innovate achieve lead

BITS Pilani

Machine Learning (IS ZC464) Session 5:

Bayes' Optimal Classifier, Gibbs Algorithm, Naïve Bayes' Classifier, Problem Solving on Bayesian Learning

Slides adapted from:

https://fenix.tecnico.ulisboa.pt/downloadFile/3779571251548/bayes2.ppt

www.cs.bu.edu/fac/gkollios/ada01/LectNotes/Bayesian.ppt

www.doc.ic.ac.uk/~yg/course/ida2002/ida-2002-5.ppt

https://cse.sc.edu/~rose/587/PPT/NaiveBayes.ppt

Review

- Prior and posterior probabilities
- Conditional probabilities
- Bayes' theorem
- MAP hypothesis
- Minimum Description Length Principle
- Entropy

Basic Approach

Bayes Rule:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D (posterior probability)
- P(D|h) = probability of D given h (likelihood of D given h)

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis h_{map})

$$\begin{aligned} & h_{map} \\ & = \max_{h \in H} P(h \mid D) \\ & = \max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ & = \max_{h \in H} P(D \mid h)P(h) \end{aligned}$$

MAP Learner

For each hypothesis h in H, calculate the posterior probability

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Output the hypothesis h_{map} with the highest posterior probability

$$h_{map} = \max_{h \in H} P(h \mid D)$$

- Computational intensive
- Provides a standard for judging the performance of learning algorithms
- Choosing P(h) and P(D|h) reflects our prior knowledge about the learning task

Entropy

- Based on the probability of each source symbol to be communicated, the Shannon entropy *H*, in units of bits (per symbol), is given by

$$Entropy = -\sum_i p_i \log_2 (p_i)$$

- where p_i is the probability of occurrence of the i^{th} possible value of the source symbol.

Encoding

- Less number of bits to represent frequent symbols
- Use more bits to represent less frequent symbols

symbol	Probability (p_i)	Code (unoptimized)	Code (Optimized)
a1	0.4	00	1
a2	0.25	01	010
a3	0.3	10	00
a4	0.05	11	011

Computation of entropy

- Entropy of the given data

$$Entropy = -\sum_i p_i \log_2 (p_i)$$

$$Entropy = -p_1 \log_2 (p_1) - p_2 \log_2 (p_2) - p_3 \log_2 (p_3) - p_4 \log_2 (p_4)$$

= - 0.4*$\log_2$(0.4) - 0.25*$\log_2$(0.25)- 0.3*$\log_2$(0.3) - 0.05*$\log_2$(0.05)

= 0.52877 + 0.5 + 0.521089 + 0.216096

=**1.76** bits (information content)

Average number of bits

Variable length encoding

= 1*0.4 + 3*0.25 + 2*0.3 + 3*0.05 = 1.9 bits

Fixed length encoding = 2 bits

symbol	Probability (p_i)	Code (unoptimized)	Code (Optimized)
a1	0.4	00	1
a2	0.25	01	010
a3	0.3	10	00
a4	0.05	11	011

MAP definition in logarithmic terms

$$h_{MAP} = \underset{h \in H}{Arg\max} P(D|h)P(h)$$

$$h_{MAP} = \underset{h \in H}{Arg\max} \log_2 P(D|h) + \log_2 P(h)$$

$$h_{MAP} = \underset{h \in H}{Arg\min} -\log_2 P(D|h) - \log_2 P(h)$$

MAP in the light of information theory

$$h_{MAP} = Arg\min_{h \in H} -\log_2 P(D|h) - \log_2 P(h)$$

- $-\log_2 P(h)$ is the description length of h under the optimal encoding for the hypothesis space H
- This is the size of the description of hypothesis h using this optimal representation
- $\log_2 P(D|h)$ is the description length of the training data D given hypothesis h under its optimal encoding.
- The Minimum Description length (MDL) principle recommends choosing the hypothesis that minimizes the sum of these two description lengths

Brute-Force Bayes' Concept Learning using MAP

- For each hypothesis *h* in *H*, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = _{h \in H} P(h|D)$$

Classification tasks (Concept learning)

- Examples
  - Spam Classification: Given an email, predict whether it is spam or not
  - Medical Diagnosis: Given a list of symptoms, predict whether a patient has disease X or not
  - Weather: Based on temperature, humidity, etc… predict if it will rain tomorrow
- What is classification task in machine learning?
  - Given training data and its class label C, features representing each training sample (say $X=<x_1,\ldots,x_k>$)are extracted. This represents $P(X|C)$
  - An unseen (or seen) sample is required to be classified. Let this be represented by its features $X=<x_1,\ldots,x_k>$.
  - The classification task requires the machine learning system to obtain the class label to which the sample might belong to.
- How is probability used in classification?
  - To classify we compute $P(C|X)$

Bayesian classification

- The classification problem may be formalized using a-posteriori probabilities:
- P(C|X) = prob. that the sample tuple $X=<x_1,\ldots,x_k>$ is of class C.
- Idea: assign to sample X the class label C such that P(C|X) is maximal

Techniques for classification based on posterior probabilities

- Maximum a Posteriori (MAP)
  - Maximum likelihood
  - Only one hypothesis contributes
- Bayes' Optimal Classifier
  - Weighted Majority Classifier
  - All hypotheses contribute
  - Costly classifier
- Gibbs Algorithm
  - Any one randomly picked hypothesis
  - Less costly
- Naïve Bayes' Classifier
  - uses assumption that the attributes are conditionally independent

Bayes' optimal Classifier: *A weighted majority classifier*

- What is he most probable classification of the new **instance** given the training data?
  - The most probable classification of the new instance is obtained by combining the prediction of *all hypothesis*, weighted by their *posterior probabilities*
- If the classification of new example can take any value v_j from some set V, then the probability $P(v_j|D)$ that the correct classification for the new **instance is v_j**, is just:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Bayes' Optimal Classifier

- Question: Given new instance x, what is its most probable classification?
- $H_{map}(x)$ is not the most probable classification sometimes in different situations.

Example: Let P(h1|D) = .4, P(h2|D) = .3, P(h3 |D) =.3

Given new data x, we have h1(x)=+, h2(x) = -, h3(x) = -

What is the most probable classification of x ?

Bayes optimal classification:

$$\max_{vj \in V} \sum_{hj \in H} P(v_j \mid h_i) P(h_i \mid D)$$

Example:

P(h1| D) =.4, P(-|h1)=0, P(+|h1)=1
P(h2|D) =.3, P(-|h2)=1, P(+|h2)=0
P(h3|D)=.3, P(-|h3)=1, P(+|h3)=0

$$\sum_{hi \in H} P(+ \mid h_i) P(h_i \mid D) = .4$$

$$\sum_{hi \in H} P(- \mid h_i) P(h_i \mid D) = .6$$

Naïve Bayes' Learner

Assume target function f: X-> V, where each instance x described by attributes <a1, a2, …., an>. Most probable value of f(x) is:

$$v = \max_{vj \in V} P(v_j \mid a_1, a_2 ....a_n)$$

$$= \max_{vj \in V} \frac{P(a_1, a_2 ....a_n \mid v_j) P(v_j)}{P(a_1, a_2 ....a_n)}$$

$$= \max_{vj \in V} P(a_1, a_2 ....a_n \mid v_j) P(v_j)$$

Naïve Bayes assumption:

$$P(a_1, a_2 ....a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$ (attributes are conditionally independent)

Naïve Bayes Classifier (I)

- A simplified assumption: attributes are conditionally independent:

$$P(C_j \mid D) \propto P(C_j) \prod_{i=1}^{n} P(d_i \mid C_j)$$

- Greatly reduces the computation cost, only count the class distribution.

Play-tennis example: estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14
P(n) = 5/14

outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 2/5**
windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

Naive Bayesian Classifier (II)

- Given a training set, we can compute the probabilities

Outlook	P	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	P	N
high	3/9	4/5
normal	6/9	1/5
Windy		
true	3/9	3/5
false	6/9	2/5

Play-tennis example: classifying X

- An **unseen sample X = <rain, hot, high, false>**

- $P(X|p)\cdot P(p) = P(rain|p)\cdot P(hot|p)\cdot P(high|p)\cdot P(false|p)\cdot P(p) = 3/9\cdot 2/9\cdot 3/9\cdot 6/9\cdot 9/14 = 0.010582$
- $P(X|n)\cdot P(n) = P(rain|n)\cdot P(hot|n)\cdot P(high|n)\cdot P(false|n)\cdot P(n) = 2/5\cdot 2/5\cdot 4/5\cdot 2/5\cdot 5/14 = 0.018286$

- Sample X is classified in class n (don't play)

The independence hypothesis…

- … makes computation possible
- … yields optimal classifiers when satisfied
- … but is seldom satisfied in practice, as attributes (variables) are often correlated.
- Attempts to overcome this limitation:
  - Bayesian networks, that combine Bayesian reasoning with causal relationships between attributes
- when conditional independence assumption is satisfied the naive Bayes classification is a MAP classification
-

Naïve Bayesian Classifier: Comments

- Advantages :
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence , therefore loss of accuracy
  - Practically, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history etc

    Symptoms: fever, cough etc., Disease: lung cancer, diabetes etc
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier
- How to deal with these dependencies?
  - Bayesian Belief Networks

Problem Solving on Bayesian Learning

Problem 1: Bayes' Theorem

- 90% students pass an examination
- 75% Students who study hard pass the exam
- 60% students study hard
- Let S: event that students pass the exam
- H : studies hard
- $P(S|H) = 0.75$
- $P(S) = 0.9$
- $P(H) = 0.6$
- $P(H|S) = ??$
- Solution : Use bayes' Theorem
- $P(H|S) = P(S|H)\,P(H)/P(S) = 0.75 \times 0.6 / 0.9 = 0.5$

Problem 2: Using Joint Probabilities

- P(cavity, toothache, catch) = 0.06
- P(cavity, toothache, ¬catch) = 0.19
- P(cavity, ¬ toothache, catch) = 0.05
- P(cavity, ¬ toothache, ¬ catch) = 0.10
- P(¬ cavity, toothache, catch) = 0.09
- P(¬ cavity, toothache, ¬catch) = 0.01
- P(¬ cavity, ¬ toothache, catch) = 0.22
- P(¬ cavity, ¬ toothache, ¬ catch) = 0.28

Compute
P(cavity)
P(cavity | toothache)
P(toothache)
P(Catch | cavity)

	toothache		**¬toothache**	
	Catch	¬catch	Catch	¬catch
Cavity	0.06	0.19	0.05	0.10
¬Cavity	0.09	0.01	0.22	0.28

Solution

	toothache		¬toothache	
	Catch	¬catch	Catch	¬catch
Cavity	0.06	0.19	0.05	0.10
¬Cavity	0.09	0.01	0.22	0.28

- P(cavity) = P(cavity,toothache)+P(cavity,~toothache)

(using Marginalization)

- P(cavity) = P(cavity,toothache, catch)+P(cavity,~toothache, catch)+ P(cavity,toothache, ~catch)+P(cavity,~toothache, ~catch) (using Marginalization)

  = 0.06 + 0.05 + 0.19 + 0.10 = 0.4

- Similarly

  P(toothache) = P(toothache, catch)+P(toothache, ~catch)

  = P(toothache, catch, cavity)+P(toothache, ~catch, cavity)

  + P(toothache, catch, ~cavity)+P(toothache, ~catch, ~cavity)

  (using Marginalization)

  = 0.06 + 0.19 + 0.09 + 0.01 = 0.35

Solution

	toothache		¬toothache	
	Catch	¬catch	Catch	¬catch
Cavity	0.06	0.19	0.05	0.10
¬Cavity	0.09	0.01	0.22	0.28

- P(cavity | toothache) = P(cavity, toothache) . P(toothache) (using Bayes' theorem)
- Where

  P(cavity, toothache) = P(cavity, toothache, catch) + P(cavity, toothache, ~catch)

  (using Marginalization Rule)

  = 0.06 + 0.19 = 0.25
- Therefore

  P(cavity | toothache) = P(cavity, toothache) . P(toothache)

  = 0.25 * 0.35 = 0.0875

Problem 3: Bayes' Theorem

- Suppose that Bob can decide to go to work by one of three modes of transport car, bus, or commuter train. Because of high traffic, if he decides to go by car, there is a 50% chance he will be late. If he goes by bus, which has special reserved lanes but is sometimes overcrowded, the probability of being late is only 20%. The commuter train is almost never late, with a probability of only 1%, but is more expensive than the bus.

Example source
http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-607/BayesEx.pdf

Example contd..

- Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car. Since he does not know which mode of transportation Bob usually uses, he gives a prior probability of 1/3 to each of the three possibilities. What is the boss' estimate of the probability that Bob drove to work?

Solution

$$
\begin{aligned}
Pr\{\text{ bus }\} = Pr\{\text{ car }\} = Pr\{\text{ train }\} &= \frac{1}{3} \\
Pr\{\text{ late } \mid \text{ car }\} &= 0.5 \\
Pr\{\text{ late } \mid \text{ train }\} &= 0.01 \\
Pr\{\text{ late } \mid \text{ bus }\} &= 0.2
\end{aligned}
$$

We want to calculate $Pr\{\text{ car } \mid \text{ late }\}$.

By Bayes Theorem, this is

$$
\begin{aligned}
& Pr\{\text{ car } \mid \text{ late }\} \\
= & \frac{Pr\{\text{ late } \mid \text{ car }\}Pr\{\text{ car }\}}{Pr\{\text{ late } \mid \text{ car }\}Pr\{\text{ car }\} + Pr\{\text{ late } \mid \text{ bus }\}Pr\{\text{ bus }\} + Pr\{\text{ late } \mid \text{ train }\}Pr\{\text{ train }\}} \\
= & \frac{0.5 \times 1/3}{0.5 \times 1/3 + 0.2 \times 1/3 + 0.01 \times 1/3} \\
= & \ 0.7042
\end{aligned}
$$

http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-607/BayesEx.pdf

Problem 4: Minimum Description Length (MDL) Principle

- Compute the MDL encoding for the problem given below

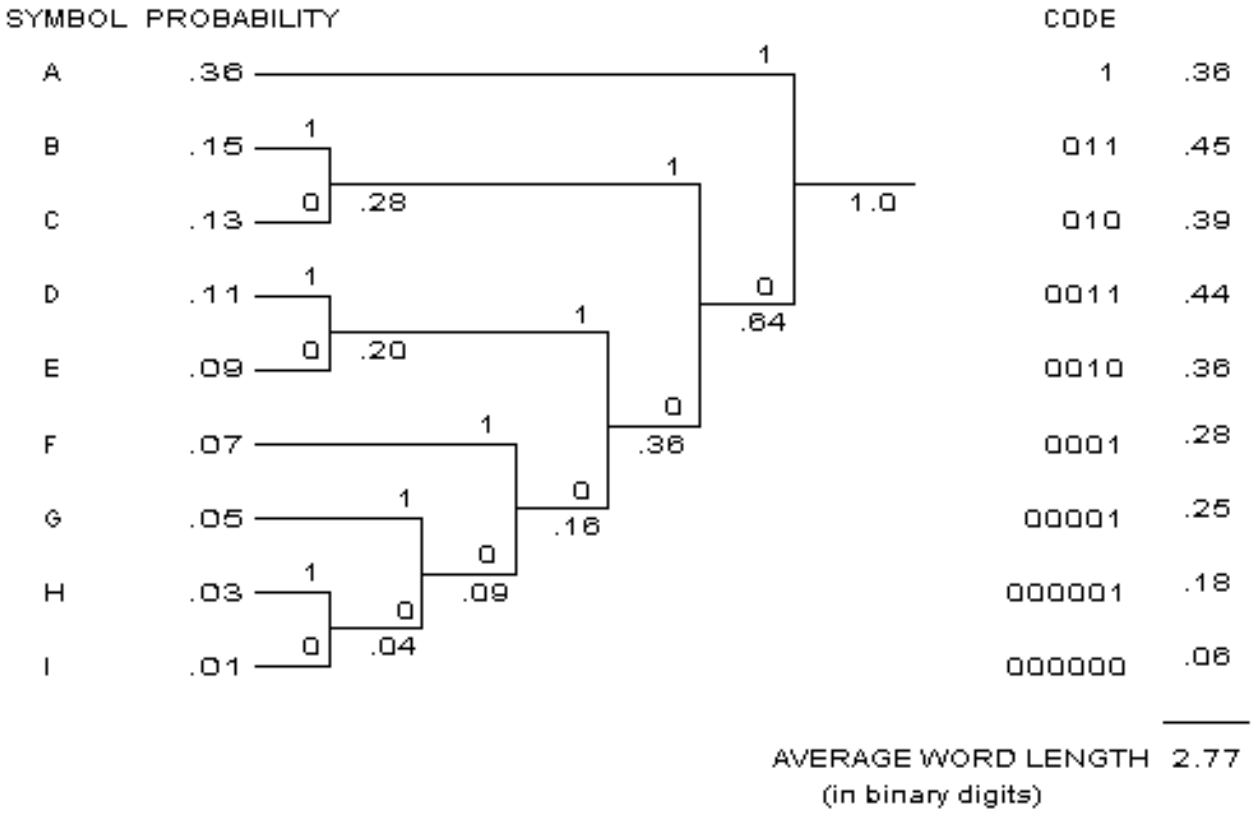
symbol	pi
A	0.36
B	0.15
C	0.13
D	0.11
E	0.09
F	0.07
G	0.05
H	0.03
I	0.01

Solution

- Arrange the symbols in sorted order
- Pair them by adding their probabilities and reach the end
- Assign smallest code to the symbol with highest probability
- Assign incremental length code to other symbols depending upon their probabilities
- Compute the total number of expected bits
- Compute the entropy of the given information

Probabilities and Codelengths

- Huffman coding



Source: http://star.itc.it/caprile/teaching/algebra-superiore-2001/

Problem 5: MAP classifier

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$$P(cancer) = .008, P(\neg cancer) = .992$$
$$P(+ \mid cancer) = .98, P(- \mid cancer) = .02$$
$$P(+ \mid \neg cancer) = .03, P(- \mid \neg cancer) = .97$$
$$P(cancer \mid +) = \frac{P(+ \mid cancer)P(cancer)}{P(+)}$$
$$P(\neg cancer \mid +) = \frac{P(+ \mid \neg cancer)P(\neg cancer)}{P(+)}$$

Problem 6: Naïve Bayes classifier

Class:
C1:buys_computer='yes'
C2:buys_computer='no'

Data sample:

X =
(age<=30,
Income=medium,
Student=yes
Credit_rating=Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayesian Classifier: Example

- Compute $P(X|C_i)$ for each class

P(age="<30" | buys_computer="yes") = 2/9=0.222
P(age="<30" | buys_computer="no") = 3/5 =0.6
P(income="medium" | buys_computer="yes")= 4/9 =0.444
P(income="medium" | buys_computer="no") = 2/5 = 0.4
P(student="yes" | buys_computer="yes)= 6/9 =0.667
P(student="yes" | buys_computer="no")= 1/5=0.2
P(credit_rating="fair" | buys_computer="yes")=6/9=0.667
P(credit_rating="fair" | buys_computer="no")=2/5=0.4

P(buys_computer=„yes“)=9/14

P(buys_computer=„no“)=5/14

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

$P(X|C_i)$: P(X|buys_computer="yes")= 0.222 x 0.444 x 0.667 x 0.0.667 =0.044
P(X|buys_computer="no")= 0.6 x 0.4 x 0.2 x 0.4 =0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer="yes") * P(buys_computer="yes")=0.028
P(X|buys_computer="no") * P(buys_computer="no")=0.007

- X belongs to class "buys_computer=yes"

Home Work

- A box contains 10 red and 15 blue balls. Two balls are selected at random and are discarded without their colors being seen. If a third ball is drawn randomly and observed to be red, what is the probability that both of the discarded balls were blue?
- Solution Hint:

$$\frac{P(R \mid BB)P(BB)}{P(R \mid BB)P(BB) + P(R \mid BR)P(BR) + P(R \mid RR)P(RR)}$$