Bayesian Classifiers with Applications to Text

Professor Daphne Koller
(substituting for Prof. Manning)

Joint Distribution

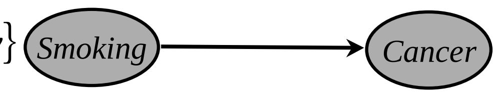
Smoking and Cancer

$S \in \{no, light, heavy\}$ $C \in \{none, benign, malignant\}$

$S\Downarrow$ $C\Rightarrow$	*none*	*benign*	*malignant*
no	0.768	0.024	0.008
light	0.132	0.012	0.006
heavy	0.035	0.010	0.005

Joint Distribution

$S \in \{no, light, heavy\}$



$C \in \{none, benign, malignant\}$

P(S=no)	0.80
P(S=light)	0.15
P(S=heavy)	0.05

P(C | S)

	Smoking		
	no	*light*	*heavy*
C=none	0.96	0.88	0.60
C=benign	0.03	0.08	0.25
C=malig	0.01	0.04	0.15

Product Rule

- *P(C,S) = P(C|S) P(S)*

S⇓ *C*⇒	*none*	*benign*	*malignant*
no	0.768	0.024	0.008
light	0.132	0.012	0.006
heavy	0.035	0.010	0.005

Marginalization

$S\Downarrow$ $C\Rightarrow$	*none*	*benign*	*malig*	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

P(Smoke) (row totals)

P(Cancer) (column totals)

Bayes Rule

$$P(S|C) = \frac{P(C,S)}{P(C)} = \frac{P(C|S)P(S)}{P(C)}$$

$S\Downarrow$ $C\Rightarrow$	*none*	*benign*	*malig*
no	0.768/.935	0.024/.046	0.008/.019
light	0.132/.935	0.012/.046	0.006/.019
heavy	0.030/.935	0.015/.046	0.005/.019

Cancer=	*none*	*benign*	*malignant*
P(S=no)	0.821	0.522	0.421
P(S=light)	0.141	0.261	0.316
P(S=heavy)	0.037	0.217	0.263

Bayes Rule

$$P(C,X) = P(C \mid X)P(X) = P(X \mid C)P(C)$$

$$\boxed{P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}}$$

The Classification Problem

- From a data set describing objects by vectors of *features* and a *class*

Age Sex ChestPain RestBP Cholesterol BloodSugar ECG MaxHeartRt Angina OldPeak Heart Disease

Vector$_1$= <49, 0, 2, 134, 271, 0, 0, 162, 0, 0, 2, 0, 3> Presence
Vector$_2$= <42, 1, 3, 130, 180, 0, 0, 150, 0, 0, 1, 0, 3> Presence
Vector$_3$= <39, 0, 3, 94, 199, 0, 0, 179, 0, 0, 1, 0, 3 > Presence
Vector$_4$= <41, 1, 2, 135, 203, 0, 0, 132, 0, 0, 2, 0, 6 > Absence
Vector$_5$= <56, 1, 3, 130, 256, 1, 2, 142, 1, 0.6, 2, 1, 6 > Absence
Vector$_6$= <70, 1, 2, 156, 245, 0, 2, 143, 0, 0, 1, 0, 3 > Presence
Vector$_7$= <56, 1, 4, 132, 184, 0, 2, 105, 1, 2.1, 2, 1, 6 > Absence

- Find a function *F*: *features* $\rightarrow$ *class* to <u>*classify*</u> a new object

Bayes-Optimal Classifiers

- **Assumption:** The data instances we see are generated from some probability distribution
  $P(X_1,\ldots,X_n,C)$
- Consider instance *x*, let
  - *c* be its ***true*** class,
  - ℓ be the class returned by the classifier *F*.
- The classifier is <u>correct</u> if $c = \ell$, and in <u>error</u> if $c \neq \ell$.
  - define $\lambda(c = \ell) = 0$ if $c = \ell$ and 1 otherwise
- The expected error incurred by choosing label ℓ is

$$\sum_{i=1}^{n} \lambda(c_i = \ell) P(c_i \mid \vec{x}) = 1 - P(\ell \mid \vec{x})$$

Bayes-Optimal Classifiers

- The expected error incurred by choosing label ℓ is

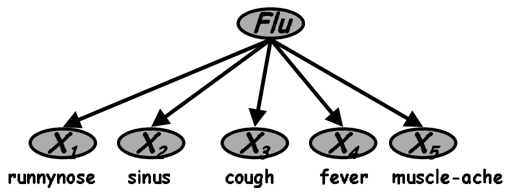
$$\sum_{i=1}^{n} \lambda(c_i = \ell) P(c_i \mid \vec{x}) = 1 - P(\ell \mid \vec{x})$$

- Thus, if we knew *P*, we could minimize error rate by choosing ℓ_i when

$$P(c_i \mid \vec{x}) > P(c_j \mid \vec{x}) \forall j \neq i$$

- Bayes Optimal Classifier:
  - Given a new instance $\langle x_1,\ldots,x_n \rangle$
    Set: $c = argmax_C\, P(C = c \mid x_1,\ldots,x_n)$

The Naïve Bayes Classifier



- **Assumption:** features are independent of each other given the class.

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

Naïve Bayes Classification

$$\arg\max_c P(c \mid x_1,\ldots,x_n)$$

$$\frac{P(x_1,\ldots,x_n \mid c)P(c)}{P(x_1,\ldots,x_n)}$$

$$= \arg\max_c P(x_1,\ldots,x_n \mid c)P(c)$$

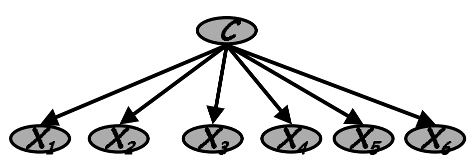
$$= \arg\max_c P(x_1 \mid c) \bullet \cdots \bullet P(x_n \mid c)P(c)$$

$$= \arg\max_c P(c)\prod_i P(x_i \mid c)$$

Naïve Bayes Algorithm

- Learn: Input = Data Set, output =
  - For each class c_j :
    - estimate
      $\hat{P}(c_j)$
    - For each attribute value x_i of each attribute X_i
      - estimate
        $\hat{P}(x_i \mid c_j)$
- Classify new instance $\langle x_1, \ldots, x_n \rangle$ as
  $$\ell = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Learning the Model

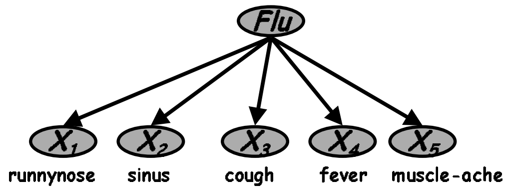


- Common practice:maximum likelihood
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

Problem with Max Likelihood



$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

- What if we have seen no training cases where patient had no flu and muscle aches?

$$\hat{P}(X_5 = t \mid C = nf) = \frac{N(X_5 = t, C = nf)}{N(C = nf)} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$\ell = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Smoothing to Avoid Overfitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

\# of values of X_i

- Somewhat more subtle version

overall fraction in data where $X_i = x_{i,k}$

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

extent of "smoothing"

Conditional Independence

- Conditional independence assumption is typically false
  - Sinus condition not independent of runny nose, even given flu
- Nevertheless, it works surprisingly well
  - Reason 1: small number of parameters
    - if we try to fit too many parameters with sparse data, can get really strange models
  - Reason 2: Don't need probabilities to be correct, only argmax

$$\arg\max_c \hat{P}(c)\prod_i \hat{P}(x_i \mid c) = \arg\max_c P(c)\prod_i P(x_i \mid c)$$

Text Classification

- Input: Document consisting of words
- Output: Classification into a set of classes

- Examples:
  - learn which news articles are "interesting"
  - learn to classify webpages by topic

- Naïve Bayes is surprisingly good at this task

Two Models

- Model 1: Multi-variate binomial
  - One feature X_w for each word in dictionary
  - X_w = true in document *d* if *w* appears in *d*
  - Naïve Bayes assumption:
    - Given the document's topic, appearance of one word in document tells us nothing about chances that another word appears

Two Models

- Model 2: Multinomial
  - One feature X_i for each word in document
    - feature values are all words in dictionary
  - Value of X_i is the word in position *i*
  - Naïve Bayes assumption:
    - Given the document's topic, word in one position in document tells us nothing about value of words in other positions
  - Second assumption:
    - word appearance does not depend on position

$$P(X_i = w \mid c) = P(X_j = w \mid c)$$

for all positions *i,j*, word *w*, and class *c*

Parameter estimation

- Binomial model:

$$\hat{P}(X_w = t \mid c_j) = \text{fraction of documents of topic } c_j \text{ in which word } w \text{ appears}$$

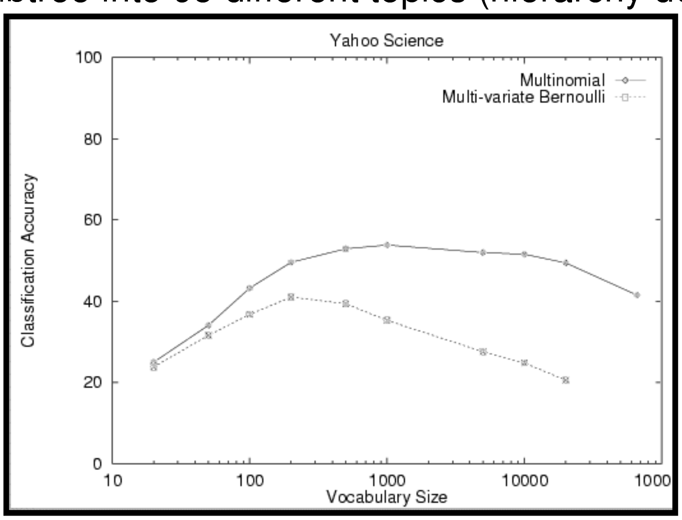
- Multinomial model:

$$\hat{P}(X_i = w \mid c_j) = \text{fraction of times in which word } w \text{ appears across all documents of topic } c_j$$

  - creating a mega-document for topic *j* by concatenating all documents in this topic
  - use frequency of *w* in mega-document

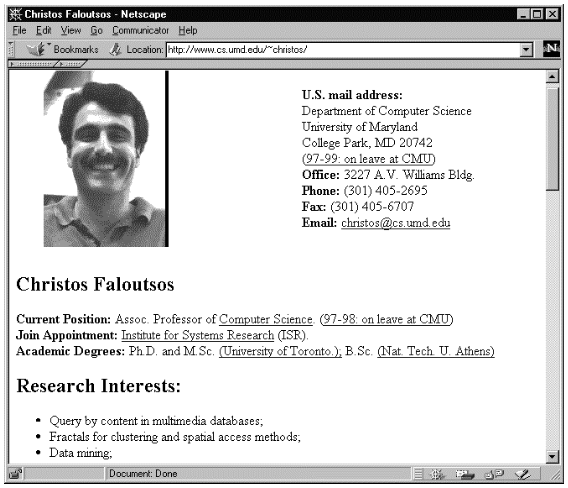
Example: AutoYahoo!

- Classify 13589 Yahoo! webpages in "Science" subtree into 95 different topics (hierarchy depth 2)



Example: WebKB (CMU)

- Classify webpages from CS departments into:
  - student, faculty, course,project

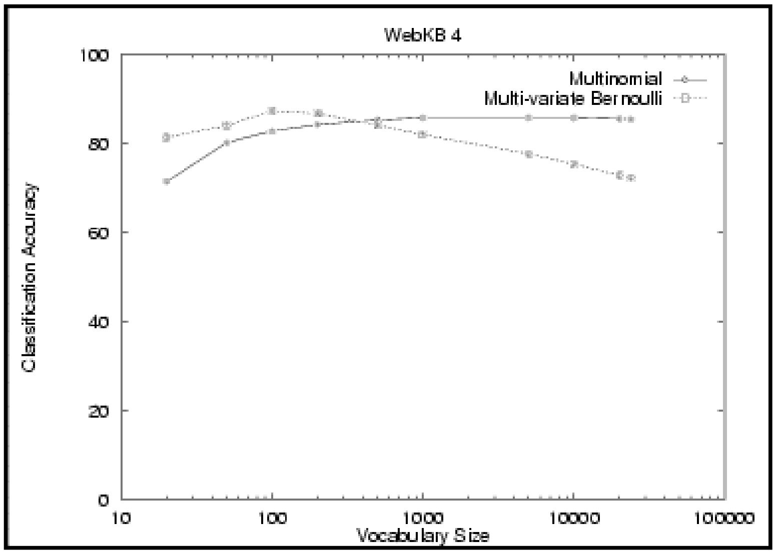


WebKB Experiment

- Train on ~5,000 hand-labeled web pages
  - Cornell, Washington, U.Texas, Wisconsin
- Crawl and classify a new site (CMU)
- Results:

	Student	Faculty	Person	Project	Course	Departmt
Extracted	180	66	246	99	28	1
Correct	130	28	194	72	25	1
Accuracy:	72%	42%	79%	73%	89%	100%

NB Model Comparison



Faculty

associate	0.00417
chair	0.00303
member	0.00288
ph	0.00287
director	0.00282
fax	0.00279
journal	0.00271
recent	0.00260
received	0.00258
award	0.00250

Students

resume	0.00516
advisor	0.00456
student	0.00387
working	0.00361
stuff	0.00359
links	0.00355
homepage	0.00345
interests	0.00332
personal	0.00332
favorite	0.00310

Courses

homework	0.00413
syllabus	0.00399
assignments	0.00388
exam	0.00385
grading	0.00381
midterm	0.00374
pm	0.00371
instructor	0.00370
due	0.00364
final	0.00355

Departments

departmental	0.01246
colloquia	0.01076
epartment	0.01045
seminars	0.00997
schedules	0.00879
webmaster	0.00879
events	0.00826
facilities	0.00807
eople	0.00772
postgraduate	0.00764

Research Projects

investigators	0.00256
group	0.00250
members	0.00242
researchers	0.00241
laboratory	0.00238
develop	0.00201
related	0.00200
arpa	0.00187
affiliated	0.00184
project	0.00183

Others

type	0.00164
jan	0.00148
enter	0.00145
random	0.00142
program	0.00136
net	0.00128
time	0.00128
format	0.00124
access	0.00117
begin	0.00116