Naïve Bayes

Dr. Xiaowei Huang

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

Up to now,

- Four machine learning algorithms:
  - decision tree learning
  - k-nn
  - linear regression
  - Gradient descent

Topics

- MLE (maximum Likelihood Estimation) and MAP
- Naïve Bayes

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data D

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

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

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

Recall: MAP Queries (Most Probable Explanation)

- Finding a high probability assignment to some subset of variables
- Most likely assignment to all non-evidence variables $W = \chi - Y$

$$MAP(W \mid e) = \arg\max_{w} P(w, e)$$

P(w,e) = P(w|e) P(e)

i.e., value of *w* for which *P(w,e)* is maximum

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

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

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

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

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

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 many parameters must we estimate?

feature vector

- Suppose $X = <X_1, \ldots X_n>$ where X_i and Y are Boolean RV s
- To estimate $P(Y|X_1, X_2, \ldots X_n)$

  2^n quantities need to be estimated!

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

- 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*

Can we reduce params using Bayes Rule?

- Suppose X =<X_1,... X_n> where X_i and Y are boolean RV's
- By Bayes rule:

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

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

- 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)

Naïve Bayes

- 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)

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

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

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

Conditional Independence

- A is conditionally independent of B given C, if the probability distribution governing A is independent of the value of B, given the value of C

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

- Which we often write $P(A|B, C) = P(A|C)$

- Example: $P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$

Assumption for Naïve Bayes

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

Chain rule

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

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

Conditional Independence

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

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

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

Reducing the number of parameters to estimate

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

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

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

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

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

Reducing the number of parameters to estimate

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

- Without conditional indep assumption?
  - $(2^n-1)x2+1$
- With conditional indep assumption?
  - 2n+1

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (given data for X and Y)
- for each value y_k
  - Estimate $\pi_k \equiv P(Y = y_k)$
- 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 Using MLE

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

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

- From the data, estimate the conditional probabilities
  - If every X_i has values $x_{i1},...,x_{ik}$
    - 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 $Y=y_k$

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

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (given data for X and Y)
- for each value y_k
  - Estimate $\pi_k \equiv P(Y = y_k)$
- for each value x_{ij} of each attribute X_i
  - estimate $\theta_{ijk} = P(X_i = x_{ij} | Y = y_k)$

- Classify (X_{new})

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

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

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

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

$P(L=1)$:	$P(L=0)$:
$P(D=1 \mid L=1)$:	$P(D=0 \mid L=1)$:
$P(D=1 \mid L=0)$:	$P(D=0 \mid L=0)$:
$P(T=1 \mid L=1)$:	$P(T=0 \mid L=1)$:
$P(T=1 \mid L=0)$:	$P(T=0 \mid L=0)$:
$P(E=1 \mid L=1)$:	$P(E=0 \mid L=1)$:
$P(E=1 \mid L=0)$:	$P(E=0 \mid L=0)$:

Extended Materials

Naïve Bayes: Subtlety #1

- If unlucky, our MLE estimate for *P(*X_i *| Y)* might be zero. (e.g., nobody in your sample has X_i <= 40.5 and Y=rich)

- Why worry about just one parameter out of many?

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

If one of these terms is 0...

- What can be done to avoid this?

Estimating Parameters: *Y*, X_i discrete-valued

- Maximum likelihood estimates:

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

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

- MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

it is common to use a "smoothed" estimate which effectively adds in a number of additional "hallucinated" examples, and which assumes these hallucinated examples are spread evenly over the possible values of X_i.

Only difference: "hallucinated" examples

Naïve Bayes: Subtlety #2

- Often the X_i are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- What is effect on estimated P(Y|X)?
  - Special case: what if we add two copies: $X_i = X_k$

Special case: what if we add two copies: $X_i = X_k$

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

Redundant
terms

About Naïve Bayes

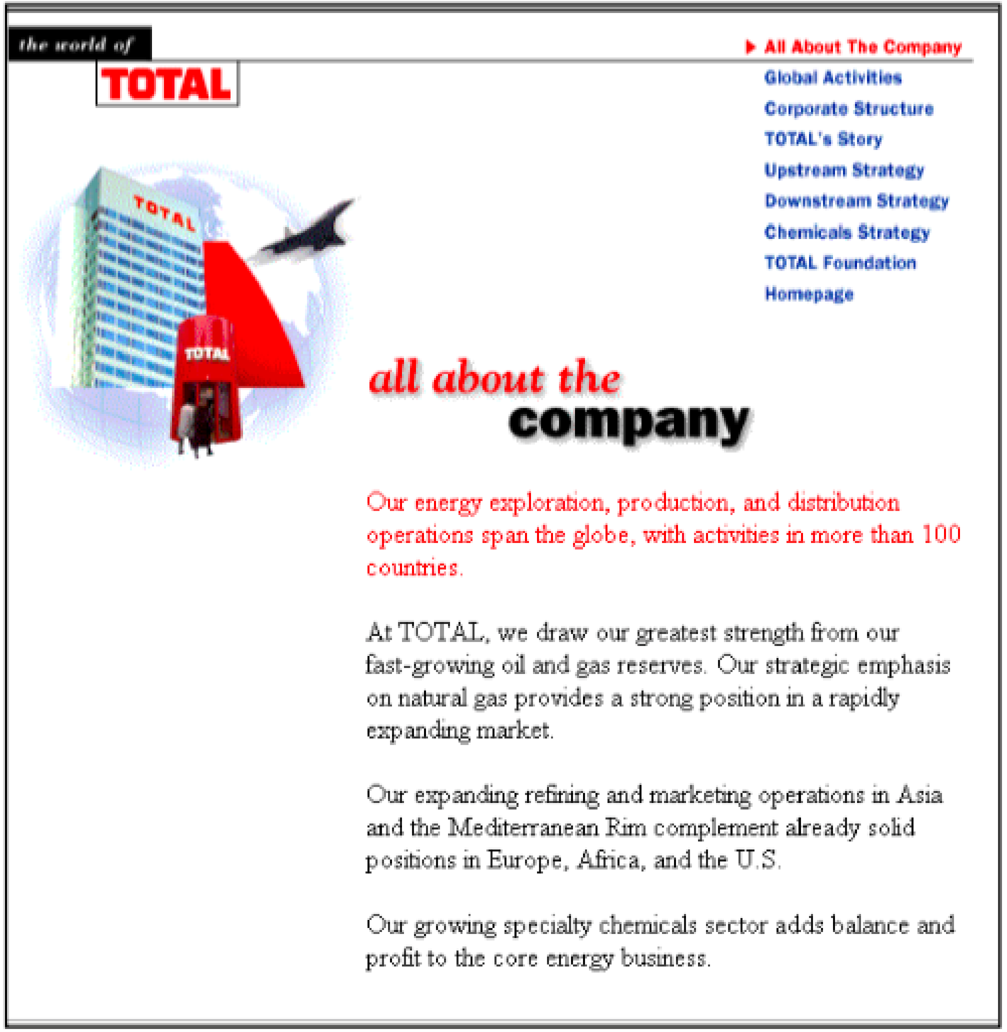
- Naïve Bayes is blazingly fast and quite robust!

Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

- How shall we represent text documents for Naïve Bayes?

Baseline: Bag of Words Approach



aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Learning to classify document: P(Y|X) the Bag of Words model

- Y discrete valued. e.g., Spam or not
- $X = <X_1, X_2, \ldots X_n>$ = document

- X_i is a random variable describing the word at position i in the document
- possible values for X_i : any word w_k in English

- Document = bag of words: the vector of counts for all w_k's
  - (like #heads, #tails, but we have more than 2 values)

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (given data for X and Y)
- for each value y_k
  - Estimate $\pi_k \equiv P(Y = y_k)$
- for each value x_{ij} of each attribute X_i
  - estimate $\theta_{ijk} = P(X_i = x_{ij} | Y = y_k)$

prob that word x_j appears in position i, given $Y=y_k$

- Classify (X_{new})

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

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

Additional assumption: word probabilities are position independent

$$\theta_{ijk} = \theta_{mjk} \;\; \text{for all } i, m$$

MAP estimates for bag of words

- MAP estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}$$

$$\theta_{aardvark} = P(X_i = \text{aardvark}) = \frac{\#\text{ observed 'aardvark'} \;+ \#\text{ hallucinated 'aardvark'} - 1}{\#\text{ observed words} \;+ \#\text{ hallucinated words} - k}$$

- What β s should we choose?

Twenty NewsGroups

Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

What you should know:

- Training and using classifiers based on Bayes rule

- Conditional independence
    - What it is
    - Why it’s important

- Naïve Bayes
    - What it is
    - Why we use it so much
    - Training using MLE, MAP estimates