Classification

Bayesian Classifiers

Bayesian classification

- A probabilistic framework for solving classification problems.
  – Used where class assignment is not deterministic, i.e. a particular set of attribute values will sometimes be associated with one class, sometimes with another.
  – Requires estimation of posterior probability for each class, given a set of attribute values:

$$p(C_i \mid x_1, x_2, \cdots, x_n) \quad \text{for each class } C_i$$

  – Then use decision theory to make predictions for a new sample **x**

Bayesian classification

- Conditional probability:

$$p(C \mid \mathbf{x}) = \frac{p(\mathbf{x}, C)}{p(\mathbf{x})} \qquad p(\mathbf{x} \mid C) = \frac{p(\mathbf{x}, C)}{p(C)}$$

- Bayes theorem:

$$p(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C)\, p(C)}{p(\mathbf{x})}$$

likelihood: $p(\mathbf{x} \mid C)$

prior probability: $p(C)$

posterior probability: $p(C \mid \mathbf{x})$

evidence: $p(\mathbf{x})$

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

- Treat each attribute and class label as random variables.

- Given a sample **x** with attributes ($x_1, x_2, \ldots, x_n$):
  - Goal is to predict class *C*.
  - Specifically, we want to find the value of C_i that maximizes $p(C_i \mid x_1, x_2, \ldots, x_n)$.

- Can we estimate $p(C_i \mid x_1, x_2, \ldots, x_n)$ directly from data?

Bayesian classifiers

Approach:

- Compute the posterior probability $p(C_i \mid x_1, x_2, \ldots , x_n)$ for each value of C_i using Bayes theorem:

$$p(C_i \mid x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n \mid C_i)\, p(C_i)}{p(x_1, x_2, \ldots, x_n)}$$

- Choose value of C_i that maximizes
  $p(C_i \mid x_1, x_2, \ldots , x_n)$
- Equivalent to choosing value of C_i that maximizes
  $p(x_1, x_2, \ldots , x_n \mid C_i)\, p(C_i)$

  (*We can ignore denominator – why?*)
- Easy to estimate priors $p(C_i)$ from data. (*How?*)
- The real challenge: how to estimate $p(x_1, x_2, \ldots , x_n \mid C_i)$?

Bayesian classifiers

- How to estimate $p(x_1, x_2, \ldots , x_n \mid C_i)$?
- In the general case, where the attributes x_j have dependencies, this requires estimating the full joint distribution $p(x_1, x_2, \ldots , x_n)$ for each class C_i.
- There is almost never enough data to confidently make such estimates.

Naïve Bayes classifier

- Assume independence among attributes x_j when class is given:

  $$p(x_1, x_2, \ldots , x_n \mid C_i) = p(x_1 \mid C_i)\, p(x_2 \mid C_i) \ldots p(x_n \mid C_i)$$

- Usually straightforward and practical to estimate $p(x_j \mid C_i)$ for all x_j and C_i.

- New sample is classified to C_i if

  $$p(C_i) \, \Pi \, p(x_j \mid C_i)$$

  is maximal.

How to estimate $p(x_j \mid C_i)$ 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 priors:
  $p(C_i) = N_i / N$
  - $p(\text{No}) = 7/10$
  - $p(\text{Yes}) = 3/10$
- For discrete attributes:
  $p(x_j \mid C_i) = |x_{ji}| / N_i$
  where $|x_{ji}|$ is number of instances in class C_i having attribute value x_j

Examples:

$p(\text{Status} = \text{Married} \mid \text{No}) = 4/7$

$p(\text{Refund} = \text{Yes} \mid \text{Yes}) = 0$

How to estimate $p(x_j \mid C_i)$ from data?

- For continuous attributes:
  - Discretize the range into bins
    - replace with an ordinal attribute
  - Two-way split: $(x_i < v)$ or $(x_i > v)$
    - replace with a binary attribute
  - Probability density estimation:
    - assume attribute follows some standard parametric probability distribution (usually a Gaussian)
    - use data to estimate parameters of distribution (e.g. mean and variance)
    - once distribution is known, can use it to estimate the conditional probability $p(x_j \mid C_i)$

How to estimate $p(x_j \mid C_i)$ 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

- Gaussian distribution:

$$P(x_j \mid C_i) = \frac{1}{\sqrt{2\pi\sigma_{ji}^2}} e^{-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2}}$$

  - one for each (x_j, C_i) pair

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

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

Example of using naïve Bayes classifier

Given a Test Record:

$$\mathbf{x} = (\text{Refund} = \text{No}, \text{Status} = \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes classifier:

p(Refund = Yes | No) = 3/7
p(Refund = No | No) = 4/7
p(Refund = Yes | Yes) = 0
p(Refund = No | Yes) = 1
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) = 2/7
p(Marital Status = Divorced | Yes) = 1/7
p(Marital Status = Married | Yes) = 0

For Taxable Income:
If Class = No: sample mean = 110
sample variance = 2975
If Class = Yes: sample mean = 90
sample variance = 25

- $p(\mathbf{x} \mid \text{Class} = \text{No}) = p(\text{Refund} = \text{No} \mid \text{Class} = \text{No}) \times p(\text{Married} \mid \text{Class} = \text{No}) \times p(\text{Income} = 120\text{K} \mid \text{Class} = \text{No}) = 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $p(\mathbf{x} \mid \text{Class} = \text{Yes}) = p(\text{Refund} = \text{No} \mid \text{Class} = \text{Yes}) \times p(\text{Married} \mid \text{Class} = \text{Yes}) \times p(\text{Income} = 120\text{K} \mid \text{Class} = \text{Yes}) = 1 \times 0 \times 1.2 \times 10^{-9} = 0$

$p(\mathbf{x} \mid \text{No})\, p(\text{No}) > p(\mathbf{x} \mid \text{Yes})\, p(\text{Yes})$

therefore $p(\text{No} \mid \mathbf{x}) > p(\text{Yes} \mid \mathbf{x})$

=> Class = No

Naïve Bayes classifier

- Problem: if one of the conditional probabilities is zero, then the entire expression becomes zero.
- This is a significant practical problem, especially when training samples are limited.
- Ways to improve probability estimation:

$$\text{Original}: p(x_j \mid C_i) = \frac{N_{ji}}{N_i}$$

$$\text{Laplace}: p(x_j \mid C_i) = \frac{N_{ji} + 1}{N_i + c}$$

$$\text{m-estimate}: p(x_j \mid C_i) = \frac{N_{ji} + mp}{N_i + m}$$

c: number of classes

p: prior probability

m: parameter

Example of Naïve Bayes classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

X: attributes

M: class = mammal

N: class = non-mammal

$$p(X \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$p(X \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$p(X \mid M)\, p(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$p(X \mid N)\, p(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$p(X \mid M)\, p(M) > p(X \mid N)\, p(N)$

=> mammal

Summary of naïve Bayes

- Robust to isolated noise samples.
- Handles missing values by ignoring the sample during probability estimate calculations.
- Robust to irrelevant attributes.
- *NOT* robust to redundant attributes.
  - Independence assumption does not hold in this case.
  - Use other techniques such as Bayesian Belief Networks (BBN).