Generative Models

CS478 – Machine Learning
Spring 2008

Thorsten Joachims
Cornell University

Reading:
Mitchell, Chapter 6.9 - 6.10
Duda, Hart & Stork, Pages 20-27

Outline

- **Bayes decision rule**
- **Bayes theorem**
- **Generative vs. discriminative learning**
- **Two generative learning algorithms**
  - naive Bayes
  - linear discriminant analysis
- **Selecting hypotheses based on training data**
  - maximum likelihood
  - maximum a posteriori (MAP)
  - bayesian

Bayes Decision Rule

- **Assumption:**
  - learning task P(X,Y) is know
- **Question:**
  - Given instance x, how should it be classified to minimize prediction error?
- **Bayes Decision Rule:**

$$h_{bayes}(\vec{x}) = argmax_{y \in Y} \left[P(Y=y|X=\vec{x}) \right]$$

Generative vs. Discriminative Models

Process:
- Generator: Generate descriptions according to distribution *P(X)*.
- Teacher: Assigns a value to each description based on *P(Y|X)*.

Training Examples $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \sim P(X,Y)$

Discriminative Model
- Select classification rules *H* to consider (hypothesis space)
- Find *h* from *H* with lowest training error
- Argument: low training error leads to low prediction error
- Examples: SVM, decision trees, Perceptron

Generative Model
- Select set of distributions to consider for modeling *P(X,Y)*.
- Find distribution that best matches *P(X,Y)* on training data
- Argument: if match close enough, we can use Bayes' Decision rule
- Examples: naive Bayes, HMM

Bayes Theorem

- **It is possible to "switch" conditioning according to the following rule**
- **Given any two random variables X and Y, it holds that**

$$P(Y=y|X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$$

- **Note that**

$$P(X=x) = \sum_{y \in Y} P(X=x|Y=y)P(Y=y)$$

Naïve Bayes' Classifier (Multivariate)

- **Model for each class**

$$P(X=\vec{x}|Y=+1) = \prod_{i=1}^{N} P(X_i=x_i|Y=+1)$$

$$P(X=\vec{x}|Y=-1) = \prod_{i=1}^{N} P(X_i=x_i|Y=-1)$$

- **Prior probabilities**

$P(Y=+1)$ $\quad P(Y=-1)$

fever (3)	cough (2)	pukes (2)	flu?
high	yes	no	1
high	no	yes	1
low	yes	no	-1
low	yes	yes	1
high	no	yes	???

- **Classification rule:**

$$h_{naive}(\vec{x}) = argmax_{y \in \{+1,-1\}} \left\{ P(Y=y) \prod_{i=1}^{N} P(X_i=x_i|Y=y) \right\}$$

Estimating the Parameters of Naïve Bayes

- **Count frequencies in training data**
  - n: number of training examples
  - n_+ / n_- : number of pos/neg examples
  - $\#(X_i=x_i, y)$: number of times feature X_i takes value x_i for examples in class y
  - $|X_i|$: number of values attribute X_i can take

fever (3)	cough (2)	pukes (2)	flu?
high	yes	no	1
high	no	yes	1
low	yes	no	-1
low	yes	yes	1
high	no	yes	???

- **Estimating P(Y)**
  - Fraction of positive / negative examples in training data

$$\hat{P}(Y=1)=\frac{n_+}{n} \qquad \hat{P}(Y=-1)=\frac{n_-}{n}$$

- **Estimating P(X|Y)**
  - Maximum Likelihood Estimate

$$\hat{P}(X_i=x_i|Y=y)=\frac{\#(X_i=x_i,y)}{n_y}$$

  - Smoothing with Laplace estimate

$$\hat{P}(X_i=x_i|Y=y)=\frac{\#(X_i=x_i,y)+1}{n_y+|X_i|}$$

Linear Discriminant Analysis

- **Spherical Gaussian model with unit variance for each class**

$$P(X=\vec{x}|Y=+1)\sim e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_+)^2}$$

$$P(X=\vec{x}|Y=-1)\sim e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_-)^2}$$

- **Prior probabilities**

$$P(Y=+1) \qquad P(Y=-1)$$

- **Classification rule:**

$$\begin{aligned} h_{LDA}(\vec{x}) &= argmax_{y\in\{+1,-1\}}\left\{P(Y=y)e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_y)^2}\right\} \\ &= argmax_{y\in\{+1,-1\}}\left\{\log(P(Y=y))-\frac{1}{2}(\vec{x}-\vec{\mu}_y)^2\right\} \end{aligned}$$

- **Often called "Rocchio Algorithm" in Inform. Retrieval**

Estimating the Parameters of LDA

- **Count frequencies in training data**
  - $(\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n)\sim P(X,Y)$: training data
  - n: number of training examples
  - n_+ / n_-: number of positive/negative training examples
- **Estimating P(Y)**
  - Fraction of positive / negative examples in training data

$$\hat{P}(Y=1)=\frac{n_+}{n} \qquad \hat{P}(Y=-1)=\frac{n_-}{n}$$

- **Estimating class means**

$$\vec{\mu}_+=\frac{1}{n_+}\sum_{\{i:y_i=+1\}}\vec{x}_i \qquad \vec{\mu}_-=\frac{1}{n_-}\sum_{\{i:y_i=-1\}}\vec{x}_i$$

Naïve Bayes Classifier (Multinomial)

- **Application: Text classification**

text	CS?
$\vec{x}_1=(The, art, of, Programming)$	+1
$\vec{x}_2=(Introduction, to, Calculus)$	-1
$\vec{x}_3=(Introduction, to, Complexity, Theory)$	+1
$\vec{x}_4=(Introduction, to, Programming)$	??

- **Assumption (*l* words in document)**

$$P(X=\vec{x}|Y=+1)=\prod_{i=1}^{l}P(W=w_i|Y=+1)$$

$$P(X=\vec{x}|Y=-1)=\prod_{i=1}^{l}P(W=w_i|Y=-1)$$

- **Classification Rule**

$$h_{text}(\vec{x})=argmax_{y\in\{+1,-1\}}\left\{P(Y=y)\prod_{i=1}^{l}P(W=w_i|Y=y)\right\}$$

Estimating the Parameters of Naïve Bayes

- **Count frequencies in training data**
  - n: number of training examples
  - n_+ / n_- : number of pos/neg examples
  - $\#(W=w_i, y)$: number of times word w_i occurs in examples of class y
  - l_+ / l_-: total number of words in pos/neg examples
  - $|V|$: size of vocabulary
- **Estimating P(Y)**

$$\hat{P}(Y=1)=\frac{n_+}{n} \qquad \hat{P}(Y=-1)=\frac{n_-}{n}$$

- **Estimating P(X|Y)**
  - Smoothing with Laplace estimate

$$\hat{P}(W=w_i|Y=y)=\frac{\#(W=w_i,y)+1}{l_y+|V|}$$

text	CS?
$\vec{x}_1=(The, art, of, Programming)$	+1
$\vec{x}_2=(Introduction, to, Calculus)$	-1
$\vec{x}_3=(Introduction, to, Complexity, Theory)$	+1
$\vec{x}_4=(Introduction, to, Programming)$	??

Test Collections

- **Reuters-21578**
  - Reuters newswire articles classified by topic
  - 90 categories (multi-label)
  - 9603 training documents / 3299 test documents (ModApte)
  - ~27,000 features
- **WebKB Collection**
  - WWW pages classified by function (e.g. personal HP, project HP)
  - 4 categories (multi-class)
  - 4183 training documents / 226 test documents
  - ~38,000 features
- **Ohsumed MeSH**
  - Medical abstracts classified by subject heading
  - 20 categories from "disease" subtree (multi-label)
  - 10,000 training documents/ 10,000 test documents
  - ~38,000 features

Example: Reuters Article (Multi-Label)

Categories: COFFEE, CRUDE

KENYAN ECONOMY FACES PROBLEMS, PRESIDENT SAYS

The Kenyan economy is heading for difficult times after a boom last year, and the country must tighten its belt to prevent the balance of payments swinging too far into deficit, President Daniel Arap Moi said.

In a speech at the state opening of parliament, Moi said high coffee prices and cheap oil in 1986 led to economic growth of five pct, compared with 4.1 pct in 1985. The same factors produced a two billion shilling balance of payments surplus and inflation fell to 5.6 pct from 10.7 pct in 1985, he added.

"But both these factors are no longer in our favour ... As a result, we cannot expect an increase in foreign exchange reserves during the year," he said.

…

Example: Ohsumed Abstract

Categories: Animal, Blood_Proteins/Metabolism, DNA/Drug_Effects, Mycotoxins/Toxicity, …

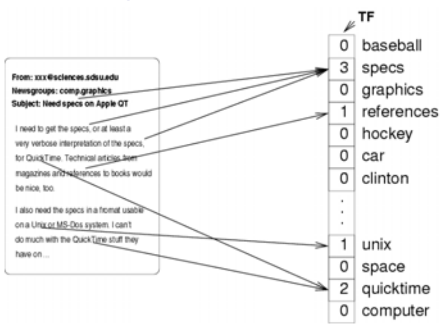
How aspartame prevents the toxicity of ochratoxin A.

Creppy EE, Baudrimont I, Anne-Marie

Toxicology Department, University of Bordeaux, France

The ubiquitous mycotoxin ochratoxin A (OTA) is found as a frequent contaminant of a large variety of food and feed and beverage such as beer, coffee and win. It is produced as a secondary metabolite of moulds from Aspergillus and Penicillium genera. Ochratoxin A has been shown experimentally to inhibit protein synthesis by competition with phenylalanine its structural analogue and also to enhance oxygen reactive radicals production. The combination of these basic mechanisms with the unusual long plasma half-life time (35 days in non-human primates and in humans), the metabolisation of OTA into still active derivatives and glutathione conjugate both potentially reactive with cellular macromolecules including DNA could explain the multiple toxic effects, cytotoxicity, teratogenicity, genotoxicity, mutagenicity and carcinogenicity. A relation was first recognised between exposure to OTA in the Balkan geographical

Representing Text as Attribute Vectors



=> Ignore ordering of words

Multi-Class / Multi-Label

- **Cannot learn multi-label rules directly**
  - Most classifiers assume that each document is in exactly one class
  - Many classifiers can only learn binary classification rules
- **Most common solution: Multi-Label**
  - Learn one binary classifier for each label
  - Attach all labels, for which some classifier says positive
- **Most common solution: Multi-Class**
  - Learn one binary classifier for each label
  - Put example into the class with the highest probability (or some approximation thereof)

Performance Measures

- **Precision/Recall Break-Even Point**
  - Intersection of PR-curve with the identity line
- **Macro-averaging**
  - First compute the measure, then compute average
  - Results in average over tasks
- **Micro-averaging**
  - First average the elements of the contingency table, then compute the measure
  - Results in average over each individual classification decision

Experimental Results

Reuters Newswir e	**WebKB Collection**	**Ohsumed MeSH**
• 90 cate gories	• 4 cate gories	• 20 cate gories
• 9603 training doc.	• 4183 training doc.	• 10000 training doc.
• 3299 test doc.	• 226 test doc.	• 10000 test doc.
• ~27000 features	• ~38000 features	• ~38000 features

microaveraged precision/recall breakeven-point [0..100]	**Reuters**	**WebKB**	**Ohsumed**
Naive Bayes	72.3	82.0	62.4
Rocchio Algorithm (LDA)	79.9	74.1	61.5
C4.5 Decision Tree	79.4	79.1	56.7
k-Nearest Neighbors	82.6	80.5	63.4
SVM	**87.5**	**90.3**	**71.6**

Comparison of Methods for Text Classification

	Naïve Bayes	Rocchio (LDA)	TDIDT C4.5	k-NN	SVM
Simplicity (conceptual)	++	++	-	++	-
Efficiency at training	+	+	--	++	-
Efficiency at prediction	++	++	+	--	++
Handling many classes	+	+	-	++	-
Theoretical validity	-	-	-	o	+
Prediction accuracy	-	o	-	+	++
Stability and robustness	-	-	--	+	++