Big Data Science

Naïve Bayes Classifier

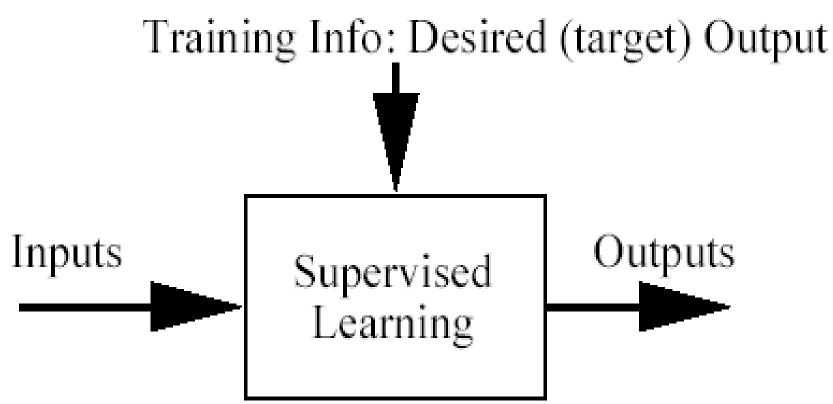
Incheon Paik

Contents

- Why Bayesian Classifier
- Bayes Theorem
- Naïve Bayes Classifier
- Another Example
- For Text Classification

Classification problem

- <u>Training data</u>: examples of the form (d,h(d))
  - where d are the data objects to classify (inputs)
  - and h(d) are the correct class info for d, h(d)$\in${1,…K}
- <u>Goal</u>: given d_{new}, provide h(d_{new})



Error = (target output - actual output)

Why Bayesian?

- Provides practical learning algorithms
  - E.g. Naïve Bayes
- Prior knowledge and observed data can be combined
- It is a generative (model based) approach, which offers a useful conceptual framework
  - E.g. sequences could also be classified, based on a probabilistic model specification
  - Any kind of objects can be classified, based on a probabilistic model specification

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

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

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

- Bayes theorem:

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

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,…,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,…,A_n)

- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

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

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n |C) = P(A_1| C_j) P(A_2| C_j)\ldots P(A_n| C_j)$
  - Can estimate $P(A_i| C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \Pi P(A_i| C_j)$ is maximal.

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:

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

  - 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:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption k
  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a 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_i) pair

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

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

Example of Naïve Bayes Classifier

Given a Test Record:

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

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(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)
=> Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- 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

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

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

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

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Naïve Bayes (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)

Naïve Bayes Classifier

- What can we do if our data *d* has several attributes?
- Naïve Bayes assumption: Attributes that describe data instances are conditionally independent given the classification hypothesis

$$P(\mathbf{d} \mid h) = P(a_1, \ldots, a_T \mid h) = \prod_t P(a_t \mid h)$$

  - it is a simplifying assumption, obviously it may be violated in reality
  - in spite of that, it works well in practice
- The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier
- One of the most practical learning methods
- Successful applications:
  - Medical Diagnosis
  - Text classification

Example. 'Play Tennis' data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Naïve Bayes solution

Classify any new datum instance $\mathbf{x}=(a_1,\ldots a_T)$ *as:*

$$h_{Naive\,Bayes} = \arg\max_h P(h)P(\mathbf{x}\,|\,h) = \arg\max_h P(h)\prod_t P(a_t\,|\,h)$$

- To do this based on training examples, we need to estimate the parameters from the training examples:
  - For each target value (hypothesis) *h*

    $$\hat{P}(h) := \text{estimate } P(h)$$

  - For each attribute value a_t of each datum instance

    $$\hat{P}(a_t\,|\,h) := \text{estimate } P(a_t\,|\,h)$$

Based on the examples in the table, classify the following datum ***x***:

x=(Outl=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

$$h_{NB} = \underset{h\in[yes,no]}{\arg\max}\, P(h)P(\mathbf{x}\,|\,h) = \underset{h\in[yes,no]}{\arg\max}\, P(h)\prod_t P(a_t\,|\,h)$$

$$= \underset{h\in[yes,no]}{\arg\max}\, P(h)P(Outlook = sunny\,|\,h)P(Temp = cool\,|\,h)P(Humidity = high\,|\,h)P(Wind = strong\,|\,h)$$

- Working:

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong\,|\,PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong\,|\,PlayTennis = no) = 3/5 = 0.60$$

$$etc.$$

$$P(yes)P(sunny\,|\,yes)P(cool\,|\,yes)P(high\,|\,yes)P(strong\,|\,yes) = 0.0053$$

$$P(no)P(sunny\,|\,no)P(cool\,|\,no)P(high\,|\,no)P(strong\,|\,no) = \mathbf{0.0206}$$

$$\Rightarrow answer : PlayTennis(x) = no$$

Learning to classify text

- Learn from examples which articles are of interest
- The attributes are the words
- Observe the Naïve Bayes assumption just means that we have a random sequence model within each class!
- NB classifiers are one of the most effective for this task
- Resources for those interested:
  - Tom Mitchell: Machine Learning (book) Chapter 6.

Naive Bayes Text Classifier

- Probability of class C for given some document vector

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})}$$

- P(x|c) approximation

$$P(\mathbf{x}|c) \approx \prod_{w_i \in \mathbf{X}} P(w_i|c) \times \prod_{w_i \notin \mathbf{X}} (1 - P(w_i|c))$$

$$P(c|\mathbf{x}) = \frac{P(c) \prod_{w_i \in \mathbf{X}} P(w_i|c) \times \prod_{w_i \notin \mathbf{X}} (1 - P(w_i|c))}{P(\mathbf{x})}$$

$$P(c) \prod_{w_i \in \mathbf{X}} P(w_i|c) \times \prod_{w_i \notin \mathbf{X}} (1 - P(w_i|c))$$

の大小でクラスを決定すればよいことになる．これが「ナイーブ・ベイズ分類器」である．

Naive Bayes Text Classifier

- Multivariate Bernoulli Model and Multinomial Model in Naive Bayes Classifier

Multinomial Model in Naive Bayes Classifier

$$P(\mathbf{x}|c) \quad = \quad P(|\mathbf{x}|)|\mathbf{x}|! \prod_i \frac{P(w_i|c)^{N(i,\mathbf{x})}}{N(i,\mathbf{x})!},$$

ここで, $P(|\mathbf{x}|)$ は長さ $|\mathbf{x}|$ の文書が起こる確率を表し, $N(i,\mathbf{x})$ は文書 $\mathbf{x}$ 内での単語 w_i の頻度を表す.
しかし、$|\mathbf{x}|$ や $N(i,\mathbf{x})$ は分類結果に影響しないので、モデル化の際は無視されることが多い。つまり、

$$P(\mathbf{x}|c) \quad \propto \quad \prod_i P(w_i|c)^{N(i,\mathbf{x})},$$

P(w|c) in Multinomial Model

$$P(w|c) \quad = \quad \frac{\text{クラス } c \text{ に属する訓練文書全体での } w \text{ の出現回数}}{\text{クラス } c \text{ に属する訓練文書全体での全単語の出現回数}}$$