Generative Models for Classification

CS4780/5780 – Machine Learning
Fall 2013

Thorsten Joachims
Cornell University

Reading:
Mitchell, Chapter 6.9 - 6.10
Duda, Hart & Stork, Pages 20-39

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

Linear Discriminant Analysis

- Spherical Gaussian model with unit variance for each class

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

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

- Prior probabilities

$$P(Y=+1), P(Y=-1)$$

- Classification rule

$$h_{LDA}(\vec{x}) = \underset{y\in\{+1,-1\}}{\operatorname{argmax}} \left\{P(Y=y) exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}_y\right)^2\right)\right\}$$

$$\underset{y\in\{+1,-1\}}{\operatorname{argmax}} \left\{\log(P(Y=y)) - \frac{1}{2}\left(\vec{x}-\vec{\mu}_y\right)^2\right\}$$

- Often called "Rocchio Algorithm" in Information Retrieval

Estimating the Parameters of LDA

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

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

- 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 ($x = (w_1, \ldots, w_l)$ sequence)

text	CS?
$x_1 = (The, art, of, Programming)$	+1
$x_2 = (Introduction, to, Calculus)$	-1
$x_3 = (Introduction, to, Complexity, Theory)$	+1
$x_4 = (Introduction, to, Programming)$	??

- Assumption

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

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

- Classification Rule

$$h_{naive}(x) = \underset{y\in\{+1,-1\}}{\operatorname{argmax}} \left\{P(Y=y)\prod_{i=1}^{l} P(W=w_i|Y=y)\right\}$$

Estimating the Parameters of Multinomial Naïve Bayes

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

text	CS?
$x_1 = (The, art, of, Programming)$	+1
$x_2 = (Introduction, to, Calculus)$	-1
$x_3 = (Introduction, to, Complexity, Theory)$	+1
$x_4 = (Introduction, to, Programming)$	??

- Estimating P(Y)

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

- Estimating P(X|Y) (smoothing with Laplace estimate):

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

Test Collections

- Reuters-21578
  - Reuters newswire articles classified by topic
  - 90 categories (multi-label)
  - 9603 training documents / 3299 test documents (ModApte)
  - ~27,000 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
- 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

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 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

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…

0	aardvark
1	announced
1	apple
…	…
0	get
0	goal
2	google
1	gross
…	…
0	zero
0	zilch
0	zztop

Multi-Class via "One-against-rest"

- Goal: $h: X \rightarrow Y$ with $Y = \{1,2,3,\dots,k\}$
- Problem:
  - Many classifiers can only learn binary classification rules
- Most common solution
  - Learn one binary classifier for each class
  - Put example into the class with the highest probability (or some approximation thereof)
- Example
  - Binarize: (x, 2) → (x, (-1,+1,-1,-1,-1,-1,-1,-1,-1,-1))
  - Training: one binary classifier $h_i(x)$ for each class i
  - Prediction on new x: $h(x) = \underset{i \in \{1..k\}}{\operatorname{argmax}} \{h_i(x)\}$
    - Assumes that classifier outputs a confidence score and not just +1 / -1

Multi-Label via "One-against-rest"

- Goal: $h: X \rightarrow Y$ with $Y = 2^{\{1,2,3,\dots,k\}}$
- Problem:
  - Many classifiers can only learn binary classification rules
- Most common solution
  - Learn one binary classifier for each label
  - Attach all labels, for which its binary classifier says positive
- Example
  - Binarize: (x, {2,7,9}) → (x, (-1,+1,-1,-1,-1,-1,+1,-1,+1,-1))
  - Training: one binary classifier $h_i(x)$ for each class i
  - Prediction on new x: $h(x) = (h_1(x), h_2(x), \dots, h_k(x))$

Contingency Table

	$h(x) = +1$	$h(x) = -1$
$y = +1$	a	b
$y = -1$	c	d

- Commonly used performance measures
  - ErrorRate = (b+c) / (a+b+c+d)
  - Accuracy = (a+d) / (a+b+c+d)
  - WeightedErrorRate = (w_b*b + w_c*c) / (a+b+c+d)
  - Precision = a / (a+c)
  - Recall = a / (a+b)
  - F_1Score = 2 * (Precision * Recall) / (Precision + Recall)

Performance Measures for Text Classification

- Precision/Recall Curve for Binary Classification
  - Sweep “threshold” from no positive classifications to only positive classifications
  - Plot precisions vs. recall at each step of the sweep
- 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

Also see http://www.scholarpedia.org/article/Text_categorization.

Experiment Results

Reuters Newswire
- 90 categories
- 9603 training docs
- 3299 test docs
- ~27000 features

WebKB Collection
- 4 categories
- 4183 training docs
- 226 test docs
- ~38000 features

Ohsumed MeSH
- 20 categories
- 10000 training docs
- 10000 test docs
- ~38000 features

Microaveraged precision/recall breakeven point [0..100]	Reuters	WebKB	Ohsumed
Naïve Bayes (multinomial)	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 (linear)	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	-	-	--	+	++