Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

September 18, 2012

Today:
- Naïve Bayes
  - discrete-valued X_i's
  - Document classification
- Gaussian Naïve Bayes
  - real-valued X_i's
  - Brain image classification
- Form of decision surfaces

Readings:

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

Optional
- Bishop 1.2.4
- Bishop 4.2

Recently:
- Bayes classifiers to learn P(Y|X)
- MLE and MAP estimates for parameters of P
- Conditional independence
- Naïve Bayes → make Bayesian learning practical

Next:
- Text classification
- Naïve Bayes and continuous variables X_i:
  - Gaussian Naïve Bayes classifier
- Learn P(Y|X) directly
  - Logistic regression, Regularization, Gradient ascent
- Naïve Bayes or Logistic Regression?
  - Generative vs. Discriminative classifiers

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:

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

Another way to view Naïve Bayes (Boolean Y):

Decision rule: is this quantity greater or less than 1?

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

$$0 \gtrless \ln[\quad] = \ln[\downarrow] + \sum_i \ln[\downarrow]$$

$$0 \gtrless \ln\frac{P(Y=1)}{P(Y=0)} + \sum_i \left[X_i \log\frac{\theta_{i1}}{\theta_{i0}} + (1-X_i)\log\frac{(1-\theta_{i1})}{(1-\theta_{i0})}\right]$$

$$\hat{\theta}_{ik} = \hat{P}(X_i=1|Y=y_k)$$

Naïve Bayes: classifying text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?

I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.

I would like to thank Frank Pfenning, who has served ably in this role for the past two years.

****
Randal E. Bryant
Dean and University Professor

How shall we represent text documents for Naïve Bayes?

Learning to classify documents: P(Y|X)

- Y discrete valued.
  - e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document

I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.

I would like to thank Frank Pfenning, who has served ably in this role for the past two years.

****
Randal E. Bryant
Dean and University Professor

- X_i is a random variable describing…

Learning to classify documents: P(Y|X)

- Y discrete valued.
  - e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document

> I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.
>
> I would like to thank Frank Pfenning, who has served ably in this role for the past two years.
>
> ******************************
>
> Randal E. Bryant
> Dean and University Professor

- X_i is a random variable describing…

Answer 1: X_i is boolean, 1 if word i is in document, else 0

e.g., $X_{pleased} = 1$

Issues?

Learning to classify documents: P(Y|X)

- Y discrete valued.
  - e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document

> I am pleased to announce that Bob Frederking of the Language Technologies Institute is our new Associate Dean for Graduate Programs. In this role, he oversees the many issues that arise with our multiple masters and PhD programs. Bob brings to this position considerable experience with the masters and PhD programs in the LTI.
>
> I would like to thank Frank Pfenning, who has served ably in this role for the past two years.
>
> ******************************
>
> Randal E. Bryant
> Dean and University Professor

- X_i is a random variable describing…

Answer 2:

- X_i represents the i^{th} *word position* in document
- X_1 = "I", X_2 = "am", X_3 = "pleased"
- and, let's assume the X_i are iid (indep, identically distributed)

$$P(X_i|Y) = P(X_j|Y) \quad (\forall i, j)$$

Learning to classify document: P(Y|X) the "Bag of Words" model

- Y discrete valued. e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document

- X_i are iid random variables. Each represents the word at its position i in the document
- Generating a document according to this distribution = rolling a 50,000 sided die, once for each word position in the document

- The observed counts for each word follow a ??? distribution

Multinomial Distribution

- $P(\theta)$ and $P(\theta|D)$ have the same form

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,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

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.

21

Multinomial Bag of Words

→

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

MAP estimates for bag of words

Map estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}$$

$$\theta_{aardvark} = P(X_i = \text{aardvark}) = \frac{\#\ \text{observed 'aardvark'} + \#\ \text{hallucinated 'aardvark'} - 1}{\#\ \text{observed words} + \#\ \text{hallucinated words} - k}$$

{γ =span)

What β's should we choose?

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

prob that word x_{ij} 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

$$\theta_{ijk} = \theta_{mjk} \;\; \text{for} \;\; i \neq m$$

Twenty NewsGroups

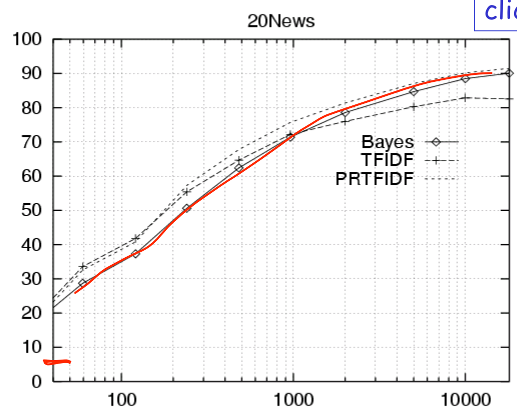
Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

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

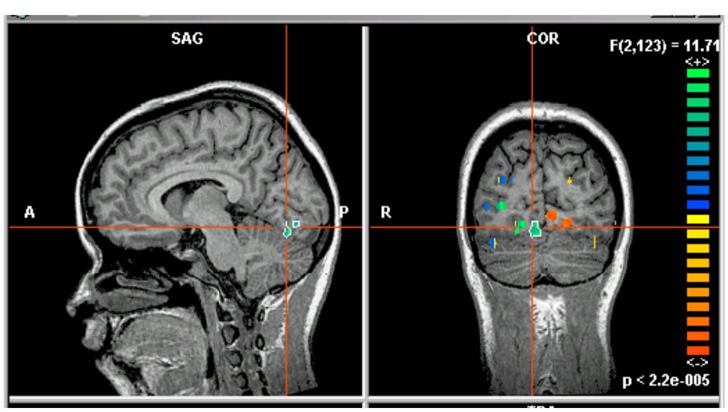
For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"



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

What if we have continuous X_i ?

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



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)

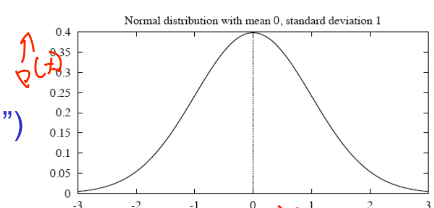
$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

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



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

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

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

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}, variance σ_{ik}
- Classify (X^{new})

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

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

* 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

δ()=1 if (Y^j=y_k) else 0

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

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?

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

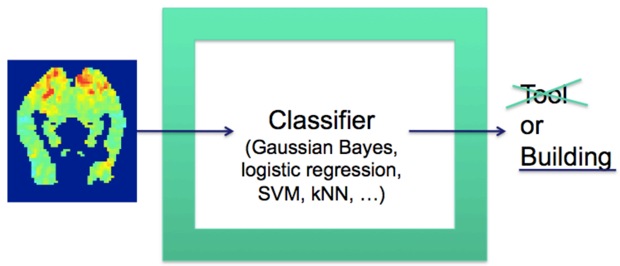
Simple Picture for GNB for $P(Y|X_1)$

What is form of decision surface for Gaussian Naïve Bayes classifier?

eg., if we assume attributes have same variance, indep of Y ($\sigma_{ik} = \sigma$)

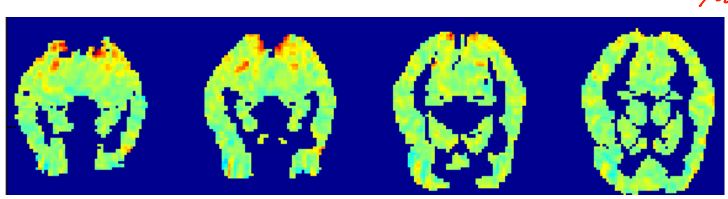
GNB Example: Classify a person's cognitive state, based on brain image

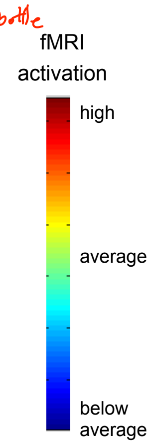
- reading a sentence or viewing a picture?
- reading the word describing a "Tool" or "Building"?
- answering the question, or getting confused?



Mean activations over all training examples for Y="bottle"

$\mu_{i\ y=bottle}$

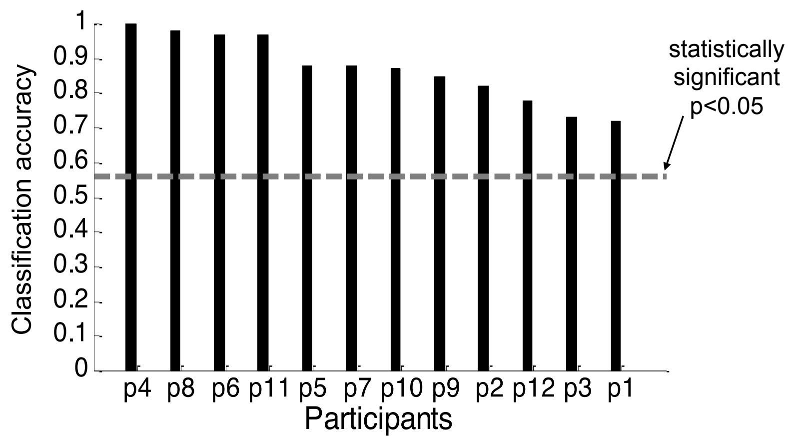




Y is the mental state (reading "house" or "bottle")
X_i are the voxel activities,

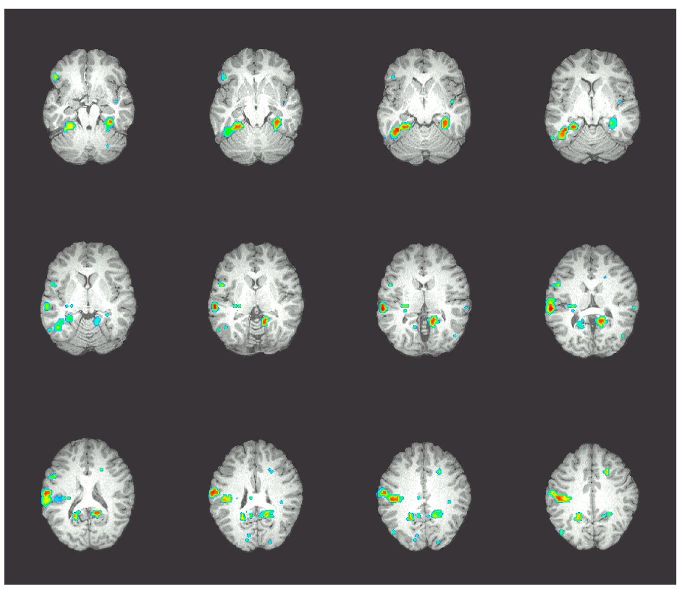
this is a plot of the μ's defining $P(X_i \mid Y=\text{"bottle"})$

Classification task: is person viewing a “tool” or “building”?



Where is information encoded in the brain?

Accuracies of cubical 27-voxel classifiers centered at each significant voxel [0.7-0.8]

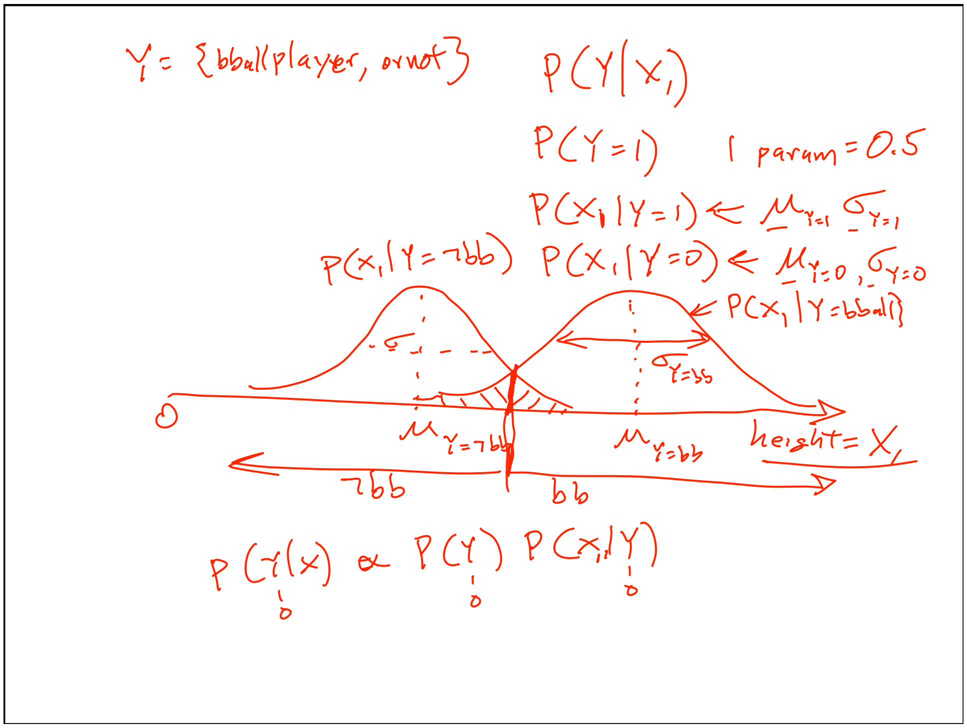


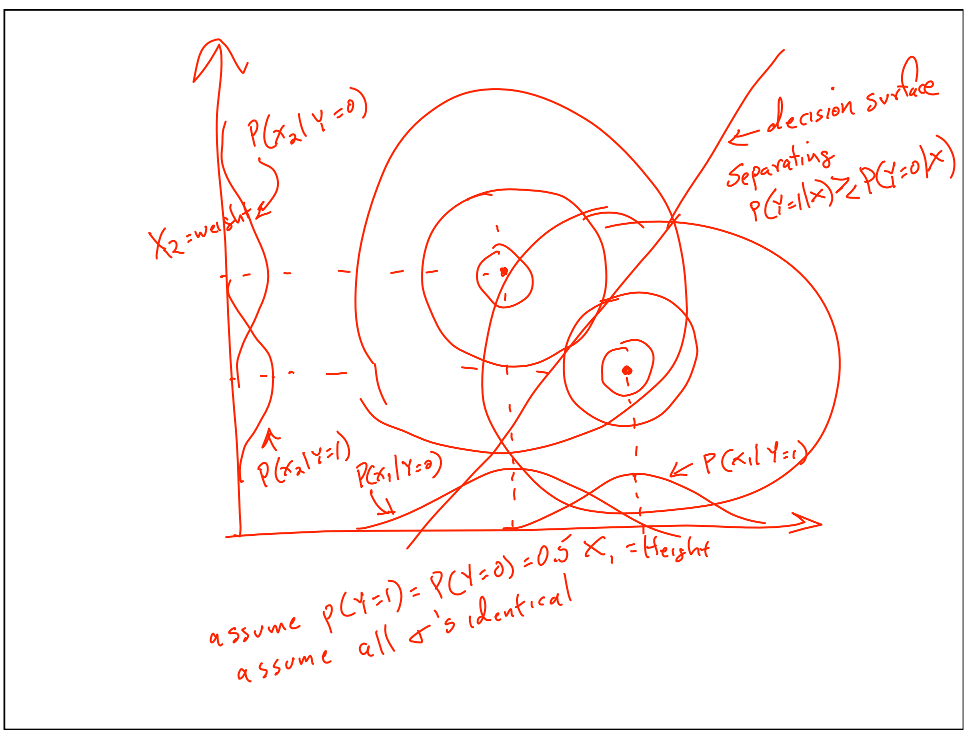
Naïve Bayes: What you should know

- Designing classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important
- Naïve Bayes assumption and its consequences
  - Which (and how many) parameters must be estimated under different generative models (different forms for P(X|Y))
    - and why this matters
- How to train Naïve Bayes classifiers
  - MLE and MAP estimates
  - with discrete and/or continuous inputs X_i

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?


Y = {bball player, or not}
P(Y|X1)
P(Y=1)
1 param = 0.5
P(X1|Y=1) ← μY=1 σY=1
P(X1|Y=¬bb)
P(X1|Y=0) ← μY=0, σY=0
P(X1|Y=bball)
σY=bb
0
μY=¬bb
μY=bb
height = X1
¬bb
bb
P(Y|X) ∝ P(Y) P(X1|Y)


P(X2|Y=0)
X2 = weight
decision surface
Separating
P(Y=1|X) ≷ P(Y=0|X)
P(X2|Y=1)
P(X1|Y=0)
P(X1|Y=1)
X1 = Height
assume P(Y=1) = P(Y=0) = 0.5
assume all σ's identical