$

e.g., $P(No) = 7/10$,

$P(Yes) = 3/10$

Conditional Probabilities for Categorical Attributes

- For categorical attributes X_i
  - The conditional probability $P(X_i = x_i|Y = y)$ is estimated according to the fraction of training instances in class y that take on a particular attribute value x_i.

- Example

$$P(Martial\ Status = Single|Yes) = 2/3$$

$$P(Home\ Owner = Yes|No) = 3/7$$

Conditional Probabilities for Categorical Attributes

$$P(Home\ Owner = Yes|No) = 3/7$$

$$P(Home\ Owner = No|No) = 4/7$$

$$P(Home\ Owner = Yes|Yes) = 0$$

$$P(Home\ Owner = No|Yes) = 1$$

$$P(Marital\ Status\ = Single\ |No)$$

$$P(Marital\ Status\ = Divorced|No) = 1/7$$

$$P(Marital\ Status\ = Married|No) = 4/7$$

$$P(Marital\ Status\ = Single|Yes) = 2/3$$

$$P(Marital\ Status\ = Divorced|Yes) = 1/3$$

$$P(Marital\ Status = Married|Yes) = 0$$

Conditional Probabilities for Continuous Attributes

There are two ways to estimate the class-conditional probabilities

1. Discretize each continuous attribute and then replace the continuous attribute value with its corresponding discrete interval.
   1. Number of interval is too large, there are too few training records in each interval to provide a reliable estimate of conditional probability. If number of interval is too small, then some interval may aggregate records from different classes.
2. A Gaussian distribution is chosen to represent the class conditional probability for continuous attribute.

$$P(X_i = x_i | Y = y_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Conditional Probabilities for Continuous Attributes

- μ_{ij} is the sample mean of $(\bar{x})$ for all training records that belong to the class y_i

- σ_{ij}^2 sample variance (s^2) of training instances.

- Example, consider the attribute **annual income**

$$\bar{x} = \frac{125 + 100 + 70 + 120 + 60 + 220 + 75}{7} = 110$$

$$s^2 = \frac{(125 - 110)^2 + (100 - 110)^2 + (70 - 110)^2 + (120 - 110)^2 + (60 - 110)^2 + (220 - 110)^2 + (75 - 110)^2}{7}$$

$$s = 54.54$$

Conditional Probabilities for Continuous Attributes

- For Annual Income
  - If Class = No: Sample mean = 110,Sample Variance =2975
  - If Class = Yes: Sample mean = 90, Sample Variance = 25

- Given a test record with taxable income equal to $\$120K$, class conditional probability can be computed as follows:

$$P(Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} exp^{-\frac{(120-110)^2}{2*2975}}$$

$$= 0.0072$$

Test record $X = (Home\ Owner = No, Marital\ Status = Married, Income = \$120K)$, compute the posterior probability $\boldsymbol{P(No|X) and\ P(Yes|X)}$

Prior probabilities of each class (Yes and No)

P(Yes) = 0.3 and P(No) = 0.7

Conditional Probabilities for Continuous Attributes

$$
\begin{aligned}
P(X|No) \\
&= P(Home\ Owner = No|No) \\
&\times P(Status = Married|No) \\
&\times P(Annual\ Income = \$120K|No) \\
&= \frac{4}{7} \times \frac{4}{7} \times 0.0072 = 0.0024
\end{aligned}
$$

$$
\begin{aligned}
P(X|Yes) \\
&= P(Home\ Owner = No|Yes) \\
&\times P(Status = Married|Yes) \\
&\times P(Annual\ Income = \$120K|Yes) \\
&= 1 \times 0 \times 1.2 \times 10^{-9} = 0
\end{aligned}
$$

Conditional Probabilities for Continuous Attributes

The posterior probability for class No is $P(No|X) = 0.7 \times 0.0024 = 0.0016$

The posterior probability for class No is $P(Yes|X) = 0.3 \times 0 = 0$

$P(X|No)$>$P(X|Yes)$, the record is classified as No

Issues with Naïve Bayes Classifier

Test record $X = (Home\ Owner = Yes, Marital\ Status = Divorced, Income = \$120K)$, compute the posterior probability $\boldsymbol{P(No|X) and\ P(Yes|X)}$

$$P(X|No) = \frac{3}{7} \times 0 \times 0.0072 = 0$$

$$P(X|Yes) = 0 \times \frac{1}{3} \times 1.2 \times 10^{-9} = 0$$

Naïve Bayes will not be able to classify X as Yes or No!

- If one of the conditional probabilities is zero, then the entire expression becomes zero
  - Need to use other estimates of conditional probabilities than simple fractions

Issues with Naïve Bayes Classifier

- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability of the class (p=1/k, for k possible value of A_i)

m: parameter

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

Zero conditional probability

- Example: P(Marital Status=Married|Yes)=0
  - Adding ***m*** "virtual" examples (*m*: tunable but up to 1% of #training examples)
    - In this dataset, # of training examples for the "Yes" class is 3.
    - Assume that we add ***m*=1** "virtual" example in our m-estimate treatment.
  - The "Martial Status" feature can takes only 3 values. So ***p=1/3***.
  - Re-estimate P(Martial Status=Married|Yes) with the m-estimate

$$P(Marital\ Status = Married|Yes) = \frac{0 + 3 \times {}^{1}/_{3}}{3 + 3} = \frac{1}{6}$$

Zero conditional probability

$$
\begin{aligned}
P(X|No) \\
&= P(Home\ Owner = No|No) \\
&\times P(Martial\ Status = Married|No) \\
&\times P(Annual\ Income = \$120K|No) \\
&= \frac{6}{10} \times \frac{6}{10} \times 0.0072 = 0.0026
\end{aligned}
$$

$$
\begin{aligned}
P(X|Yes) \\
&= P(Home\ Owner = No|Yes) \\
&\times P(Status = Married|Yes) \\
&\times P(Annual\ Income = \$120K|Yes)
\end{aligned}
$$

$$= 4/6 \times 1/6 \times 1.2 \times 10^{-9} = 1.3 \times 10^{-10}$$

Zero conditional probability

The posterior probability for class No is $P(No|X) = \frac{7}{10} \times 0.0026 = 0.0018$

The posterior probability for class No is $P(Yes|X) = \frac{3}{10} \times 1.3 \times 10^{-10} = 4.0 \times 10^{-11}$

$\boldsymbol{P(X|No) > P(X|Yes)}$**, the record is classified as No**

Naïve Bayes Classifiers (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)