Naïve Bayes Classifier

Machine Learning 10-601B

Seyoung Kim

Many of these slides are derived from Tom Mitchell, William Cohen, Eric Xing. Thanks!

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 or Carpool to CMU
- E=1 iff Even # letters last name

What probability parameters must we estimate?

P(S=1) :	P(S=0) :
P(D=1 | S=1) :	P(D=0 | S=1) :
P(D=1 | S=0) :	P(D=0 | S=0) :
P(G=1 | S=1) :	P(G=0 | S=1) :
P(G=1 | S=0) :	P(G=0 | S=0) :
P(E=1 | S=1) :	P(E=0 | S=1) :
P(E=1 | S=0) :	P(E=0 | S=0) :

Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i \mid Y)$ 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?

$$P(X_1 \ldots X_n|Y) = \prod_i \underbrace{P(X_i|Y)}_{\text{If one of these terms is 0...}}$$

- 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:

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

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

MAP estimates (Beta, Dirichlet priors):

Only difference:
"imaginary" examples

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

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

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: $X_i = X_k$

$$P(X_1 \ldots X_n|Y) = \prod_i \underbrace{P(X_i|Y)}_{\text{Redundant terms}}$$

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
- $X = <X_1, X_2, \dots X_n>$ = 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 (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 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

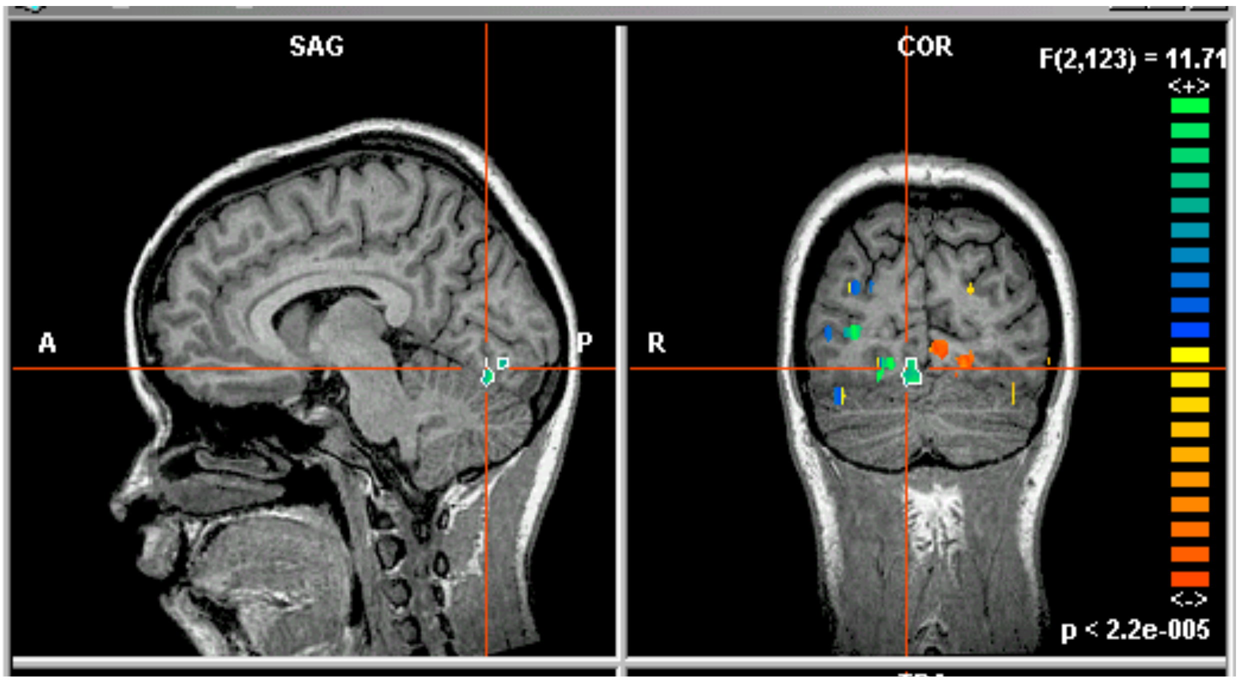
Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

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 if we have continuous X_i ?

Eg., image classification: X_i is real-valued ith pixel



Given input images X

- Classify whether this is from a normal or schizophrenic brain
- Classify which tasks he/she is performing?
- Classify which word he/she is reading?

What if we have continuous X_i ?

Eg., image classification: X_i is real-valued i[th] pixel

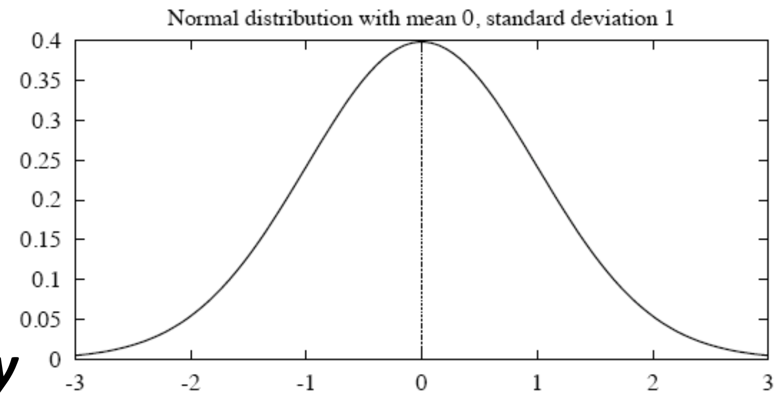
Naïve Bayes requires $P(X_i \mid Y=y_k)$, but X_i is real (continuous)

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

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution

Gaussian Distribution
(also called "Normal")

p(x) is a *probability density function*, whose integral (not sum) is 1



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

What if we have continuous X_i?

Gaussian Naïve Bayes (GNB): assume

$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2}$$

Sometimes assume variance

- 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 $P(X_i|Y = y_k)$

  - class conditional mean μ_{ik}, standard deviation σ_{ik}

- Classify (X^{new})

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

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

* 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

$\delta()$=1 if (Y^j=y_k) else 0

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

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 to think about:

- 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?
- How would you select a subset of X_i's?
- How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?

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