Generative Models

CS4780/5780 – Machine Learning
Fall 2011

Thorsten Joachims
Cornell University

Reading:
Mitchell, Chapter 6.9 - 6.10
Duda, Hart & Stork, Pages 20-39 (see CoursePack at Campus Store)
• Bayes decision rule
• Bayes theorem
• Generative vs. discriminative learning
• Two generative learning algorithms
  – naive Bayes
  – linear discriminant analysis
• Estimating models from training data
  – maximum likelihood
  – maximum a posteriori (MAP)
• Assumption:
  – learning task P(X,Y) is known

• Question:
  – Given instance x, how should it be classified to minimize prediction error?

• Bayes Decision Rule:

$$h_{bayes}(\bar{x}) = \arg\max_{y \in Y} [P(Y = y | X = \bar{x})]$$
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 matches $P(X, Y)$ on training data
• Argument: if match close enough, we can use Bayes’ Decision rule
• Examples: naive Bayes, HMM
• It is possible to “switch” conditioning according to the following rule
• Given any two random variables $X$ and $Y$, it holds that

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

• Note that

$$P(X = x) = \sum_{y \in Y} P(X = x \mid Y = y)P(Y = y)$$
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
  \[ P(Y = +1) \quad P(Y = -1) \]

- Classification rule:
  \[ h_{naive}(\vec{x}) = \arg\max_{y \in \{+1, -1\}} \left\{ P(Y = y) \prod_{i=1}^{N} P(X_i = x_i | Y = y) \right\} \]

<table>
<thead>
<tr>
<th></th>
<th>fever (3)</th>
<th>cough (2)</th>
<th>pukes (2)</th>
<th>flu?</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td>high</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>low</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>-1</td>
</tr>
<tr>
<td>low</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>high</td>
<td>no</td>
<td>yes</td>
<td>???</td>
<td>????</td>
</tr>
</tbody>
</table>
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

- **Estimating $P(Y)$**
  - Fraction of positive / negative examples in training data
  \[
  \hat{P}(Y = 1) = \frac{n_+}{n}, \quad \hat{P}(Y = -1) = \frac{n_-}{n}
  \]

- **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
  \[
  \hat{P}(X_i = x_i|Y = y) = \frac{\#(X_i = x_i, y) + 1}{n_y + |X_i|}
  \]

<table>
<thead>
<tr>
<th></th>
<th>fever</th>
<th>cough</th>
<th>pukes</th>
<th>flu?</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>high</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>-1</td>
</tr>
<tr>
<td>low</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>high</td>
<td>no</td>
<td>yes</td>
<td>???</td>
<td>???</td>
</tr>
</tbody>
</table>
• 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

\[ P(Y = +1) \quad P(Y = -1) \]

• Classification rule

\[ h_{LDA}(\vec{x}) = \arg\max_{y \in \{+1, -1\}} \left\{ P(Y = y) e^{-\frac{1}{2}(\vec{x} - \vec{\mu}_y)^2} \right\} \]

\[ = \arg\max_{y \in \{+1, -1\}} \left\{ \log(P(Y = y)) - \frac{1}{2}(\vec{x} - \vec{\mu}_y)^2 \right\} \]

• Often called “Rocchio Algorithm” in Information Retrieval
Estimating the Parameters of LDA

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

- **Estimating P(Y)**
  
  - Fraction of pos / neg examples in training data
  
  \[
  \hat{P}(Y = 1) = \frac{n_+}{n} \quad \hat{P}(Y = -1) = \frac{n_-}{n}
  \]

- **Estimating class means**
  
  \[
  \hat{\mu}_+ = \frac{1}{n_+} \sum_{i : y_i = +1} \vec{x}_i \quad \hat{\mu}_- = \frac{1}{n_-} \sum_{i : y_i = -1} \vec{x}_i
  \]
Naïve Bayes Classifier (Multinominal)

- **Application: Text classification**

<table>
<thead>
<tr>
<th>text</th>
<th>CS?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\vec{x}_1 = (The, art, of, Programming))</td>
<td>+1</td>
</tr>
<tr>
<td>(\vec{x}_2 = (Introduction, to, Calculus))</td>
<td>-1</td>
</tr>
<tr>
<td>(\vec{x}_3 = (Introduction, to, Complexity, Theory))</td>
<td>+1</td>
</tr>
<tr>
<td>(\vec{x}_4 = (Introduction, to, Programming))</td>
<td>??</td>
</tr>
</tbody>
</table>

- **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**

  \[
h_{text}(\vec{x}) = \arg\max_{y \in \{+1, -1\}} \left\{ P(Y = y) \prod_{i=1}^{l} P(W = w_i|Y = y) \right\}
  \]
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 \( P(Y) \)
  \[ \hat{P}(Y = 1) = \frac{n_+}{n}, \quad \hat{P}(Y = -1) = \frac{n_-}{n} \]

- Estimating \( P(X|Y) \)
  - Smoothing with Laplace estimate
  \[ \hat{P}(W = w_i|Y = y) = \frac{#(W = w_i, y) + 1}{l_y + |V|} \]

<table>
<thead>
<tr>
<th>text</th>
<th>CS?</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 = \text{(The, art, of, Programming)} )</td>
<td>+1</td>
</tr>
<tr>
<td>( x_2 = \text{(Introduction, to, Calculus)} )</td>
<td>-1</td>
</tr>
<tr>
<td>( x_3 = \text{(Introduction, to, Complexity, Theory)} )</td>
<td>+1</td>
</tr>
<tr>
<td>( x_4 = \text{(Introduction, to, Programming)} )</td>
<td>??</td>
</tr>
</tbody>
</table>
- **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
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.

...
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 wine. 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 metabolism 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 Feature Vectors

Vector Space Representation of text

- Each word is a feature
- Feature value is either
  - Term Frequency
  - TFIDF  (*→ Wikipedia*)
- Normalize vectors to unit length

⇒ Ignore ordering of words.

Google announced today that they acquired Apple for the amount equal to the gross national product of Switzerland. Google officials stated that they first wanted to buy Switzerland, but eventually were turned off by the mountains and the snowy winters…
• 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 class
  – 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
### Experiment Results

<table>
<thead>
<tr>
<th></th>
<th>Reuters Newswire</th>
<th>WebKB Collection</th>
<th>Ohsumed MeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>– 90 categories</td>
<td>– 4 categories</td>
<td>– 20 categories</td>
</tr>
<tr>
<td></td>
<td>– 9603 training</td>
<td>– 4183 training</td>
<td>– 10000 training</td>
</tr>
<tr>
<td></td>
<td>docs</td>
<td>docs</td>
<td>docs</td>
</tr>
<tr>
<td></td>
<td>– 3299 test docs</td>
<td>– 226 test docs</td>
<td>– 10000 test docs</td>
</tr>
<tr>
<td></td>
<td>– ~27000 features</td>
<td>– ~38000 features</td>
<td>– ~38000 features</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Reuters</th>
<th>WebKB</th>
<th>Ohsumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes (multinomial)</td>
<td>72.3</td>
<td>82.0</td>
<td>62.4</td>
</tr>
<tr>
<td>Rocchio Algorithm (LDA)</td>
<td>79.9</td>
<td>74.1</td>
<td>61.5</td>
</tr>
<tr>
<td>C4.5 Decision Tree</td>
<td>79.4</td>
<td>79.1</td>
<td>56.7</td>
</tr>
<tr>
<td>k-Nearest Neighbors</td>
<td>82.6</td>
<td>80.5</td>
<td>63.4</td>
</tr>
<tr>
<td>SVM (linear)</td>
<td>87.5</td>
<td>90.3</td>
<td>71.6</td>
</tr>
</tbody>
</table>

**Microaveraged precision/recall breakeven point [0..100]**
<table>
<thead>
<tr>
<th></th>
<th>Naïve Bayes</th>
<th>Rocchio (LDA)</th>
<th>TDIDT C4.5</th>
<th>k-NN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplicity (conceptual)</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>Efficiency at training</td>
<td>+</td>
<td>+</td>
<td>--</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>Efficiency at prediction</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>--</td>
<td>++</td>
</tr>
<tr>
<td>Handling many classes</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>Theoretical validity</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>Prediction accuracy</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>++</td>
</tr>
<tr>
<td>Stability and robustness</td>
<td>-</td>
<td>-</td>
<td>--</td>
<td>+</td>
<td>++</td>
</tr>
</tbody>
</table>