Machine Learning 10-601, 10-301

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

April 7, 2021

Today:

- Finish MAP estimate
- Bayes Classifiers
- Conditional Independence
- Naïve Bayes

Required Reading:

Mitchell:
"Naïve Bayes and Logistic Regression"

http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

Principles for Estimating Probabilities

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

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

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

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} P(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})} \\ &= \arg\max_{\theta} P(\mathcal{D} \mid \theta) P(\theta) \end{aligned}$$

Maximum Likelihood Estimate

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum A Posteriori (MAP) Estimate

X=1 X=0

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

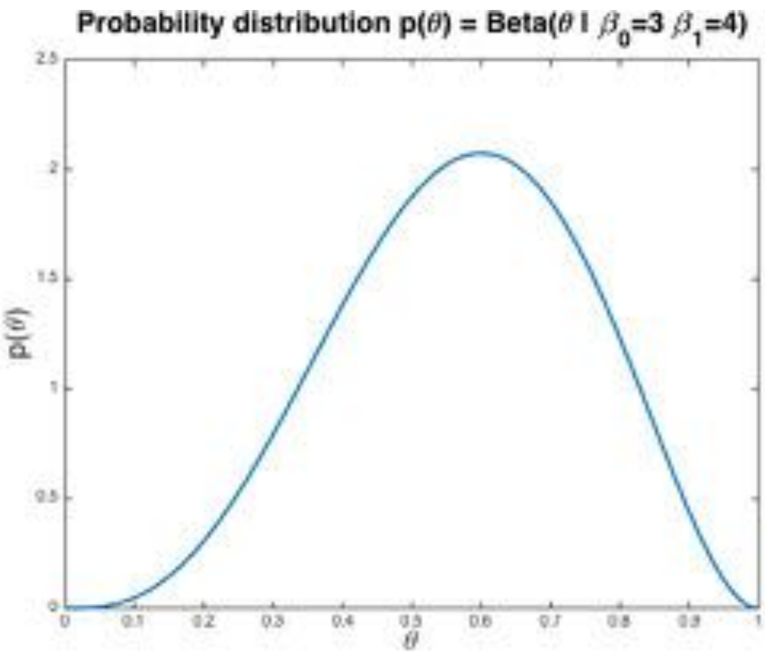
- Assume prior $P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1,\beta_0)}\,\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}$
- Then

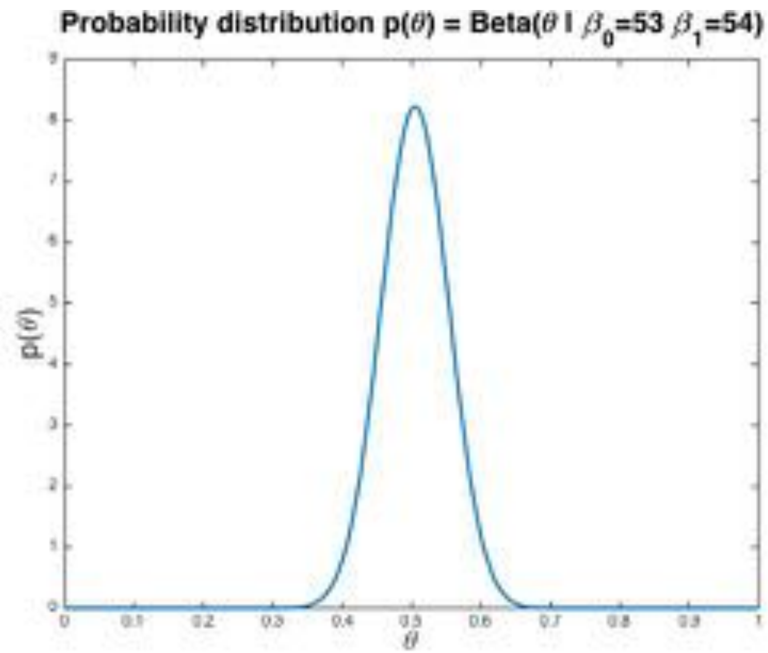
$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{\alpha_1+\beta_1-1}{(\alpha_1+\beta_1-1)+(\alpha_0+\beta_0-1)}$$

(like MLE, but hallucinating $\beta_1 - 1$ additional heads, $\beta_0 - 1$ additional tails)

Beta prior distribution – P(θ)

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





We say P(θ) is the *conjugate prior* for P(D|θ), if P(θ|D) has same form as P(θ)

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

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

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]

We say P(θ) is the *conjugate prior* for P(D|θ), if P(θ|D) has same form as P(θ)

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,

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

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.

[A. Singh]

You should know

- Probability basics
  - random variables, conditional probs, …
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution
- Estimating parameters from data
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – Bernoulli, Binomial, Beta, Dirichlet, …
  - conjugate priors

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

Consider 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 =<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

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

If we have 100 boolean X_i's: $P(Y | X_1, X_2, \ldots X_{100})$

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 $X = \langle X_1, \ldots X_n \rangle$
where X_i and Y are boolean RV's

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

How many parameters to define $P(X_1, \ldots X_n \mid Y)$?

How many parameters to define $P(Y)$?

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 $i \neq j$

Conditional Independence

Definition: X is conditionally independent 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. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

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

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$
\begin{aligned}
P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\
&= P(X_1|Y)P(X_2|Y)
\end{aligned}
$$

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

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

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?
- With conditional indep assumption?

Naïve Bayes in a Nutshell

Bayes rule:

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

Assuming conditional independence among X_i' s:

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

So, to pick most probable Y for $X^{new} = \langle X_1, \dots, X_n \rangle$

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

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):

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

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

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

Example: Taking 10-601 or 10-301? $P(G|F,B,U)$

- G=1 iff you're taking 10-601
- U=1 iff taking undergrad class
- F=1 iff first year at CMU
- B=1 iff Birthday is before July 1

What probability parameters must we estimate from training data?

Example: Taking 10-601 or 10-301? P(G|F,B,U)

- G=1 iff you're taking 10-601
- U=1 iff taking undergrad class
- F=1 iff first year at CMU
- B=1 iff Birthday is before July 1

P(G=1) :

P(F=1 | G=1) :

P(F=1 | G=0) :

P(B=1 | G=1) :

P(B=1 | G=0) :

P(U=1 | G=1) :

P(U=1 | G=0) :

P(G=0) :

P(F=0 | G=1) :

P(F=0 | G=0) :

P(B=0 | G=1) :

P(B=0 | G=0) :

P(U=0 | G=1) :

P(U=0 | G=0) :

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

$$G^{new} \leftarrow \arg \max_{g_k \in \{0,1\}} P(G = g_k) \; P(F = f^{new} | G = g_k) \; P(B = b^{new} | G = g_k) \; P(U = u^{new} | G = g_k)$$

Example: Taking 10-601 or 10-301? $P(G|F,B,U)$

- G=1 iff you're taking 10-601
- U=1 iff taking undergrad class
- F=1 iff first year at CMU
- B=1 iff Birthday is before July 1

$P(G=1)$:

$P(F=1 \mid G=1)$:

$P(F=1 \mid G=0)$:

$P(B=1 \mid G=1)$:

$P(B=1 \mid G=0)$:

$P(U=1 \mid G=1)$:

$P(U=1 \mid G=0)$:

$P(G=0)$:

$P(F=0 \mid G=1)$:

$P(F=0 \mid G=0)$:

$P(B=0 \mid G=1)$:

$P(B=0 \mid G=0)$:

$P(U=0 \mid G=1)$:

$P(U=0 \mid G=0)$:

Naïve Bayes: Subtlety #1

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)?
  - Extreme case: what if we add two copies: $X_i = X_k$

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

Naïve Bayes: Subtlety #2

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero.
(for example, X_i = *birthdate.* X_i = *Jan_25_2002*)

- Why worry about just one parameter out of many?

- What can be done to address 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:

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

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

MAP estimates (Beta, Dirichlet priors):

Only difference: "imaginary" examples

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

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

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 many more than 2 values
  - assume word probabilities are position independent (i.i.d. rolls of a 50,000-sided die)

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_j of each attribute X_i

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

  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 <u>position</u> independent

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

MAP estimates for bag of words

Map estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\underbrace{\sum_{m=1}^{k} \alpha_m}_{\text{Observed count for word m}} + \underbrace{\sum_{m=1}^{k} (\beta_m - 1)}_{\text{Hallucinated count for word m}}}$$

What β' s should we choose?

Twenty NewsGroups

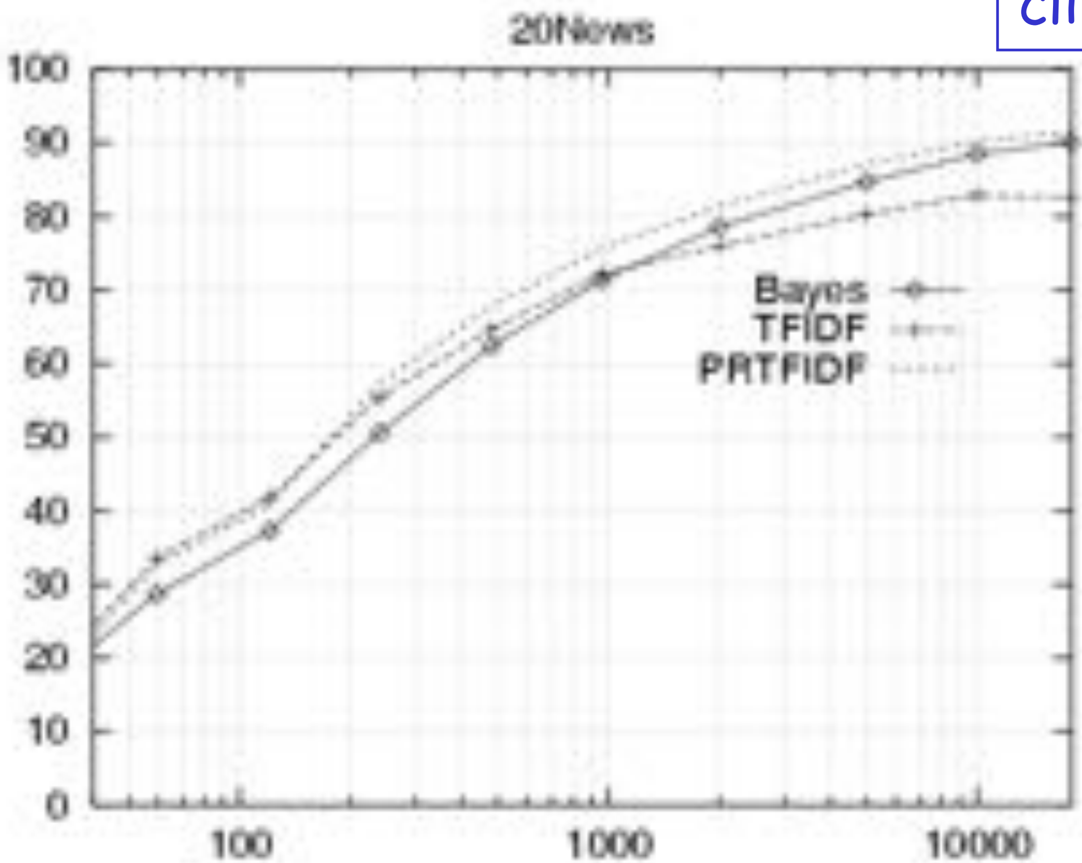
Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

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

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)

Questions:

- How can we extend Naïve Bayes if just 2 of the X_i's are <u>dependent</u>?

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

- What error will the 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?