Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 13, 2012

Today:
- Bayes Classifiers
- Naïve Bayes
- Gaussian Naïve Bayes

Readings:
Mitchell:
"Naïve Bayes and Logistic Regression"
(available on class website)

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

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

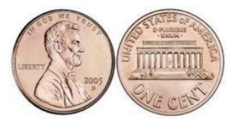
$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Conjugate priors

$X \in \{0,1\}$
$P(X=1) \equiv \theta$

- P(θ) and P(θ|D) have the same form

Eg. 1 Coin flip problem



Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1}$$

If prior is Beta distribution,

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

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

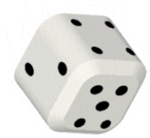
[A. Singh]

Conjugate priors

- P(θ) and P(θ|D) have the same form

$P(X=1)$
$P(X=2)$

Eg. 2 Dice roll problem (6 outcomes instead of 2)



Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\ldots\theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1 - 1}\theta_2^{\beta_2 - 1}\ldots\theta_k^{\beta_k - 1}}{B(\beta_1, \beta_2, \ldots \beta_K)} \sim \text{Dirichlet}(\beta_1 \ldots \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

21

[A. Singh]

Conjugate priors

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of

Likelihood is ~ Multinomial(θ = {θ₁, θ₂, ... , θ

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\dots\theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1-1}\theta_2^{\beta_2-1}\dots\theta_k^{\beta_k-1}}{B(\beta_1, \beta_2, \dots \beta_K)} \sim \text{Dirichlet}$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \dots, \beta_k +$$

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

For Multinomial, conjugate prior is Dirichlet distribution.

21

[A. Singh]

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

f : X → Y

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

$$P(r \mid f, <40.5) = \frac{P(r, f, <40.5)}{P(f, <40.5)}$$

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?

$P(Y|X)$

Suppose X =<X_1,… X_n>
where X_i 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

X_1 X_2 4

$+ \ldots = 1$

To estimate $P(Y| X_1, X_2, \ldots X_n)$

2^n params

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

$2^{30} \sim 1B$

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:

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

Can we reduce params using Bayes Rule?

Suppose X =<X_1,… X_n>
where X_i and Y are boolean RV's

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

$P(X_1 \cdots X_n | Y) \Rightarrow (2^n - 1)2$ params

$\hat{\theta}_{f,<40,r} = P(X_1 = f,\ X_2 = <40 \mid Y = r)$

$P(Y) \Rightarrow 1$ param

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

$X_1 \perp X_2 | Y$

Given this assumption, then:

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

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

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

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

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

$P(X_1|Y) \Rightarrow 2$ params

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, to pick most probable Y for $X^{new} = < X_1, \ldots, X_n >$

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

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 of these...

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,E)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive to CMU
- E=1 iff even # of letters in last name

What probability parameters must we estimate?

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

test examples

$P(S=1|D,E,S)$

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive or Carpool to CMU
- E=1 iff Even # letters last name

P(S=1) :	26/110	P(S=0) :	1 − P(S=1)
P(D=1 | S=1) :	2/26	P(D=0 | S=1) :	24/26
P(D=1 | S=0) :	2/84	P(D=0 | S=0) :	
P(G=1 | S=1) :	24/26	P(G=0 | S=1) :	2/26
P(G=1 | S=0) :	14/84	P(G=0 | S=0) :	
P(E=1 | S=1) :	13/26	P(E=0 | S=1) :	13/26
P(E=1 | S=0) :	30/84	P(E=0 | S=0) :	54/84

	G	D	E	S
⊕	1	0	1	.7
⊕	1	1	1	.89
⊕	0	0	0	.02
⊖	1	0	1	.7
⊖	1	0	0	.4
⊕	1	0	0	.4

$$P(S=1|G=1,D=0,E=1) = \frac{\underbrace{P(G=1|S=1)\,P(D=0|S=1)\,P(E=1|S=1)\,P(S=1)}_{A}}{A + P(G=1|S=0)\,P(D=0|S=0)\,P(E=1|S=0)\,P(S=0)}$$

0.73

⊕ = correct test example

⊖ = incorrect

Naïve Bayes: Subtlety #1

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

- Why worry about just one parameter out of many?
- 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:

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

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

MAP estimates (Beta, Dirichlet priors):

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

Only difference: "imaginary" examples

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

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: $X_i = X_k$