Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 26, 2015

Today:
- Bayes Classifiers
- Conditional Independence
- Naïve Bayes

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

Two Principles for Estimating Parameters

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

$$\frac{\alpha_1}{\alpha_1 + \alpha_0}$$

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

P(Data | Model)

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

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

P(Model | Data)

$$\frac{\alpha_1 + \beta_1}{(\alpha_1 + \beta_1) + (\alpha_0 + \beta_0)}$$

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

Maximum Likelihood Estimate

- 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

- 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

$\theta^{\alpha_1+\beta_1-1}(1-\theta)^{\alpha_0+\beta_0-1}$

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

(like MLE, but hallucinating β_1-1 additional heads, β_0-1 additional tails)

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

Consider Y=Wealth, X=<Gender, HoursWorked> P(W|G,H)

$2^3 = 8$

#iparams = 7

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

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

1-p

#iparams = 4

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

$G(x) = \arg\max_{y \in \{0,1\}} P(Y|X)$

To estimate P(Y| X_1, X_2, … X_n)

$Y=1 : 2^n$

$Y=0 : 2^n$

$\#\text{params} = 2^n$

If we have 30 boolean X_i's: P(Y | X_1, X_2, … X_{30})

$2^{30} = (2^{10})^3 \approx 10^9$

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

$$\underbrace{P(Y|X)}_{2^n} = \frac{P(X|Y)P(Y)}{P(X)}$$

$2^{n+1} - 1$

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

$Y = 1: \quad 2^n - 1$

$Y = 0: \quad 2^n - 1$

$\longrightarrow 2 \cdot (2^n - 1)$

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

1

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 <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

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

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

Which we often write

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

$P(Y|Z)$

E.g.,

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

$P(T, R|L) = P(T|L)\,P(R|L)$

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:

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

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

$$P(X_1 \ldots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

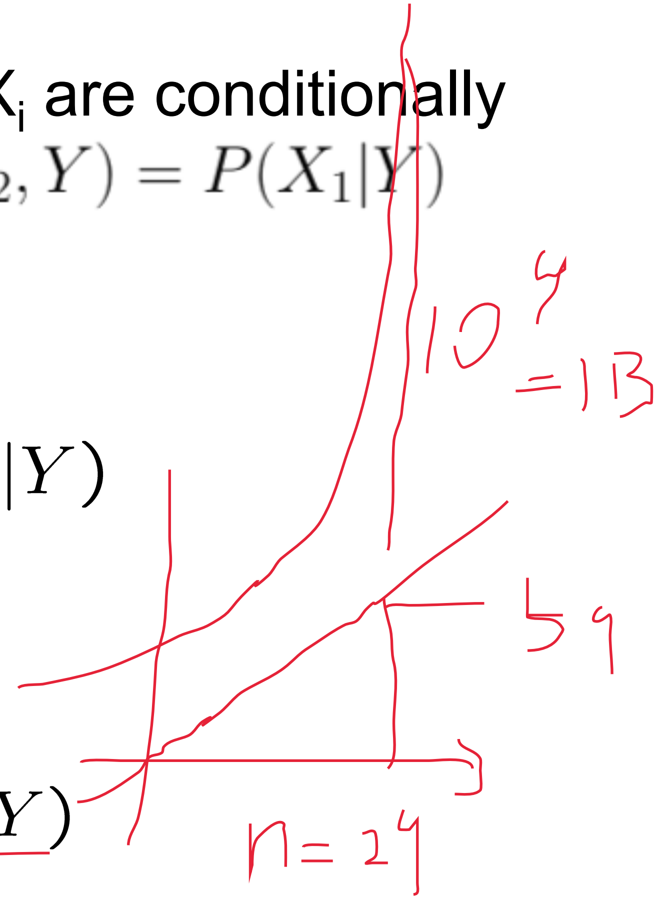
Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

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

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:

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

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? $2(2^n - 1)$ $+$ 1
- With conditional indep assumption? $2n$ $+$ 1

$Y=1: \; 2n/2 = n$

$Y=0: \; 2n/2 = n$

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 $X^{new} = < X_1, \ldots, X_n >$

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

$P(X_i^{new} = x_{ij} | Y = y_k)$

Naïve Bayes Algorithm – discrete X_i

$X_i \in \{1, \ldots, J\}, \quad Y \in \{1, \ldots, K\}$

- Train Naïve Bayes (examples)

  for each* value y_k

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

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

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

- 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):

K $$\widehat{\pi}_k = \widehat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$ $\cdot\ \ell(\theta)$ ~~max~~ $\underline{\ln P(D \mid \theta, \pi)}$

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

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

$$\frac{\partial \ell(\theta)}{\partial \pi} = 0$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

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: $X_i = X_k$

$$P(Y=1|X) \propto P(Y=1)\, \underline{P(X_1|Y=1)}\, \underline{P(X_2|Y=1)} \cdots$$

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_~~1992~~*)

$P(X_i = \_\_ \mid Y) = 0 = \frac{1}{365}$

- Why worry about just one parameter out of many?

$P(Y=1 \mid X_1 \ldots X_n) = \frac{P(Y) \prod_i P(X_i \mid Y)}{P(X)} = 0$

- What can be done to address this?

① high-dim

MLE ← prior

⇓

MAP

② Sparse

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

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

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

$l(\theta, \pi) =$

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

$\frac{\partial l}{\partial \theta} = 0 \quad \frac{\partial l}{\partial \pi} = 0$

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

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

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

$\ln P(\mathrm{Data} | \theta, \pi) \, P(\pi)$

What you should know:

- Training and using classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important

$$\left(2(2^n - 1) + 1\right) \Downarrow 2n + 1$$

- Naïve Bayes
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables and continuous (Gaussian)

$$\hat{f}(x) = \arg\max_y P(Y|X) \propto P(X|Y)\,P(Y) \quad \prod_i P(X_i|Y)$$

MLE θ π GNB

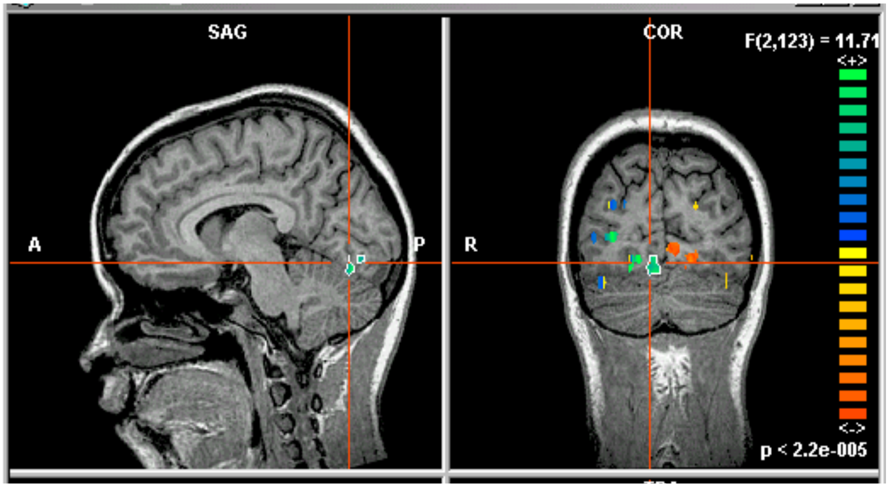
MAP θ π θ

Questions:

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

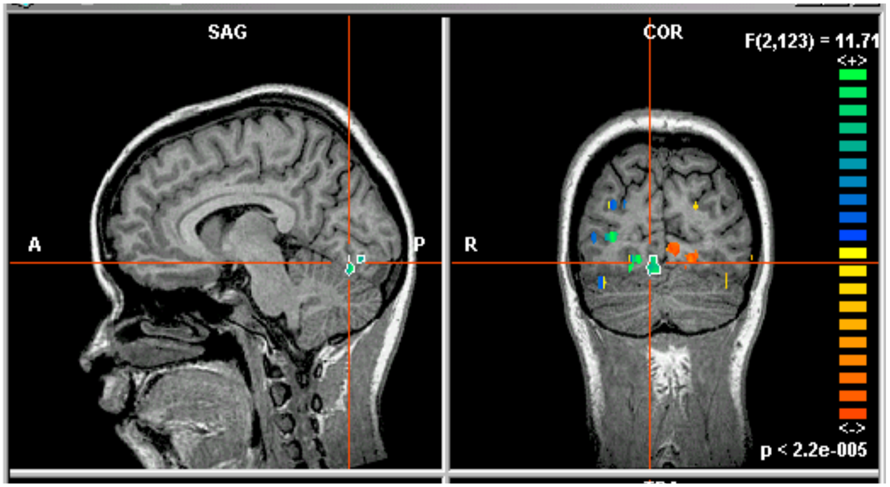
What if we have continuous X_i ?

Eg., image classification: X_i is ith pixel



What if we have continuous X_i ?

image classification: X_i is i[th] pixel, Y = mental state



Still have:

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

Just need to decide how to represent $P(X_i \mid Y)$

What if we have continuous X_i ?

Eg., image classification: X_i is i^{th} pixel

Gaussian Naïve Bayes (GNB): assume

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Sometimes assume σ_{ik}

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Gaussian Naïve Bayes Algorithm – continuous X_i (but still discrete Y)

- Train Naïve Bayes (examples)

  for each value y_k

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

  for each attribute X_i estimate

  class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

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

$$Y^{new} \leftarrow \arg\max_{y_k} \ \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

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

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\widehat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

δ(z)=1 if z true, else 0

$$\widehat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \widehat{\mu}_{ik})^2 \delta(Y^j = y_k)$$