Lecture 13 -- Naïve Bayes

Prof Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

(Attendance Code: **466202**)

Up to now,

A good connection between symbolic AI and machine learning

- Three machine learning algorithms:
    - decision tree learning
    - k-nn
    - linear regression + gradient descent

Also useful for un/semi-supervised learning

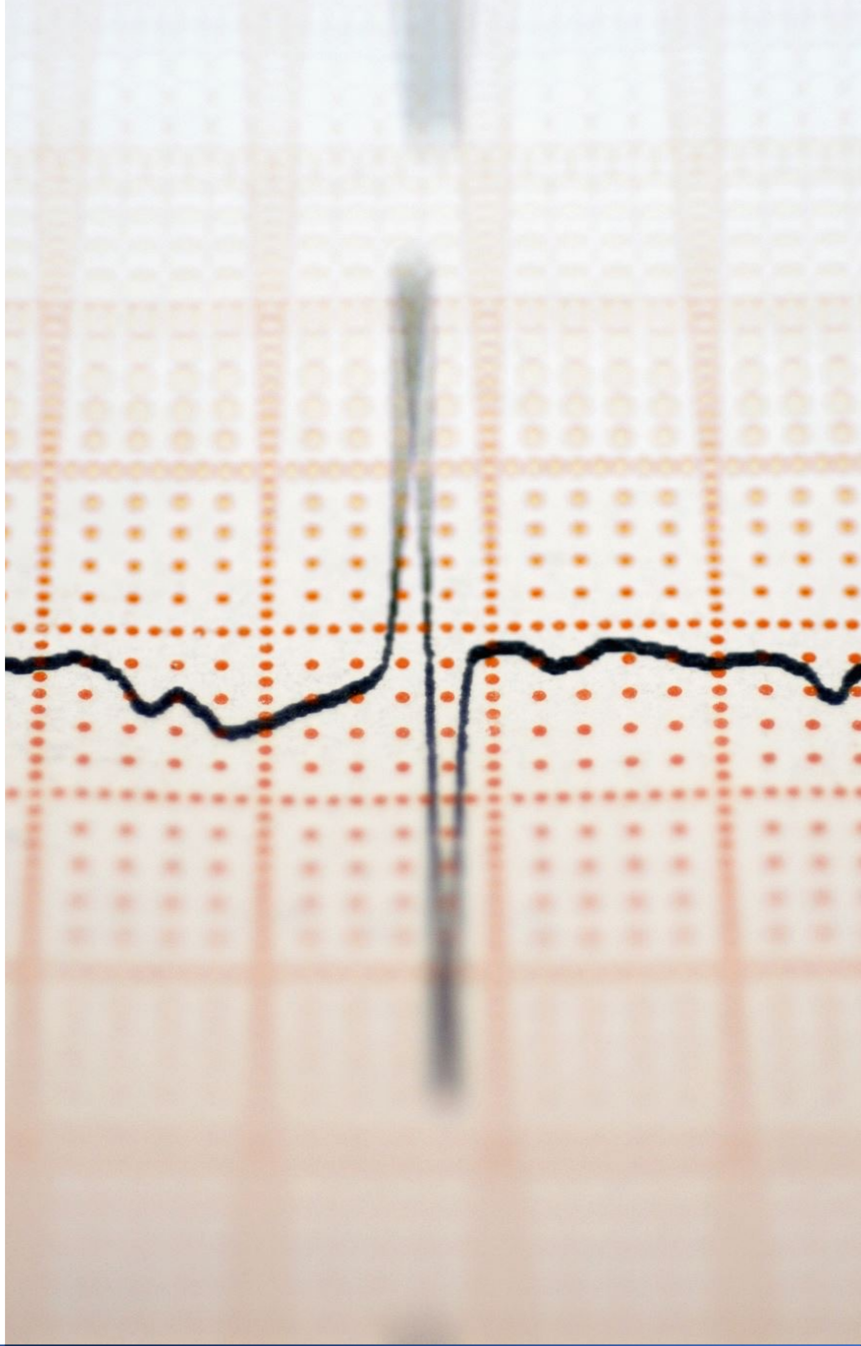
Learn how to design new objective function and how to adapt optimisation algorithm for new objective function

Topics

- $\boxed{\text{Naïve Bayes}}$
  - Theoretical Analysis on Computational Efficiency via Parameter Estimation
  - Naïve Bayes Algorithm

Bayesian learning is a key branch of machine learning

1. Preparation for graphical models;
2. Learn how assumptions can simply a complex problem and lead to useful conclusions



Naïve Bayes Algorithm

- Given data for X and Y, such that for each data instance *(x,y)*, *x* has *n* dimensions and *y* is a label.

- Goal: classify (x^{new})

Typical classification problem

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k | x^{new})$$

Bayes Rule

$$Y^{new} \leftarrow \arg\max_{y_k} \boxed{P(Y = y_k)} \boxed{P(x^{new} | Y = y_k)}$$

The essence of Naïve Bayes is to have an efficient method for the computation of this term by having an assumption.

This term can be easily estimated from the training dataset.

How to calculate these two terms?

Theorical Analysis on Computational Efficiency via Estimation of Parameters

Let's learn classifiers by learning P(Y|X)

- Consider Y=Wealth, X=<Gender, HoursWorked>

X_1 | X_2 | Y | P(gender, hoursWorked, wealth)

Gender	HoursWorked	Wealth	Probability
Female	<40.5	Poor	0.253122
		Rich	0.0245895
	>40.5	Poor	0.0421768
		Rich	0.0116293
Male	<40.5	Poor	0.331313
		Rich	0.0971295
	>40.5	Poor	0.134106
		Rich	0.105933

Can we estimate P(Y)? If so, how?

how many parameters need to estimate?

7

Let's learn classifiers by learning P(Y|X)

P(Y|X)

- P(gender, hoursWorked, wealth) => P(wealth|gender, hoursWorked)

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How about estimating P(X|Y)?

How many parameters to estimate?

4

How many parameters must we estimate?

- Suppose X =<X_1,... X_n> where X_i and Y are Boolean real variables
- To estimate $P(Y|X_1, X_2, \ldots X_n)$, how many quantitates need to be estimated or collected?

  2^n

- If we have 30 Boolean X_i's: $P(Y \mid X_1, X_2, \ldots X_{30})$

  2^{30} ~ 1 billion!

- You need lots of data or a very small *n*

n=2, and we need to estimate 4 quantities. Why 4?

These 4 are needed

These 4 can be obtained by 1-x

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

Idea: Shall we now consider P(X|Y)?

Can we reduce parameters using Bayes Rule?

- Suppose X =<X_1,... X_n> where X_i and Y are Boolean real variables
- By Bayes rule:

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

Y(X|Y)

Gender	HrsWorked	P(G,HW|rich)	P(G,HW|poor)
F	<40.5	0.1028	0.3327
F	>40.5	0.0486	0.0554
M	<40.5	0.4059	0.4355
M	>40.5	0.4427	0.1763

- How many parameters for P(X|Y) = P(X_1,... X_n|Y)?

$(2^n-1)x2$

For example, P(Gender,HrsWorked|Wealth)

- How many parameters for P(Y)?

1

For example, P(Wealth)

P(Y)

P(poor)	P(rich)
0.7607187	0.2392813

So, by now, we haven't achieved the reduction of parameters. P(Y|X) needs 2^n, while P(X|Y)P(Y) needs $(2^n-1)x2+1=2^{n+1}-1$. Any further idea?

Naïve Bayes Assumption

- Naïve Bayes assumes

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all $i \neq j$

For example,

P(Gender,HrsWorked|Wealth) = P(Gender|Wealth) * P(HrsWorked|Wealth)

Recap: Conditional independence

- Two variables A,B are *independent* if

$$P(A \wedge B) = P(A)P(B)$$

$$\forall a, b : P(A = a \wedge B = b) = P(A = a)P(B = b)$$

- Two variables A,B are *conditionally independent given* C if

$$\boxed{P(A \wedge B|C) = P(A|C)P(B|C)}$$

$$\forall a, b, c : P(A = a \wedge B = b|C = c) = P(A = a|C = c)P(B = b|C = c)$$

- We also have $\boxed{P(A|B, C) = P(A|C)}$

Two equivalent definitions.

Effect of Naïve Bayes Assumption

- Naïve Bayes uses assumption that the X_i are conditionally independent, given Y
- Given this assumption, then:

Chain rule

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$

$$= P(X_1|Y)P(X_2|Y)$$

Definition of Conditional Independence

- in general: $$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

$(2^n-1)\text{x}2$ $\rightarrow$ $2n$

Why? Every $P(X_i|Y)$ takes a parameter, and we have n X_i.

Summary of Naïve Bayes Algorithm

$$P(Y|X_1, ..., X_n) = \frac{P(X_1, ..., X_n|Y)P(Y)}{P(X_1, ..., X_n)}$$

- To make this tractable we naively assume conditional independence of the features given the class: ie

$$P(X_1, ..., X_n|Y) = P(X_1|Y)P(X_2|Y)...P(X_n|Y)$$

Step 1

- Now: I only need to estimate ... parameters:

$$P(X_1|Y), P(X_2|Y), ..., P(X_n|Y), P(Y)$$

Step 2

- Besides, we also need to estimate P(Y)

Summary of the Benefit of Naïve Bayes Algorithm

How many parameters to describe $P(X_1, ..., X_n|Y)$? $P(Y)$?

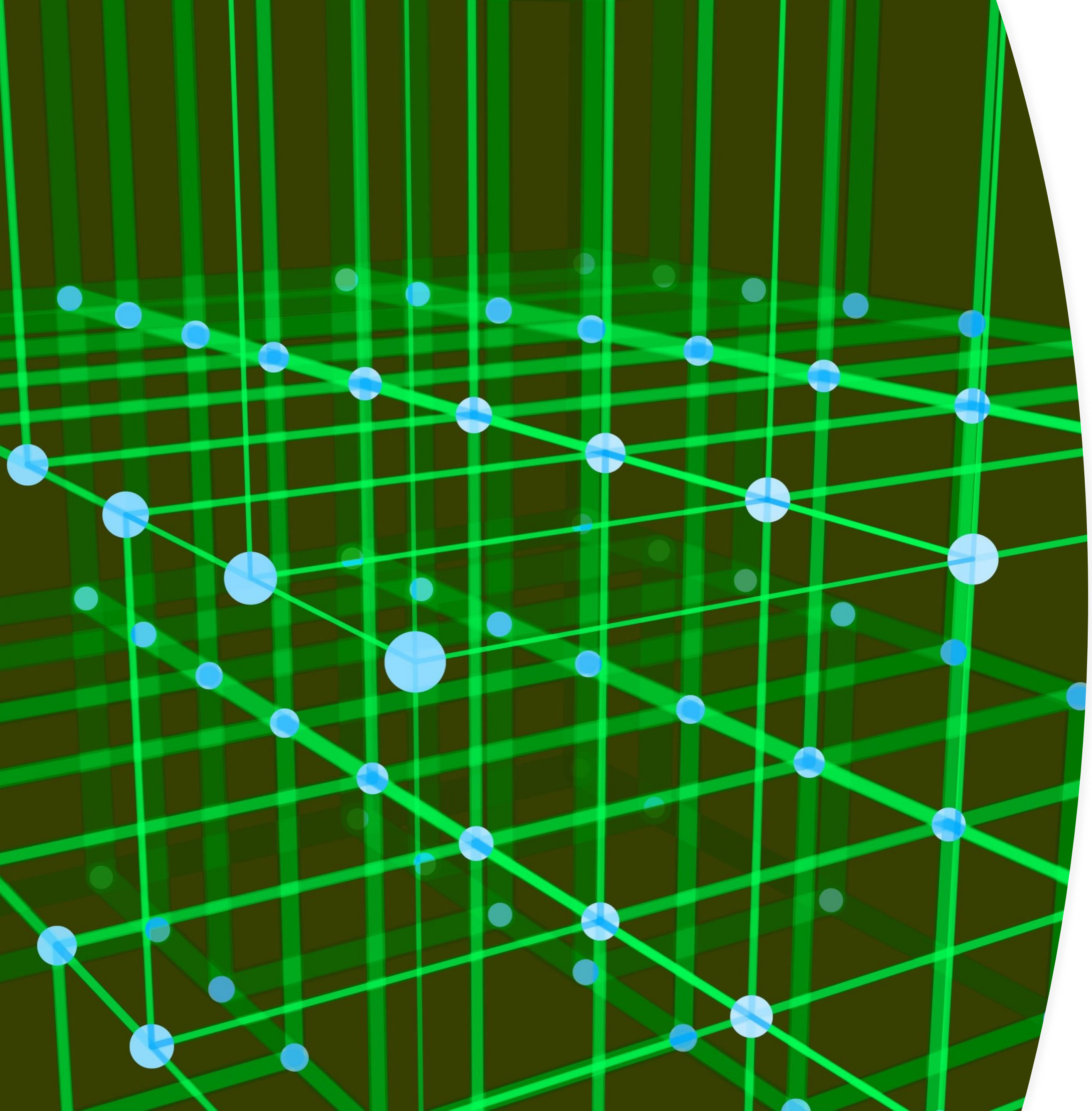
- Without conditional independent assumption?

  $(2^n-1)x2+1$

- With conditional independent assumption?

  $2n+1$

Which is opposed to 2^n before considering naïve bayes.



Naïve Bayes Algorithm

-- A 3-step procedure

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (given data for X and Y)

Step 2

- For each value y_k
  - Estimate $\pi_k \equiv P(Y = y_k)$

Step 1

- For each value x_{ij} of each attribute X_i
  - estimate $\theta_{ijk} = P(X_i = x_{ij} | Y = y_k)$

Training Naïve Bayes Classifier

Step 2

- From the data D, estimate *class priors:*
  - For each possible value of Y, estimate $Pr(Y=y_1), Pr(Y=y_2),\dots. Pr(Y=y_k)$
  - An estimate:

$$\widehat{\pi}_k = \widehat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

Number of items in dataset D for which $Y=y_k$

Step 1

- From the data, estimate the conditional probabilities
  - If every X_i has values $x_{i1},\dots,x_{ik}$
    - for each y_i and each X_i estimate $q(i,j,k)=Pr(X_i=x_{ij}|Y=y_k)$
    - $$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in dataset D for which $X_i=x_{ij}$ and $Y=y_k$

Exercise

- Consider the following dataset:
- P(Wealthy=Y) =
- P(Wealthy=N)=
- P(Gender=F | Wealthy = Y) =
- P(Gender=M | Wealthy = Y) =
- P(HrsWorked > 40.5 | Wealthy = Y) =
- P(HrsWorked < 40.5 | Wealthy = Y) =
- P(Gender=F | Wealthy = N) =
- P(Gender=M | Wealthy = N) =
- P(HrsWorked > 40.5 | Wealthy = N) =
- P(HrsWorked < 40.5 | Wealthy = N) =

Gender	HrsWorked	Wealthy?
F	39	Y
F	45	N
M	35	N
M	43	N
F	32	Y
F	47	Y
M	34	Y

Check the exercises of lecture notes for answers

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (given data for X and Y)

Step 1

- for each value y_k
  - Estimate $\pi_k \equiv P(Y = y_k)$

Step 2

- for each value x_{ij} of each attribute X_i
  - estimate $\theta_{ijk} = P(X_i = x_{ij} | Y = y_k)$

$$\hat{\theta} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(D|\theta) P(\theta)$$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Step 3

$$\boxed{Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}}$$

Exercise (Continued)

- Consider the following dataset:
- Classify a new instance
  - Gender = F /\ HrsWorked = 44

Gender	HrsWorked	Wealthy?
F	39	Y
F	45	N
M	35	N
M	43	N
F	32	Y
F	47	Y
M	34	Y

Check the exercises of lecture notes for answers

Example: Live outside of Liverpool? $P(L|T,D,E)$

- L=1 iff live outside of Liverpool
- T=1 iff shop at Tesco
- D=1 iff Drive or Carpool to Liverpool
- E=1 iff Even # letters last name

P(L=1) :

P(D=1 | L=1) :

P(D=1 | L=0) :

P(T=1 | L=1) :

P(T=1 | L=0) :

P(E=1 | L=1) :

P(E=1 | L=0) :

P(L=0) :

P(D=0 | L=1) :

P(D=0 | L=0) :

P(T=0 | L=1) :

P(T=0 | L=0) :

P(E=0 | L=1) :

P(E=0 | L=0) :