Bayesian Classifiers, Conditional Independence and Naïve Bayes

Required reading:

"Naïve Bayes and Logistic Regression"

(available on class website)

Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

Jan 20, 2010

Announcements

- Homework 1 due today
- Homework 2 out soon – watch email

- Auditors must
  - officially register to audit
  - hand in at least n-1 or the n homeworks

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

Suppose Y=Wealth, X=<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 many parameters must we estimate?

Suppose X =<X1,… Xn>
where Xi and Y are boolean RV's

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

P(poor | G, HW) = 1-P(rich | G, HW)

To estimate P(Y| X1, X2, … Xn)

need 2^n params to est

If we have 30 Xi's instead of 2?

2^{30} > 1 Billion

Can we reduce params by using Bayes Rule?

Suppose X =<X1,… Xn>
where Xi and Y are boolean RV's

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

$$1 \gtrless \frac{P(Y=r|x)}{P(Y=p|x)} = \frac{P(x|Y=r)P(Y=r)}{P(x|Y=p)P(Y=p)}$$

\# params

$2(2^n) - 2$

$P(X_1 \dots X_n | Y)$

2^n

$X_1 \; X_2 \; X_3 \; \dots \; X_n$

$P(X_1 \dots X_n | Y=r)$ $\quad P(X_1 \dots X_n | Y=p)$

don't sum to 1

Bayes Rule

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

Which is shorthand for:

$$(\forall i,j) P(Y=y_i|X=x_j) = \frac{P(X=x_j|Y=y_i)P(Y=y_i)}{P(X=x_j)}$$

Equivalently:

$$(\forall i,j) P(Y=y_i|X=x_j) = \frac{P(X=x_j|Y=y_i)P(Y=y_i)}{\sum_k P(X=x_j|Y=y_k)P(Y=y_k)}$$

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≠j

Conditional Independence

Definition: X is <u>conditionally independent</u> of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i|Y = y_j, Z = z_k) = P(X = x_i|Z = z_k)$$

Which we often write

$$P(X|Y, Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y

Given this assumption, then:

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

by cond. indep. def!

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

$P(x_i=1|Y=1)$
$P(x_i=1|Y=0)$

How many parameters to describe P(X1…Xn|Y)? P(Y)?

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

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)P(X_1 \ldots X_n|Y = y_k)}{\sum_j P(Y = y_j)P(X_1 \ldots X_n|Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \ldots, X_n >$ is:

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

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

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

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

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's):

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

$$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in dataset D for which Y=y_k

Example: Live in Sq Hill? P(S|G,D,M)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

What probability parameters must we estimate?

Example: Live in Sq Hill? P(S|G,D,M) ∝ P(S) P(G|S) P(D|S) P(M|S)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Rachel Maddow fan

$\hat{P}$(S=1) : 26/67	P(S=0) : 1 − 26/67
$\hat{P}$(D=1 | S=1) : 2/26	P(D=0 | S=1) : 24/26
$\hat{P}$(D=1 | S=0) : 5/41	P(D=0 | S=0) :
$\hat{P}$(G=1 | S=1) : 24/26	P(G=0 | S=1) :
$\hat{P}$(G=1 | S=0) : 16/41	P(G=0 | S=0) :
$\hat{P}$(M=1 | S=1) : 3/26	P(M=0 | S=1) :
$\hat{P}$(M=1 | S=0) : 3/41	P(M=0 | S=0) :

S=0, $\hat{S}$=0 (.9)
S=1, $\hat{S}$=1 (.6)
S=0, $\hat{S}$=1 (.6) d=0, m=0, g=1

$$P(S=1|G=g, M=m, D=d) \propto \frac{P(S=1)\,P(D=d|S=1)\,P(M=m|S=1)\,P(G=g|S=1)}{P(S=0)\,P(\ |S=0)\,P(\ |S=0)\,P(\ |S=0)}$$

26/67 · 24/26 · 23/26 · 24/26

Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero. (e.g., X_{373}= Birthday_Is_January_30_1990)

- Why worry about just one parameter out of many?

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

one is zero

- What can be done to avoid this?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

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

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

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

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

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

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

$$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + \alpha_k}{|D| + \sum_m \alpha_m}$$

Only difference: "imaginary" examples

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + \alpha'_k}{\#D\{Y = y_k\} + \sum_m \alpha'_m}$$

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

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

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

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

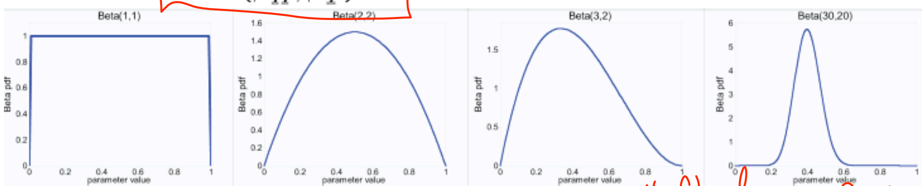
$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \sim Beta(\beta_H,\beta_T)$$

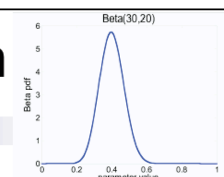
Mean:

Mode:



- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$ (# of heads, # of tails)
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H,\beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H,\beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H+\beta_H}{\alpha_H+\beta_H+\alpha_T+\beta_T}$$

- Beta prior equivalent to extra thumbtack flips
- As $N \to \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Dirichlet distribution

- number of heads in N flips of a two-sided coin
  - follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)

- what it's not two-sided, but k-sided?
  - follows a multinomial distribution
  - Dirichlet distribution is the conjugate prior

$$P(\theta_1, \theta_2, ...\theta_K) = \frac{1}{B(\alpha)} \prod_i^K \theta_i^{(\alpha_1 - 1)}$$

Lejeune Dirichlet

Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	Germany
Nationality	German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

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: $X_i = X_k$

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 Text

Target concept $Interesting? : Document \rightarrow \{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k|v_j)$ is probability that word in position i is w_k, given v_j

one more assumption: $P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

Twenty NewsGroups

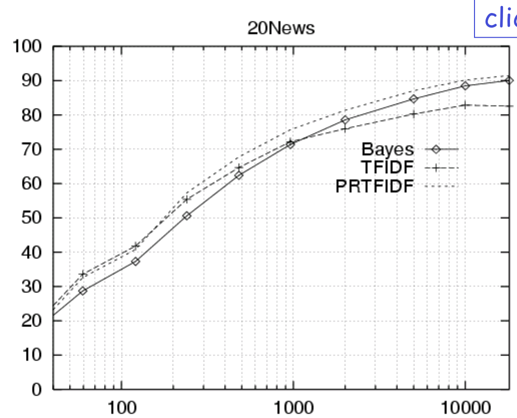
Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

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

Learning Curve for 20 Newsgroups

For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"



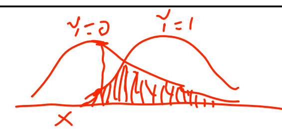
Accuracy vs. Training set size (1/3 withheld for test)

What you should know:

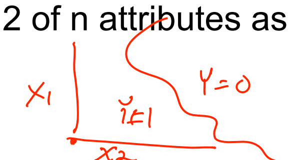
- Training and using classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important
- Naïve Bayes
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables and continuous (Gaussian)

$$P(Y|X_1 \cdots X_n) \approx \frac{P(Y)\prod_i P(X_i|Y)}{P(X_1 \cdots X_n)}$$

Questions:



- What is the error will classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?
- Can you use Naïve Bayes for a combination of discrete and real-valued X_i?
- How can we easily model just 2 of n attributes as dependent?



- What does the decision surface of a Naïve Bayes classifier look like?