

Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 25, 2011

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

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, classification rule for $X^{new} = \langle X_1, \dots, X_n \rangle$ 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): Boolean X_i

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

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

$$0 \geq \log \frac{P(Y=1|X_1 \dots X_n)}{P(Y=0|X_1 \dots X_n)} = \log \frac{P(Y=1)}{P(Y=0)} + \sum_i \log \left[\frac{P(X_i|Y=1)}{P(X_i|Y=0)} \right]$$

$$\hat{\theta}_{ik} = \hat{P}(x_i=1|Y=k) \quad 0 \geq \log \frac{P(Y=1)}{P(Y=0)} + \sum_i \left[x_i \log \frac{\theta_{i1}}{\theta_{i0}} + (1-x_i) \log \frac{(1-\theta_{i1})}{(1-\theta_{i0})} \right]$$

$$1 - \hat{\theta}_{ik} = \hat{P}(x_i=0|Y=k)$$

$P(S | D, G, M)$

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
- $X = \langle X_1, X_2, \dots, X_n \rangle = \text{document}$
- X_i is a random variable describing...

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

Learning to classify documents: $P(Y|X)$

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

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., $X_{\text{pleased}} = 1$
 $X_{\text{random}} = 0$
 \vdots

50000 of these

Issues? *cond indep assumption false!*

Learning to classify documents: $P(Y|X)$

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

X_1, X_2, X_3
 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_n*

- X_i is a random variable describing...

Answer 2:

- X_i represents the i^{th} word position in document
- $X_1 = \text{"I"}, X_2 = \text{"am"}, X_3 = \text{"pleased"}$
- 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
- $X = \langle X_1, X_2, \dots, X_n \rangle$ = 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~~⁵⁰⁰⁰⁰ outcomes instead of 2)

Likelihood is \sim Multinomial($\theta = \{\theta_1, \theta_2, \dots, \theta_k\}$)

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

count # of 1's count for side k

If prior is Dirichlet distribution,

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

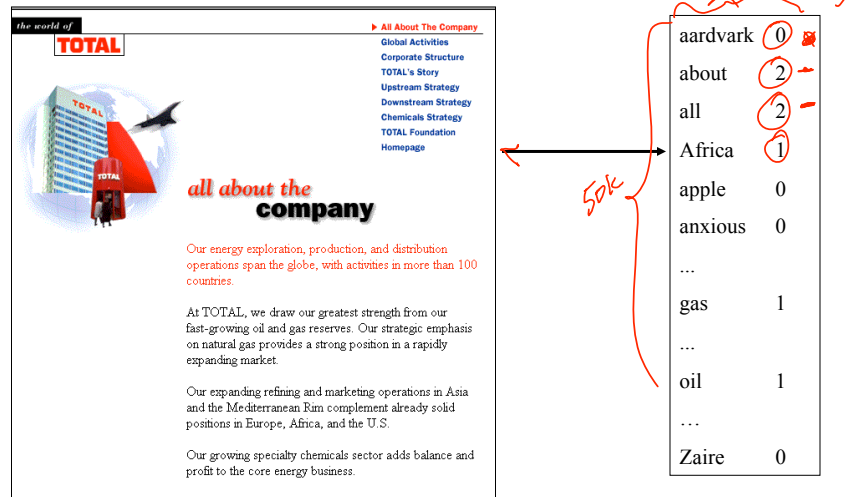
Then posterior is Dirichlet distribution

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

For Multinomial, conjugate prior is Dirichlet distribution.



Multinomial Bag of Words



MAP estimates for bag of words

Map estimate for multinomial

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

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

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

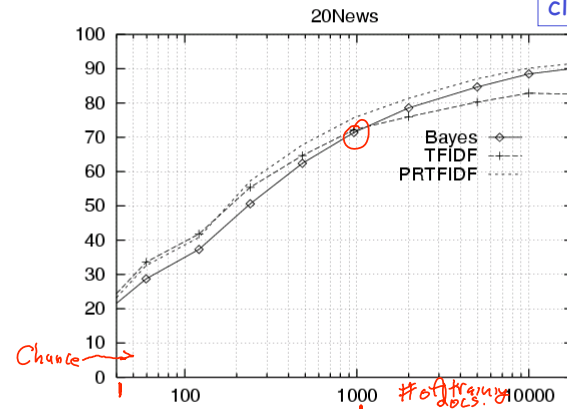
comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
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 i^{th} pixel



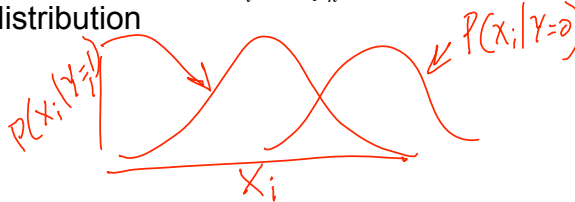
What if we have continuous X_i ?

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

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

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

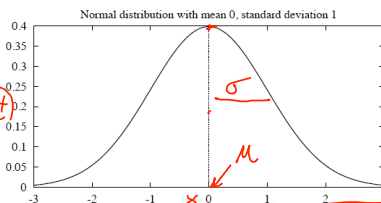
Common approach: assume $P(X_i | Y=y_k)$ follows a Normal (Gaussian) distribution



Gaussian Distribution

(also called "Normal")

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



$$E[X] = \sum_i x_i p(x_i)$$

$$E[X] = \int_{-\infty}^{\infty} x p(x) dx$$

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

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

$$V_{\text{var}}[X] = E[(X - E[X])^2]$$

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

$$E[X] = \mu$$

• Variance of X is

$$\text{Var}(X) = \sigma^2$$

• 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} , 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:

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

Diagram annotations:

- $\hat{\mu}_{ik}$ is labeled as "ith feature" and "kth class".
- X_i^j is labeled as "jth training example".
- $\delta(Y^j = y_k)$ is labeled as " $\delta()=1$ if $(Y^j=y_k)$ else 0".

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

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, $X = \langle X_1, \dots, X_n \rangle$?

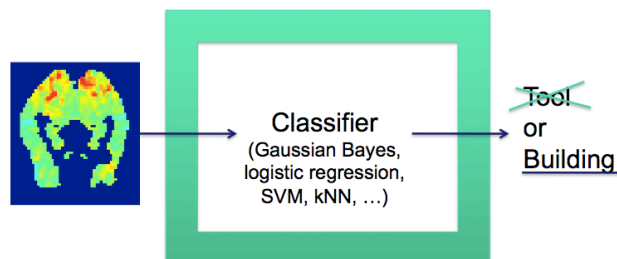
$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2}\left(\frac{x - \mu_{ik}}{\sigma_{ik}}\right)^2}$$

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

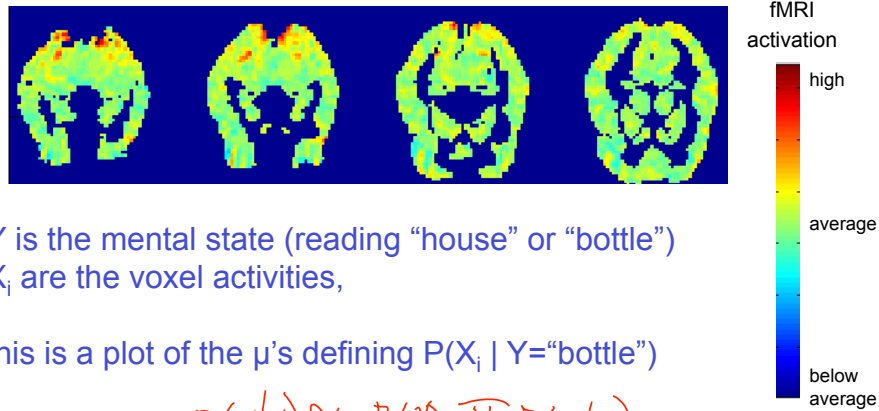
eg., if we assume attributes have same variance, indep of Y
($\sigma_{ik} = \sigma$)

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"

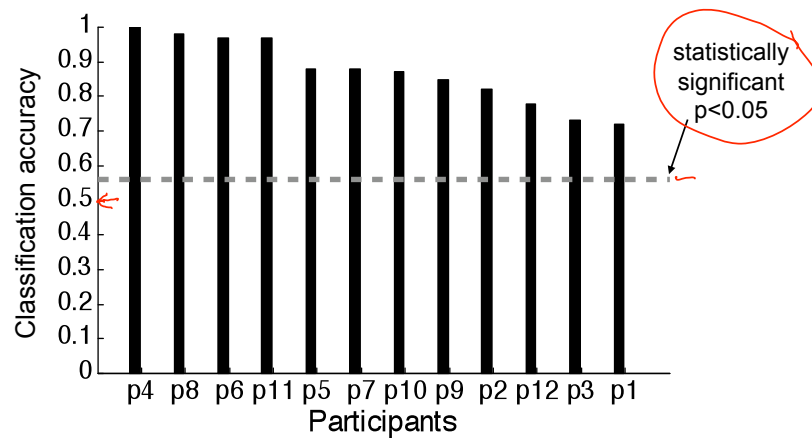


Y is the mental state (reading "house" or "bottle")
 X_i are the voxel activities,

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

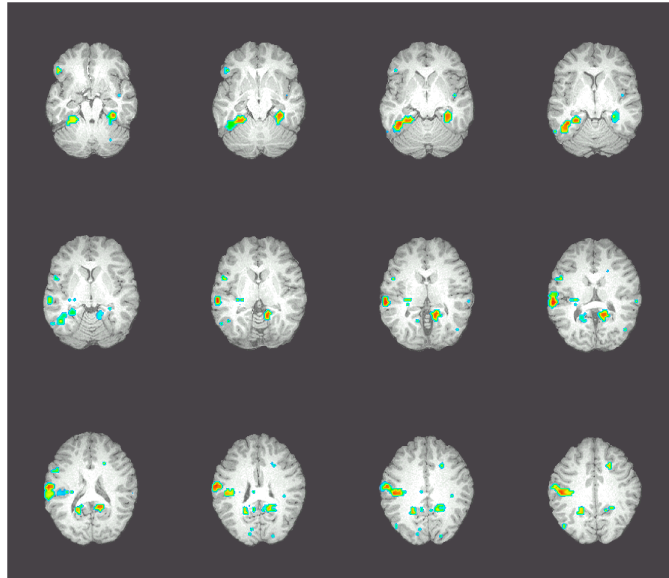
$$P(Y/X) \propto P(Y) \prod_i P(X_i | Y)$$

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?