Discriminative vs. Generative Learning

CS4780/5780 – Machine Learning
Fall 2013

Thorsten Joachims
Cornell University

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

Discriminative Learning

- Modeling Step:
  - Select classification rules *H* to consider (hypothesis space)
- Training Principle:
  - Given training sample $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$
  - Find *h* from *H* with lowest training error → Empirical Risk Minimization
  - Argument: low training error leads to low prediction error, if overfitting is controlled.
- Examples: SVM, decision trees, Perceptron

Discriminative Training of Linear Rules

$$\min_{w,b} R(w) + C \sum_{i=1}^{n} \Delta(w * x_i + b, y_i)$$

Regularizer: $R(w)$; Regularization Parameter: C; Training Loss / Empirical Risk / Training error: $\sum_{i=1}^{n} \Delta(w * x_i + b, y_i)$

- Soft-Margin SVM
  - $R(w) = \frac{1}{2} w * w$
  - $\Delta(\bar{y}, y_i) = \max(0, 1 - y_i \bar{y})$
- Perceptron
  - $R(w) = 0$
  - $\Delta(\bar{y}, y_i) = \max(0, -y_i \bar{y})$
- Linear Regression
  - $R(w) = 0$
  - $\Delta(\bar{y}, y_i) = (y_i - \bar{y})^2$
- Ridge Regression
  - $R(w) = \frac{1}{2} w * w$
  - $\Delta(\bar{y}, y_i) = (y_i - \bar{y})^2$
- Lasso
  - $R(w) = \frac{1}{2} \sum |w_i|$
  - $\Delta(\bar{y}, y_i) = (y_i - \bar{y})^2$
- Regularized Logistic Regression / Conditional Random Field
  - $R(w) = \frac{1}{2} w * w$
  - $\Delta(\bar{y}, y_i) = \log(1 + e^{-y_i \bar{y}})$

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:

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

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

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

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

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

fever (h,l,n)	cough (y,n)	pukes (y,n)	flu?
high	yes	no	1
high	no	yes	1
low	yes	no	-1
low	yes	yes	1
high	no	yes	???

- Prior probabilities

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

- Classification rule:

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

Estimating the Parameters of NB

fever (h,l,n)	cough (y,n)	pukes (y,n)	flu?
high	yes	no	1
high	no	yes	1
low	yes	no	-1
low	yes	yes	1
high	no	yes	???

- 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

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

- 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

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