10-601
Machine Learning

Naïve Bayes classifiers

Types of classifiers

- We can divide the large variety of classification approaches into three major types

1. Instance based classifiers
    - Use observation directly (no models)
    - e.g. K nearest neighbors

2. Generative:
    - build a generative statistical model
    - e.g., Bayesian networks

3. Discriminative
    - directly estimate a decision rule/boundary
    - e.g., decision tree

Bayes decision rule

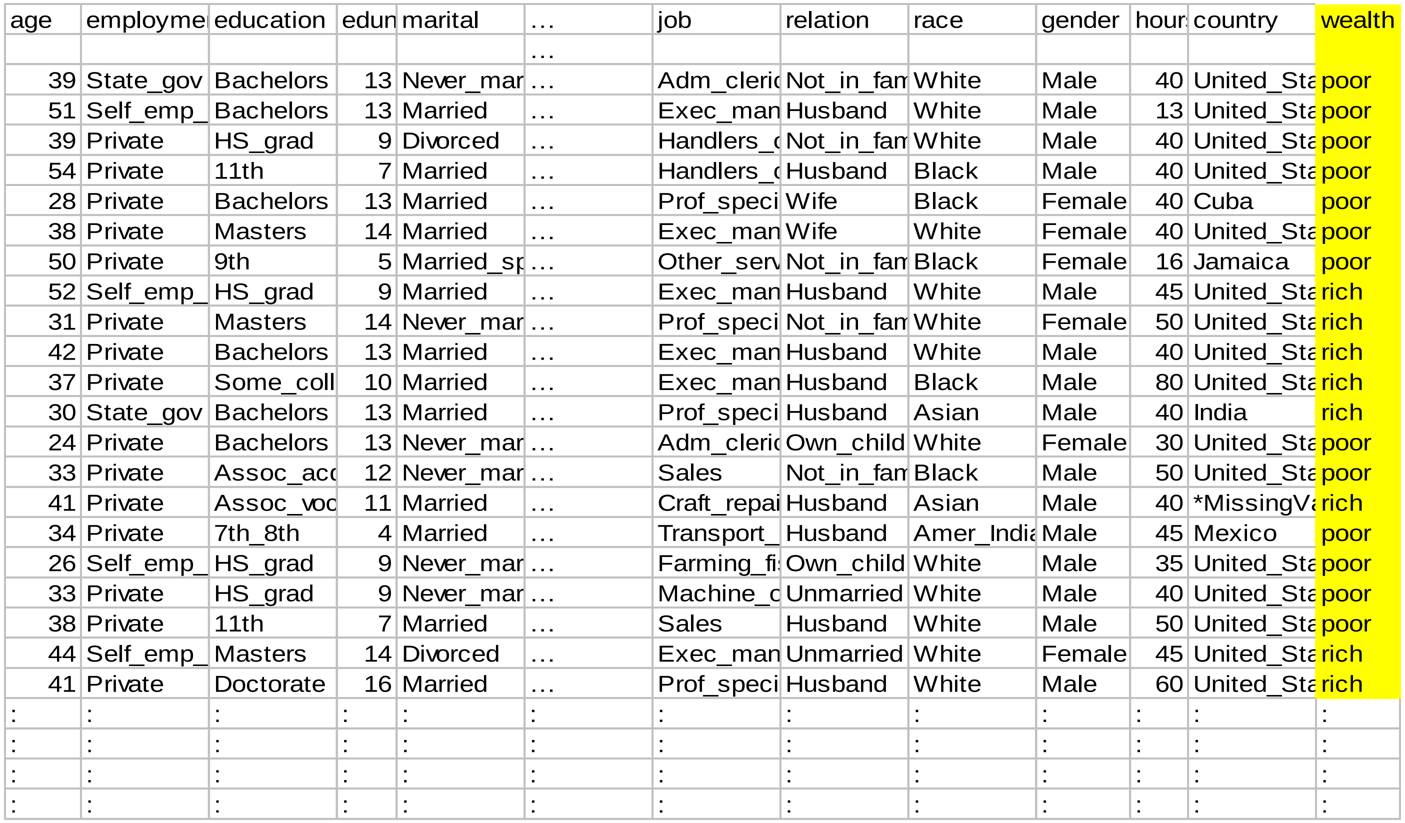
- If we know the conditional probability P(X | Y) we can determine the appropriate class by using Bayes rule:

$$P(y=i\,|\,x)=\frac{P(x\,|\,y=i)P(y=i)}{P(x)}\overset{def}{=}q_i(x)$$

But how do we determine p(X|Y)?

Computing p(x|y)

- Consider a dataset with 16 attributes (lets assume they are all binary). How many values to we need to know to fully determine p(x|y)?



age	employme	education	edun	marital	…	job	relation	race	gender	hour	country	wealth
					…							
39	State_gov	Bachelors	13	Never_mar	…	Adm_cleric	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	…	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	…	Handlers_c	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	…	Handlers_c	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	…	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	…	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	…	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	…	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	…	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	…	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	…	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	…	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	…	Adm_cleric	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acc	12	Never_mar	…	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	…	Craft_repai	Husband	Asian	Male	40	*MissingVa	rich
34	Private	7th_8th	4	Married	…	Transport_	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	…	Farming_fi	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	…	Machine_o	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	…	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	…	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	…	Prof_speci	Husband	White	Male	60	United_Sta	rich
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:

Learning the values for the full conditional probability table would require enormous amounts of data

Naïve Bayes Classifier

- Naïve Bayes classifiers assume that given the class label (Y) the attributes are conditionally independent of each other:

$$p(x \mid y) = \prod_i p_i(x_i \mid y)$$

Product of probability terms

Specific model for atribute *i*

- Using this idea the full classification rule becomes:

$$
\begin{aligned}
\hat{y} &= \arg\max_v \, p(y = v \mid x) \\
&= \arg\max_v \frac{p(x \mid y = v) p(y = v)}{p(x)} \\
&= \arg\max_v \prod_i p_i(x_i \mid y = v) p(y = v)
\end{aligned}
$$

Data likelihood

- The biggest advantage of the Naïve Bayes classifier is that it requires relatively little data for estimating the parameters.
- The conditional independence assumption means that we can to learn the parameters independently for each feature / attribute
- The global likelihood of the data can be expressed as:

$$L(X,Y) = L(X \mid Y)L(Y)$$

- Since the two parts of this product do not share parameters we can maximize them separately.

Data likelihood

- The global likelihood of the data can be expressed as:

$$L(X,Y) = L(X \mid Y)L(Y)$$

- Since the two parts of this product do not share parameters we can maximize them separately.
- For binary attributes, assume we observe n0 instances of x_i=0 and n1 instances of x_i=1 for class 1 then:

$$L(X_i \mid y=1) = \prod_{j \quad s.t. \quad y^j=1} p_i(x_i^j \mid y=1) = \theta_{i|y=1}^{n0}(1-\theta_{i|y=1})^{n1}$$

- where $\theta_{i|y=1}$ is the conditional probability that x_i = 0 given the class y=1

attribute i in sample j

Likelihood for all the i'th attributes for samples in class 1

Learning the parameters

$$L(X_i \mid y=1) = \prod_i \theta_{i|y=1}^{n0}(1-\theta_{i|y=1})^{n1}$$

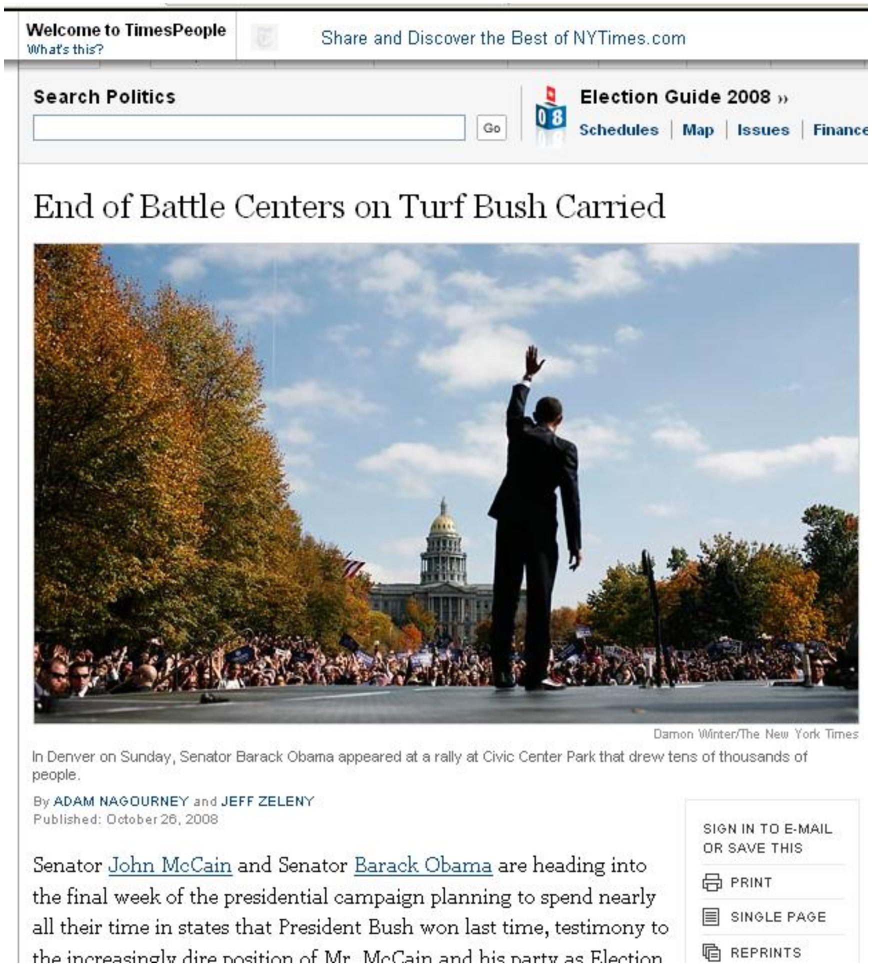
- To learn the parameters for this model we look for the value of $\theta_{i|y=1}$ that maximizes the likelihood of the data given the model.
- In turns out (see the density estimation class) that the MLE estimator is:

$$\theta_{i|y=1} = \frac{n0}{n0+n1}$$

- We repeat this for each attribute in each of the two classes and obtain the full set of conditional probabilities
- How do we compute p(y=1)?

Example: Text classification

- What is the major topic of this article?


Welcome to TimesPeople
What's this?
Share and Discover the Best of NYTimes.com
Search Politics
Go
Election Guide 2008 »
Schedules | Map | Issues | Finance
SIGN IN TO E-MAIL OR SAVE THIS
PRINT
SINGLE PAGE
REPRINTS

End of Battle Centers on Turf Bush Carried

Damon Winter/The New York Times

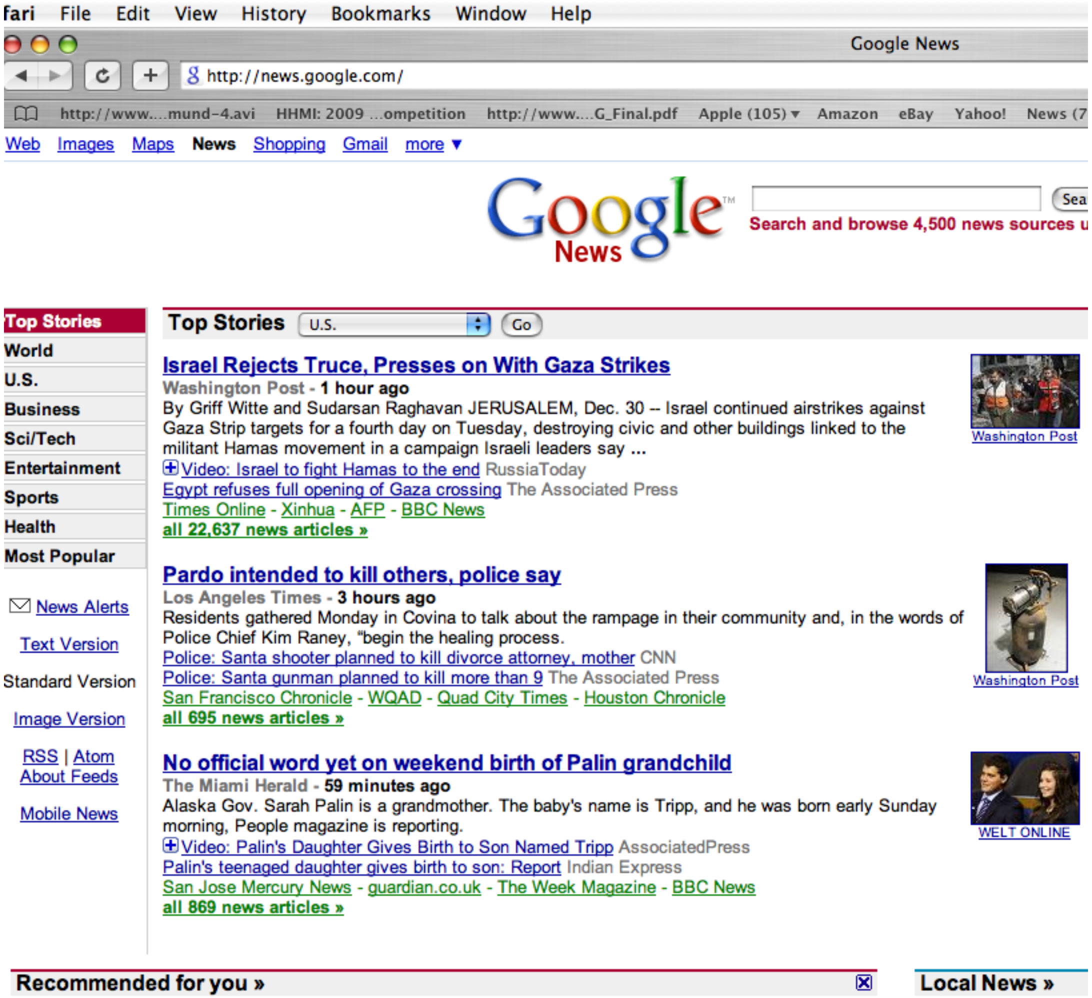
In Denver on Sunday, Senator Barack Obama appeared at a rally at Civic Center Park that drew tens of thousands of people.

By ADAM NAGOURNEY and JEFF ZELENY
Published: October 26, 2008

Senator John McCain and Senator Barack Obama are heading into the final week of the presidential campaign planning to spend nearly all their time in states that President Bush won last time, testimony to the increasingly dire position of Mr. McCain and his party as Election

Example: Text classification

- Text classification is all around us



Israel Rejects Truce, Presses on With Gaza Strikes
Washington Post - 1 hour ago
By Griff Witte and Sudarsan Raghavan JERUSALEM, Dec. 30 -- Israel continued airstrikes against Gaza Strip targets for a fourth day on Tuesday, destroying civic and other buildings linked to the militant Hamas movement in a campaign Israeli leaders say ...
Video: Israel to fight Hamas to the end RussiaToday
Egypt refuses full opening of Gaza crossing The Associated Press
Times Online - Xinhua - AFP - BBC News
all 22,637 news articles »

Pardo intended to kill others, police say
Los Angeles Times - 3 hours ago
Residents gathered Monday in Covina to talk about the rampage in their community and, in the words of Police Chief Kim Raney, "begin the healing process.
Police: Santa shooter planned to kill divorce attorney, mother CNN
Police: Santa gunman planned to kill more than 9 The Associated Press
San Francisco Chronicle - WQAD - Quad City Times - Houston Chronicle
all 695 news articles »

No official word yet on weekend birth of Palin grandchild
The Miami Herald - 59 minutes ago
Alaska Gov. Sarah Palin is a grandmother. The baby's name is Tripp, and he was born early Sunday morning, People magazine is reporting.
Video: Palin's Daughter Gives Birth to Son Named Tripp AssociatedPress
Palin's teenaged daughter gives birth to son: Report Indian Express
San Jose Mercury News - guardian.co.uk - The Week Magazine - BBC News
all 869 news articles »

Feature transformation

- How do we encode the set of features (words) in the document?
- What type of information do we wish to represent? What can we ignore?
- Most common encoding: ‘Bag of Words’
- Treat document as a collection of words and encode each document as a vector based on some dictionary
- The vector can either be binary (present / absent information for each word) or discrete (number of appearances)

- Google is a good example
- Other applications include job search adds, spam filtering and many more.

Feature transformation: Bag of Words

- In this example we will use a binary vector
- For document x_j we will use a vector of *m** indicator features $\{\phi^j(x_j)\}$ for whether a word appears in the document
  - $\phi^j(x_j) = 1$, if word *j* appears in document x_j; zero otherwise
- $\Phi(x_j) = [\phi^1(x_j) \ldots \phi^m(x_j)]^T$ is the resulting feature vector for the entire dictionary
- For notational simplicity we will replace each document x_i with a fixed length vector $\Phi_i = [\phi^1 \ldots \phi^m]^T$, where $\phi^j = \phi^j(x_j)$.

*The size of the vector for English is usually ~10000 words

Example

Assume we would like to classify documents as election related or not.

Dictionary

- Washington
- Congress

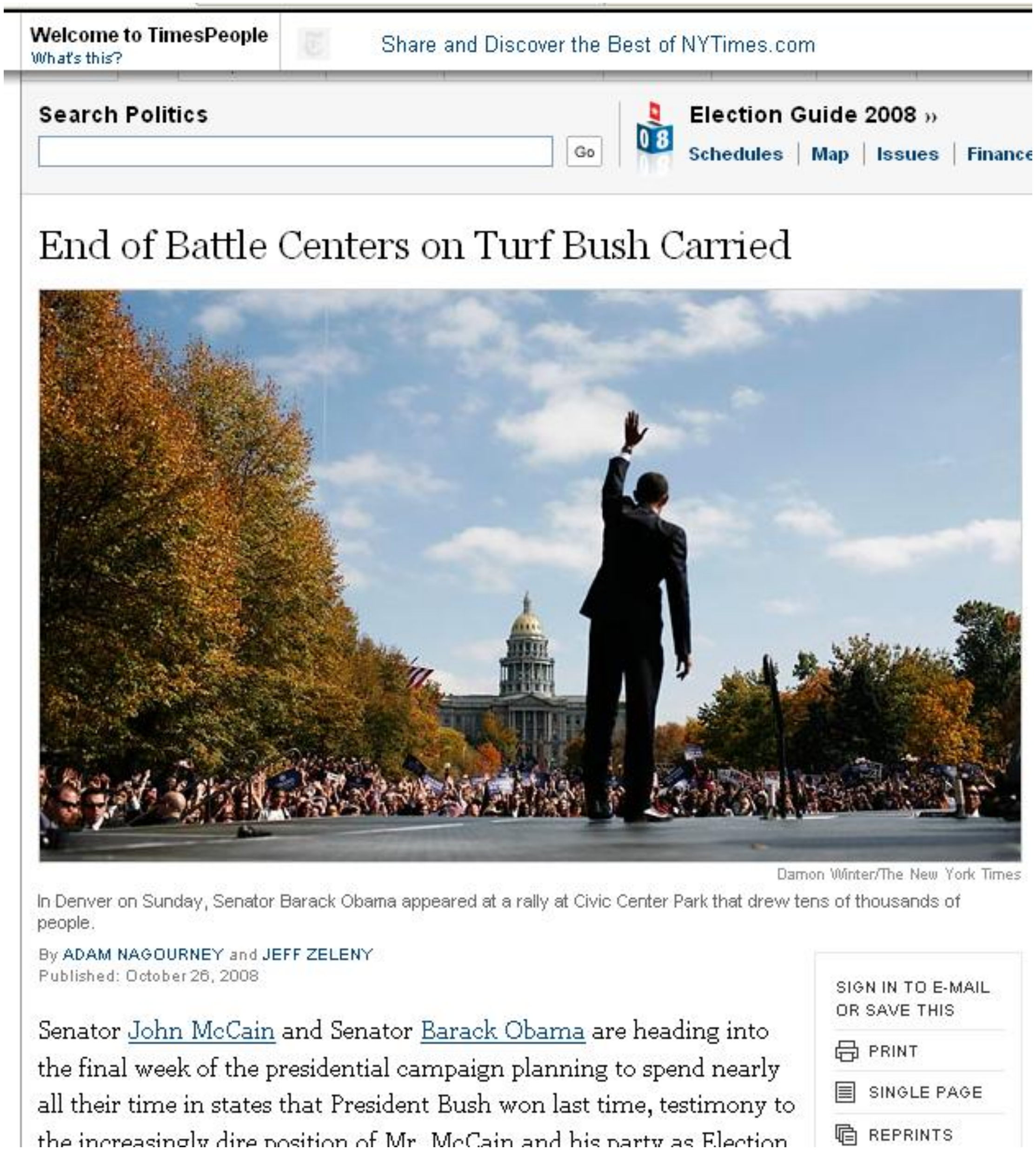
…

54. Mccain

55. Obama

56. Nader

$$\phi_{54}=\phi_{54}(x^j) = 1$$
$$\phi_{55}=\phi_{55}(x^j) = 1$$
$$\phi_{56}=\phi_{56}(x^j) = 0$$


Welcome to TimesPeople
What's this?
Share and Discover the Best of NYTimes.com
Search Politics
Go
08
Election Guide 2008 »
Schedules | Map | Issues | Finance
SIGN IN TO E-MAIL OR SAVE THIS
PRINT
SINGLE PAGE
REPRINTS

End of Battle Centers on Turf Bush Carried

Damon Winter/The New York Times

In Denver on Sunday, Senator Barack Obama appeared at a rally at Civic Center Park that drew tens of thousands of people.

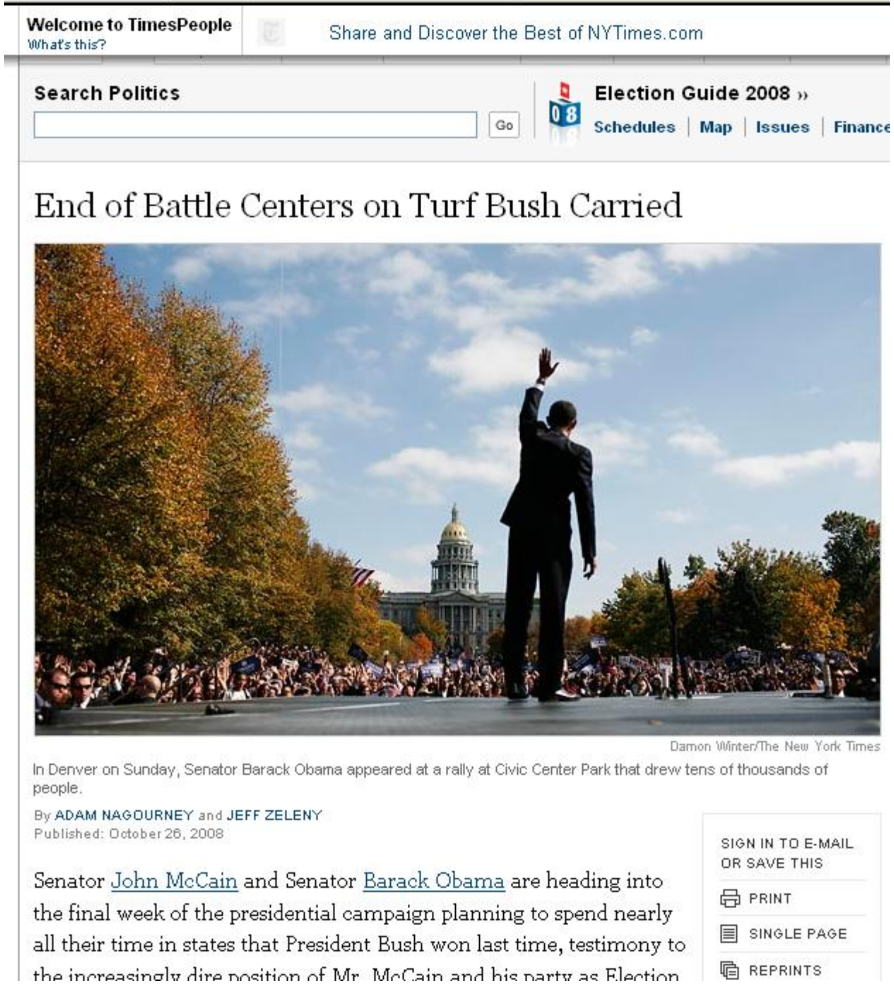
By ADAM NAGOURNEY and JEFF ZELENY
Published: October 26, 2008

Senator John McCain and Senator Barack Obama are heading into the final week of the presidential campaign planning to spend nearly all their time in states that President Bush won last time, testimony to the increasingly dire position of Mr. McCain and his party as Election

Example: cont.

We would like to classify documents as election related or not.

- Given a collection of documents with their labels (usually termed ‘training data’) we learn the parameters for our model.
- For example, if we see the word ‘Bush’ in n1 out of the n documents labeled as ‘election’ we set p(‘bush’|’election’)=n1/n
- Similarly we compute the priors (p(‘election’)) based on the proportion of the documents from both classes.


Welcome to TimesPeople
What's this?
Share and Discover the Best of NYTimes.com
Search Politics
Go
Election Guide 2008 »
Schedules | Map | Issues | Finance
SIGN IN TO E-MAIL OR SAVE THIS
PRINT
SINGLE PAGE
REPRINTS

End of Battle Centers on Turf Bush Carried

Damon Winter/The New York Times

In Denver on Sunday, Senator Barack Obama appeared at a rally at Civic Center Park that drew tens of thousands of people.

By ADAM NAGOURNEY and JEFF ZELENY
Published: October 26, 2008

Senator John McCain and Senator Barack Obama are heading into the final week of the presidential campaign planning to spend nearly all their time in states that President Bush won last time, testimony to the increasingly dire position of Mr. McCain and his party as Election

Example: Classifying Election (E) or Sports (S)

Assume learned the following model

$P(\phi_{\text{bush}} = 1 \mid E) = 0.8$, $P(\phi_{\text{bush}} = 1 \mid S) = 0.1$ $P(S) = 0.5$

$P(\phi_{\text{obama}} = 1 \mid E) = 0.9$, $P(\phi_{\text{obama}} = 1 \mid S) = 0.05$ $P(E) = 0.5$

$P(\phi_{\text{mccain}} = 1 \mid E) = 0.9$, $P(\phi_{\text{mccain}} = 1 \mid S) = 0.05$

$P(\phi_{\text{football}} = 1 \mid E) = 0.1$, $P(\phi_{\text{football}} = 1 \mid S) = 0.7$

For a specific document we have the following feature vector

$\phi_{\text{bush}} = 1$ $\phi_{\text{obama}} = 1$ $\phi_{\text{mccain}} = 1$ $\phi_{\text{football}} = 0$

$P(Y = E \mid 1,1,1,0) \propto 0.8*0.9*0.9*0.9*0.5 = 0.5832$

$P(Y = S \mid 1,1,1,0) \propto 0.1*0.05*0.05*0.3*0.5 = 0.000075$

So the document is classified as ‘Election’

Naïve Bayes classifiers for continuous values

- So far we assumed a binomial or discrete distribution for the data given the model (p(x_i|y))
- However, in many cases the data contains continuous features:

  - Height, weight

  - Levels of genes in cells

  - Brain activity
- For these types of data we often use a Gaussian model
- In this model we assume that the observed input vector X is generated from the following distribution

$$X \sim N(\mu, \Sigma)$$

Gaussian Bayes Classification

$$P(y=v \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y=v)P(y=v)}{p(\mathbf{x})}$$

- To determine the class when using the Gaussian assumption we need to compute p(x|y):

$$P(x \mid y) = \frac{1}{(2\pi)^{1/2} \mid \Sigma \mid^{1/2}} \exp\left[(X-\mu)^T \Sigma^{-1} (X-\mu)\right]$$

Once again, we need lots of data to compute the values of the covariance matrix Σ

Gaussian Bayes Classification

- Once again we use the Naïve Bayes assumption: Attributes are independent given the class label
- In the Gaussian model this means that the covariance matrix becomes a diagonal matrix with zeros everywhere except for the diagonal
- Thus, we only need to learn the values for the variance term for each attribute: $x_i \sim N(\boldsymbol{\mu}_i, \sigma_i^2)$.

$$P(X \mid y = v) = \prod_i \frac{1}{(2\pi)^{1/2}\sigma_{i,v}} \exp\left[-\frac{\left(\mathbf{x}_i - \mu_{i,v}\right)^2}{2\sigma_{i,v}^2}\right]$$

Separate means and variance for each class

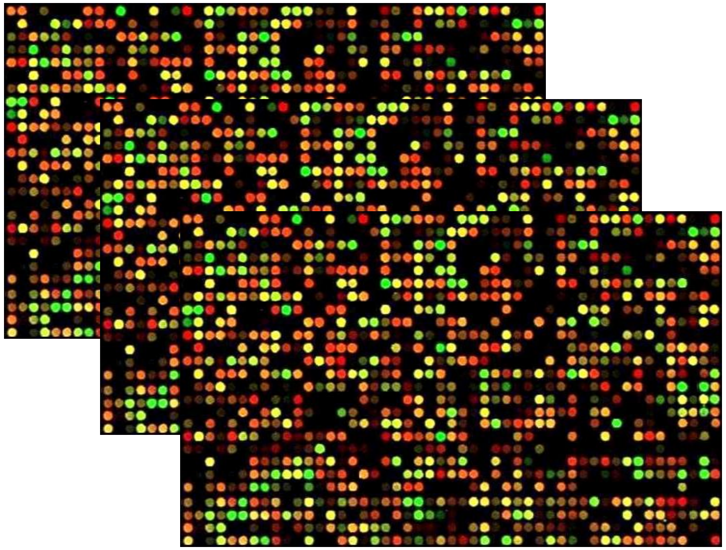
MLE for Gaussian Naïve Bayes Classifier

- For each class we need to estimate one global value (prior) and two values for each feature (mean and variance)
- The prior is computed in the same way we did before (counting) which is the MLE estimate For each feature
- The MLE for mean and variance is computed by setting:

$$\mu_{i,0} = \frac{1}{n0}\sum_{j=1}^{n} x_i^j \qquad \sigma_{i,0}^2 = \frac{1}{n0}\sum_{j=1}^{n}\left(x_i^j - \mu_{i,0}\right)^2$$

Example: Classifying gene expression data

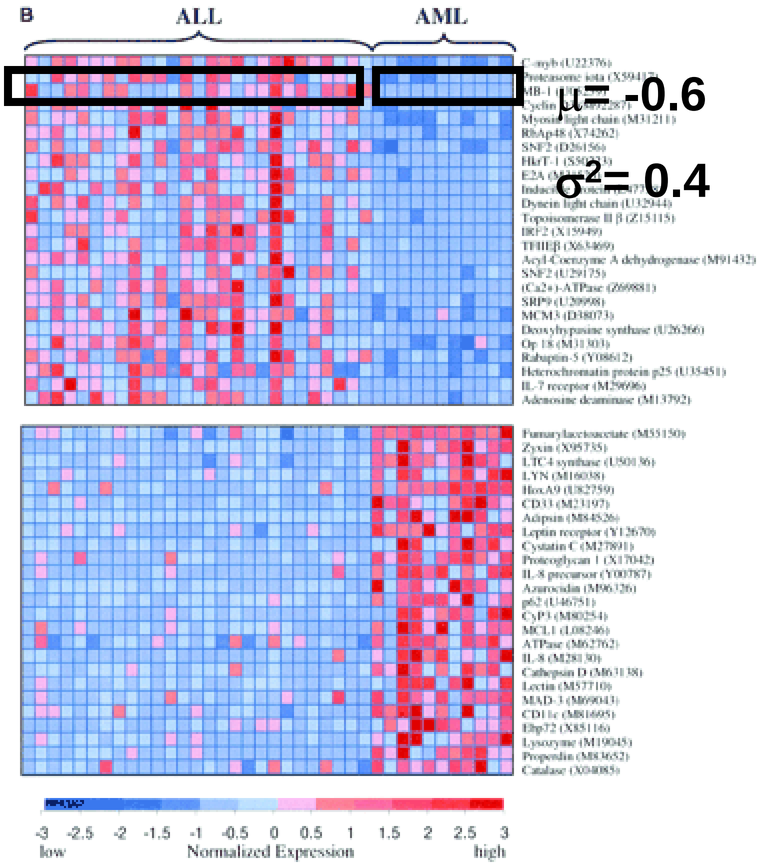
- Measures the levels (up or down) of genes in our cells
- Differs between healthy and sick people and between different disease types
- Given measurement of patients with two different types of cancer we would like to generate a classifier to distinguish between them



Classifying cancer types

- We select a subset of the genes (more in our 'feature selection' class later in the course).

- We compute the mean for each of the genes in each of the classes



μ= 1.8

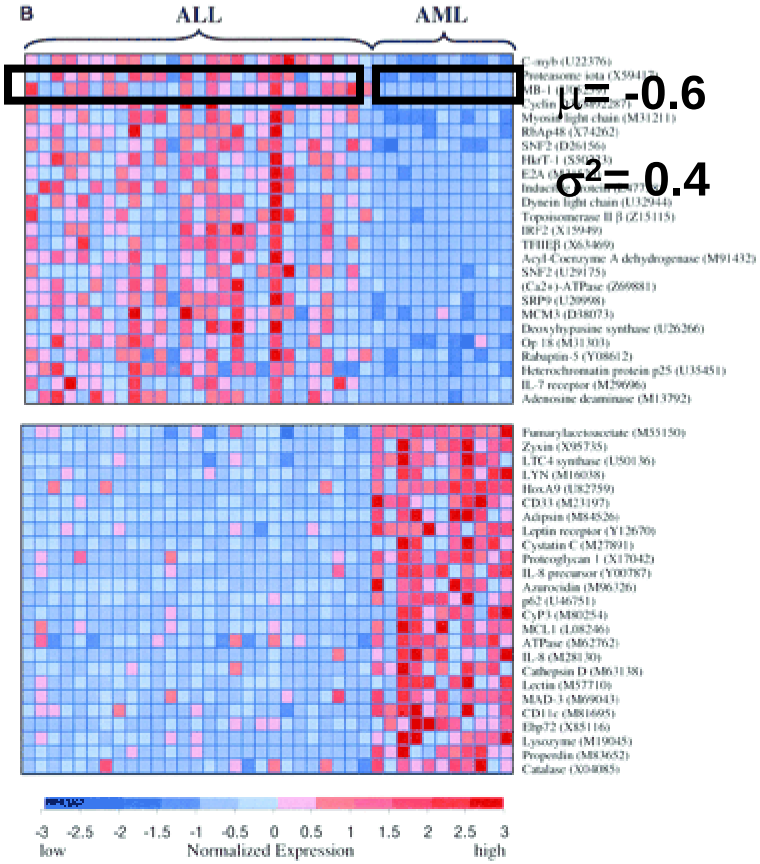
σ^2= 1.1

μ= -0.6

σ^2= 0.4

Classifying cancer types

- We compute the mean for each of the genes in each of the classes
- We determine the prior for each class
- We can compute our decision for each of the samples we have labels for



μ= 1.8

σ^2= 1.1

μ= -0.6

σ^2= 0.4

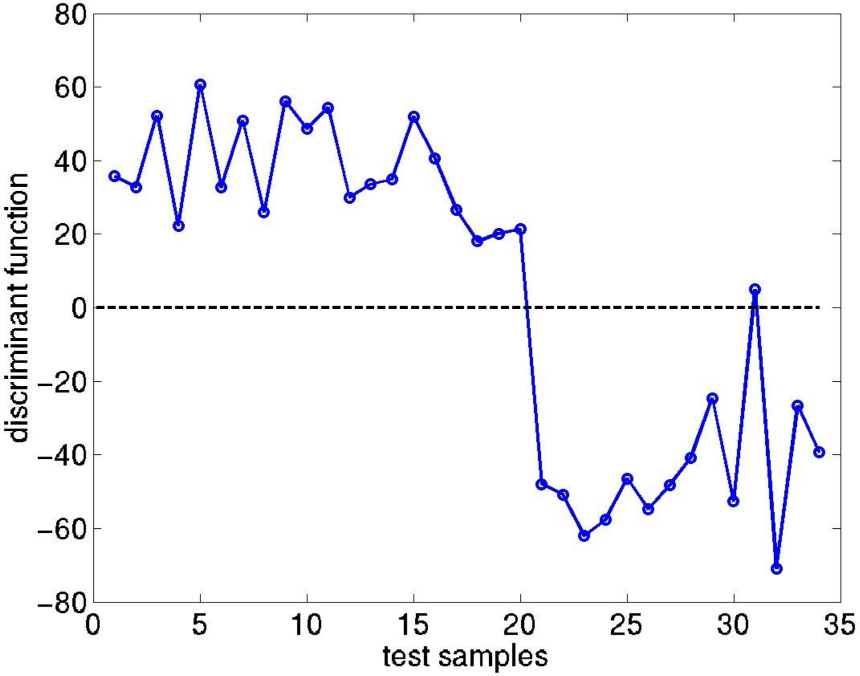
Classification accuracy

- The figure shows the value of the discriminate function

$$f(x) = \log \frac{p(Y=1 \mid X)}{p(Y=0 \mid X)}$$

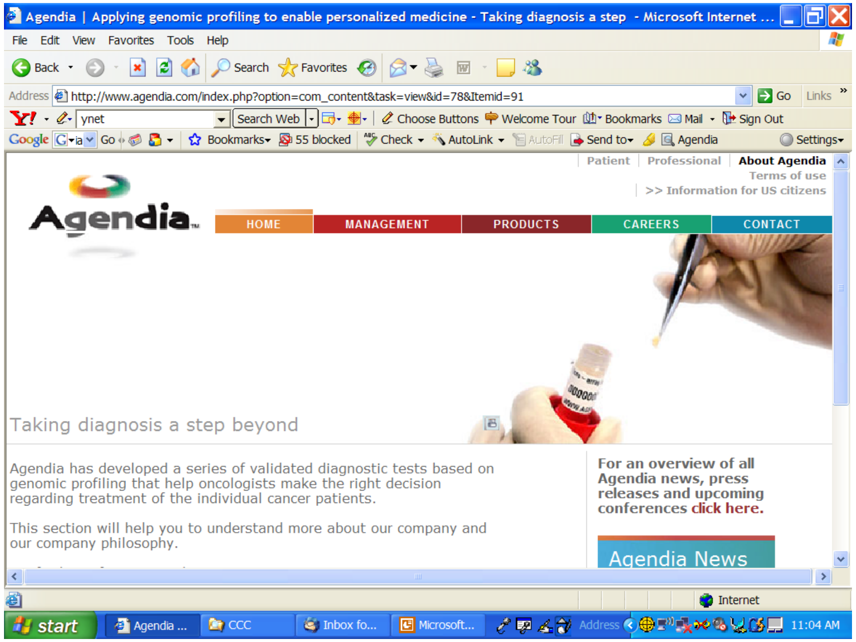
across the test examples

- The only test error is also the decision with the lowest confidence



FDA Approves Gene-Based Breast Cancer Test*

“ MammaPrint is a DNA microarray-based test that measures the activity of 70 genes... The test measures each of these genes in a sample of a woman's breast-cancer tumor and then uses a specific formula to determine whether the patient is deemed low risk or high risk for the spread of the cancer to another site.”



Agendia | Applying genomic profiling to enable personalized medicine - Taking diagnosis a step - Microsoft Internet ...

Address http://www.agendia.com/index.php?option=com_content&task=view&id=78&Itemid=91

Taking diagnosis a step beyond

Agendia has developed a series of validated diagnostic tests based on genomic profiling that help oncologists make the right decision regarding treatment of the individual cancer patients.

This section will help you to understand more about our company and our company philosophy.

For an overview of all Agendia news, press releases and upcoming conferences click here.

*Washington Post, 2/06/2007

Possible problems with Naïve Bayes classifiers: Assumptions

- In most cases, the assumption of conditional independence given the class label is violated

  - much more likely to find the word ‘George’ if we saw the word ‘Bush’ regardless of the class

- This is, unfortunately, a major shortcoming which makes these classifiers inferior in many real world applications (though not always)
- There are models that can improve upon this assumption without using the full conditional model (one such model are Bayesian networks which we will discuss later in this class).

Possible problems with Naïve Bayes classifiers: Parameter estimation

- Even though we need far less data than the full Bayes model, there may be cases when the data we have is not enough
- For example, what is p(S=1,N=1|E=2)?
- This can get worst. Assume we have 20 variables, almost all pointing in the direction of the same class except for one for which we have no record for this class.
- Solutions?

Summer?	Num > 20	Evaluation
1	1	3
1	0	3
0	1	2
0	1	1
0	0	3
1	1	1

Important points

- Problems with estimating full joints
- Advantages of Naïve Bayes assumptions
- Applications to discrete and continuous cases
- Problems with Naïve Bayes classifiers