CMSC35000-1 Introduction to Artificial Intelligence **Winter 2005**

Lecture 5: Wednesday January 19

Lecturer: Partha Niyogi *Scribe: Mike Rainey*

0.1 Learning Algorithms

1. $sign(w \bullet x)$
2. $sign(\sum_{i=0}^{n} \alpha_i \sigma(w_i \bullet x))$
   Non-parametric models: $\sum_{i=1}^{n} \alpha_i f_i(x)$ where $f_i \in H$
   Given $(x_i, y_i) \ldots (x_n, y_n)$, find $\min_{f \in H_n}(\sum_{i=1}^{n}(y_i = f(x_i))^2)$ where $H_n = \sum_{i=1}^{n} \alpha_i f_i = \sigma(w \bullet x)$ are linear combinations.
3. Kernel Based Methods

Example 0.1 *Support Vector Machines, Least Squares Regularizaion*

Definition 0.2 *A Kernel K is defined as* $K : (x \times x) \to \mathbb{R}$.

Definition 0.3 *K is (a)* ***symmetric*** *if* $\forall_{x,y} K(x,y) = K(y,x)$ *and (b)* ***positive semidefinite*** *if* $\forall_{z_1,\ldots,z_n \in X} K_{i,j} = K(z_i, z_j)$.

Definition 0.4 *A* ***positive semidefinite matrix*** *is a Hermitian matrix all of whose eigenvalues are nonnegative.*

Example 0.5 *Examples of kernels:*

(a) $K(x,y) = e^{-\dfrac{\|x-y\|}{\sigma^2}}$

   Exercise 0.6 *Check that it is positive-semidefinite.*

(b) $K(x,y) = x^T \bullet y$

(c) $K(x,y) = (x \bullet y)^d \exists_d$

$$H = \{K(x, \bullet)\} \text{where}\, K(x, \bullet) : x \to \mathbb{R} \tag{0.1}$$

$$H = \{K_x | x \in x\} \text{where} \sum_{i=1}^{n} \alpha_i K_{xi} \tag{0.2}$$

$$\min_{f\in H_n}\sum_{i=1}^{n}(y_i - f(x_i))^2 = \min_{f\in H_n}\sum_{i=1}^{n}(y_i - \sum_{i=1}^{n}\alpha_j K(x_j, x_i))^2 \tag{0.3}$$

Consider $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$.

Consider a vector $K\widehat{\alpha}$ with n numbers where i^th element is $f(x_i) = \sum_{i=1}^{n} K(x_i, y_i)$.

$$\begin{aligned} J(\alpha) &= \min_{\alpha} \|y - K\alpha\|^2 \\ &= \min_{\alpha}(y - K\alpha)^T(y - K\alpha) \\ &= y^Ty - 2\alpha^T K^T y + \alpha^T K^T K\alpha \end{aligned} \tag{0.4}$$

is minimized when $\frac{\delta}{\delta\alpha} = 0$.

Example 0.7 $-2K^Ty = 2K^TK\alpha = 0$

Definition 0.8 *If K is positive definite, K is **invertible**. So, $\alpha = K^{-1}y$, and $K\alpha = y$ is interpreted data.*

0.2 Another Algorithm

Find α to fit data as closely as possible:

$$\min_{\alpha} \|y - K\alpha\|^2 \tag{0.5}$$

$$\min_{\alpha} \|y - K\alpha\|^2 + \gamma\|K\alpha\|^2 \tag{0.6}$$

where $\min_\alpha \|y - K\alpha\|^2$ is the fit to data, and $\gamma\|K\alpha\|^2$ controls complexity.

By using this method, the error of training data goes to 0. This framework is the most successful today, and $H_n = \sum_{i=1}^{n} \alpha_i K(x_i, \bullet)$ is called **Reproducing Kernel Hilbert Space** (RKHS).

0.3 Decision Trees

The goal is to learn a function $f : x \to y$ where $y = \{-1, 1\}$. We are given a set $Q = \{\text{questions}\}$ of yes / no questions. Formally, each $q \in Q$ is $q : x \to y$. The data are labeled examples denoted as (x_i, y_i). For buidling a decision tree, we want a good q, one that divides all the data into two classes: $y_i = +1$ and $y_i = -1$.

0.3.1 Purity of the Dataset

Given $D = \{(x_i, y_i)i = 1, \ldots, n\}$, n_1 = number of data such that $y_i = +1$, $N - n_1$ = number of data such that $y_i = -1$.

Definition 0.9 *If* $n_1 = 1$ *or* $n_1 = 0$ *we have a* ***pure data set****,* $n_1 = 1/2$ *we have an impure data set.*

Given a g, D, measure purity:

$$D_1 = \{(x_i, y_i) | q(x_i) = +1\} \tag{0.7}$$

$$D_2 = \{(x_i, y_i) | q(x_i) = -1\} \tag{0.8}$$

Corollary 0.10 *The following hold:* $D_1 \bigcap D_2 = \emptyset$ *and* $D_1 \bigcup D_2 = D$.

Definition 0.11 *Then we have* $\mathbf{g}(\mathbf{D}, \mathbf{q}) = \frac{|D_1|}{|D|} I(D_1) + \frac{|D_2|}{|D|} I(D_2)$.

where I is the **impurity function**.

Example 0.12 *A possible impurity function:*

$$p(1-p) \tag{0.9}$$

Example 0.13 *Another possible impurity function, the* ***entropy*** *of* p*:*

$$H(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \tag{0.10}$$

Proposition 0.14 $\min_{q \in Q} g(D, q)$ *finds the best question.*

Example 0.15 *Common decision tree for real-valued data.*

$$x = \mathbb{R}^k \tag{0.11}$$

$$(x_i, y_i) \; where \, i = 1, \ldots, n \tag{0.12}$$

$$Q = \{look \; at \; a \; coordinate \; and \; threshhold\} \tag{0.13}$$

Pick $i \in \{1, \cdots, k\}$ *and* $t \in \mathbb{R}$*. Then* $q(x, i, t, +) = +1 \Leftrightarrow x(i) > t$ *and* $q(x, i, t, -) = -1 \Leftrightarrow x(i) > t$.

Exercise 0.16 *Convince yourself that an impure dataset always has a query that gives a nontrivial split.*