Generative Models for Classification

CS6780 – Advanced Machine Learning
Spring 2015
Thorsten Joachims
Cornell University

Reading:
Murphy 3.5, 4.1, 4.2, 8.6.1

Generative vs. Conditional vs. ERM

• Empirical Risk Minimization
  – Find $h = \arg\min_{h} \mathbb{E}_{s} \mathbb{L}(h)$ s.t. overfitting control
  – Pro: directly estimate decision rule
  – Con: committed to loss, $X$, $Y$

• Discriminative Conditional Model
  – Find $P(Y|X)$, then derive $h(x)$ via Bayes rule
  – Pro: not committed to loss
  – Con: committed to $X$, $Y$; conditional distributions more complex than decision rule

• Generative Model
  – Find $P(X,Y)$, then derive $h(x)$ via Bayes rule
  – Pro: not committed to loss function, $X$, and $Y$; often computationally easy
  – Con: Model dependencies in $X$

Bayes Decision Rule

• Assumption:
  – learning task $P(X,Y)=P(Y|X)P(X)$ is known

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

• Bayes Decision Rule:
  $h_{\text{bayes}}(x) = \arg\max_{y} [P(Y=y|X=x)P(X=x)]$

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

$$P(X=x|Y=y) = \prod_{i=1}^{n} P(X_i=x_i|Y=y)$$

• Prior probabilities

$$P(Y=y)$$

• Classification rule:

$h_{\text{naïve}}(x) = \arg\max_{y \in \{+1,-1\}} \left\{ P(Y=y) \prod_{i=1}^{n} P(X_i=x_i|Y=y) \right\}$

Estimating the Parameters of NB

• Count frequencies in training data
  – $n_y/n$: number of pos/neg examples
  – $|X_i|$: number of times pos/neg examples $X_i$ 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
  $\hat{P}(Y=+1) = \frac{n_y}{n}$
  $\hat{P}(Y=-1) = \frac{n_{-1}}{n}$

• Estimating $P(X|Y)$
  – Maximum Likelihood Estimate
  $\hat{P}(X_i=x_i|Y=y) = \frac{\#(X_i=x_i,Y=y)}{n_y}$
  – Smoothing with Laplace estimate
  $\hat{P}(X_i=x_i|Y=y) = \frac{\#(X_i=x_i,Y=y)+1}{n_y+|X_i|}$
Linear Discriminant Analysis

- Spherical Gaussian model with unit variance for each class
  \[ P(X = \pm 1) = \exp \left( -\frac{1}{2}(x - \mu_{\pm})^2 \right) \]
- Prior probabilities
  \[ P(Y = 1), P(Y = -1) \]
- Classification rule
  \[ h_LDA(x) = \arg\max_{Y \in \{\pm 1\}} P(Y = y) \exp \left( -\frac{1}{2}(x - \mu_y)^2 \right) \]

Often called "Rocchio Algorithm" in Information Retrieval

Estimating the Parameters of LDA

- Count frequencies in training data
  \[ (\chi_1, y_1), \ldots, (\chi_n, y_n) \sim P(X, Y) \]: training data
- \( n \): number of training examples
- \( n_+, n_- \): number of positive/negative training examples
- Estimating \( P(Y) \)
  \[ P(Y = 1) = \frac{n_+}{n}, \quad P(Y = -1) = \frac{n_-}{n} \]
- Estimating class means
  \[ \mu_+ = \frac{1}{n_+} \sum_{i | y_i = 1} x_i, \quad \mu_- = \frac{1}{n_-} \sum_{i | y_i = -1} x_i \]

Naïve Bayes Classifier (Multinomial)

- Application: Text classification (\( x = (w_1, \ldots, w_l) \) sequence)
  \[
  P(W = w | Y = y) = \frac{\#(W = w, Y = y) + 1}{l_y + |V|}
  \]
- 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
  \[ h_{naive}(x) = \arg\max_{Y \in \{\pm 1\}} P(Y = y) \prod_{i=1}^l P(W = w_i | Y = y) \]

Estimating the Parameters of Multinomial Naïve Bayes

- Count frequencies in training data
  \[ (\chi_1, y_1), \ldots, (\chi_n, y_n) \sim P(X, Y) \]: training data
- \( n \): number of training examples
- \( n_+, n_- \): number of positive/negative training examples
- Estimating \( P(Y) \)
  \[ P(Y = 1) = \frac{n_+}{n}, \quad P(Y = -1) = \frac{n_-}{n} \]
- Estimating \( P(W | Y) \) (smoothing with Laplace estimate):
  \[ \hat{P}(W = w | Y = y) = \frac{\#(W = w, Y = y) + 1}{l_y + |V|} \]