CS 60050
Machine Learning

Naïve Bayes Classifier

Some slides taken from course materials of Tan, Steinbach, Kumar

Bayes Classifier

- A probabilistic framework for solving classification problems
- Approach for modeling probabilistic relationships between the attribute set and the class variable
  - May not be possible to certainly predict class label of a test record even if it has identical attributes to some training records
  - Reason: noisy data or presence of certain factors that are not included in the analysis

Probability Basics

- P(A = a, C = c): joint probability that random variables A and C will take values a and c respectively
- P(A = a | C = c): conditional probability that A will take the value a, given that C has taken value c

$$P(C \mid A) = \frac{P(A, C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A, C)}{P(C)}$$

Bayes Theorem

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

- P(C) known as the prior probability
- P(C | A) known as the posterior probability

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?

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

Bayesian Classifiers

- Consider each attribute and class label as random variables

- Given a record with attributes $(A_1, A_2,\ldots,A_n)$
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes $P(C| A_1, A_2,\ldots,A_n)$

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

Class-conditional probability

Prior probability

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

Posterior probability

Evidence

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$

  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n|C)\ P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assumes all attributes A_i are conditionally independent, when class C is given:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C)\, P(A_2 | C) \ldots P(A_n | C)$
  - Can estimate $P(A_i | C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \prod P(A_i | C_j)$ is maximal.

Conditional independence: basics

- Let X, Y, Z denote three sets of random variables
- The variables in X are said to be conditionally independent of variables in Y, given Z if

$$P(X \mid Y, Z) = P(X \mid Z)$$

- An example
  - Level of reading skills of people tends to increase with length of the arm
  - Explanation: both increase with age of a person
  - If age is given, arm length and reading skills are (conditionally) independent

Conditional independence: basics

- If X and Y are conditionally independent, given Z

$$P(X, Y \mid Z) = P(X, Y, Z) / P(Z)$$
$$= P(X, Y, Z) / P(Y, Z) * P(Y, Z) / P(Z)$$
$$= P(X \mid Y, Z) * P(Y \mid Z)$$
$$= P(X \mid Z) * P(Y \mid Z)$$

$$P(X, Y \mid Z) = P(X \mid Z) * P(Y \mid Z)$$

NB assumption:

$$P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C)\, P(A_2 \mid C) \ldots P(A_n \mid C)$$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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: $P(C) = N_c/N$
  - e.g., $P(No) = 7/10$, $P(Yes) = 3/10$
- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_{c_k}$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

P(Status=Married|No) = 4/7
P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes, two options:
  - Discretize the range into bins
    - one ordinal attribute per bin
  - Probability density estimation:
    - Assume attribute follows a Gaussian / normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|c)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (A_i, c_j) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

A complete example

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

Training data:

Tid	Refund	Marital Status	Taxable Income	Evade
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

P(No) = 7/10 P(Yes) = 3/10

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0/3
P(Refund = No | Yes) = 3/3

P(Marital status = Single | No) = 2/7
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) = 0/3

For taxable income:
If class=No: sample mean = 110
sample variance = 2975
If class=Yes: sample mean = 90
sample variance = 25

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

P(No) = 7/10 **P(Yes) = 3/10**

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0/3
P(Refund = No | Yes) = 3/3

P(Marital status = Single | No) = 2/7
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) = 0/3

For taxable income:
If class=No: sample mean = 110
sample variance = 2975
If class=Yes: sample mean = 90
sample variance = 25

- P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Predicted Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- To prevent this, some variations of 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

m: parameter

Naïve Bayes: Pros and Cons

- Robust to isolated noise points

- Can handle missing values by ignoring the instance during probability estimate calculations

- Robust to irrelevant attributes

- Independence assumption may not hold for some attributes
  - Presence of correlated attributes can degrade performance of NB classifier

Example with correlated attribute

- Two attributes A, B and class Y (all binary)
- Prior probabilities:
  - $P(Y=0) = P(Y=1) = 0.5$
- Class conditional probabilities of A:
  - $P(A=0 \mid Y=0) = 0.4$ $\quad$ $P(A=1 \mid Y=0) = 0.6$
  - $P(A=0 \mid Y=1) = 0.6$ $\quad$ $P(A=1 \mid Y=1) = 0.4$
- Class conditional probabilities of B are same as that of A
- B is perfectly correlated with A when Y=0, but is independent of A when Y=1

Example with correlated attribute

- Need to classify a record with A=0, B=0
- $P(Y=0 \mid A=0,B=0)$ $$= \frac{P(A=0,B=0 \mid Y=0)\, P(Y=0)}{P(A=0,\, B=0)}$$
$$= \frac{P(A=0|Y=0)\, P(B=0|Y=0)\, P(Y=0)}{P(A=0,\, B=0)}$$
$$= (0.16 * 0.5) / P(A=0,B=0)$$
- $P(Y=1 \mid A=0,B=0)$ $$= \frac{P(A=0,B=0 \mid Y=1)\, P(Y=1)}{P(A=0,\, B=0)}$$
$$= \frac{P(A=0|Y=1)\, P(B=0|Y=1)\, P(Y=1)}{P(A=0,\, B=0)}$$
$$= (0.36 * 0.5) / P(A=0,B=0)$$
- Hence prediction is Y=1

Example with correlated attribute

- Need to classify a record with A=0, B=0
- In reality, since B is perfectly correlated to A when Y= 0
- $P(Y=0 \mid A=0,B=0)$

$$= \frac{P(A=0,B=0 \mid Y=0)\, P(Y=0)}{P(A=0,\, B=0)}$$

$$= \frac{P(A=0|Y=0)\, P(Y=0)}{P(A=0,\, B=0)}$$

$$= (0.4 * 0.5) / P(A=0,B=0)$$

- Hence prediction should have been Y=0

Other Bayesian classifiers

- If it is suspected that attributes may have correlations:

- Can use other techniques such as Bayesian Belief Networks (BBN)
  - Uses a graphical model (network) to capture prior knowledge in a particular domain, and causal dependencies among variables