CS369M: Algorithms for Modern Massive Data Set Analysis **Lecture 6 - 10/12/2009**

Reproducing Kernel Hilbert Spaces and Kernel-based Learning Methods (1 of 2)

Lecturer: Michael Mahoney *Scribes: David Fong and Ya Xu*

**Unedited Notes*

1 Kernel-based Learning Algorithms

Kernel-based Learning Algortihms are used in data analysis and machine learning. There are several types of learning mechanism:

- Unsupervised Learning - No teacher/ labels
- Supervised Learning - Teachers/ labels
- Semi-supervised Learning - The labels might be expensive and only some data point has labels.
- Online Learning - Time Series Data

2 Kernel-based Method

- A way to model a much larger class of data using a vector space model.
- A lot more descriptive flexibility without much additional computational cost.

The kernel method involves a mapping into a high (possibly infinite) dimensional space.

$$\phi(X) : X \rightarrow F$$

Given a set of vector $x_i \in \mathbf{R}$, we define the Gram matrix G:

$$G_{ij} = x_i^T x_j$$

which is a symmetric matrix of inner products.

Definition Given a matrix G, we say G is positive semi-definite if for all vectors x, we have $x^T G x \geq 0$.

We can also generalize the concept of positive number to a partial ordering on matrices. To compare two matrices A, B, we can check if $A - B$ is positive semi-definite.

In $\mathbf{R}^n$, any Gram matrix is positive semi-definite. Also, any positive semi-definite matrix is a Gram matrix for some set of vectors.

Note The set of vectors that generates a certain Gram matrix is not unique.

3 Supervised Learning - Classification

There are different ways to formalize this. One way is to say the data are $(x_i, y_i)_{i=1}^n \in R^N \times Y$, where $Y = \{-1, 1\}$. Then the goal is to find a function $f : \mathbf{R}^N \to Y$ such that if given a new example, it will classify it correctly. For example, we can say

$$\begin{cases} f(x) > 0 & \to \quad \text{assign 1} \\ f(x) < 0 & \to \quad \text{assign -1} \end{cases}$$

Question: What if the data is more representable as a graph?

3.1 Risk Minimization

Given some training data $(x_i, y_i)_{i=1}^n$ and also test data drawn from the same distribution $P(x, y)$, our goal is to find the best function f from what we already know:

- $(x_i, y_i)_{i=1}^n$
- a function class I to optimize over

We want to minimize the risk/error defined by

$$R[f] = \in L(f(x), y) dP(x, y)$$

where L denotes some loss function.

Our goal is to minimize $R[f]$ subject to to bias/variance trade-off while having flexibility generally.

3.2 Empirical Risk Minimization (ERM)

The empirical risk is defined on the test data set:

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i)$$

and we hope that if $n \to \infty$, we would have $R_{emp}[f] \to R[f]$.

3.3 Structural Risk Minimization

The idea is to restrict ourselves to some nice function class and do ERM with the following procedures:

1. Construct a nested family of function class

$$F_1 \subset F_2 \subset \cdots \subset F_k$$

2. Let $f_1, \cdots, f_k$ be the ERM solutions in F_k

3. Choose (k^*, F_{k^*}, f_{k^*}) such that upper found on generalization error is minimized.

Theorem Let h be the VC dimension of I. Then $\forall \delta > 0$, $f \in I$

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h(\ln(\frac{2n}{h} + 1)) - \ln(\delta/4)}{n}}$$

with probability $(1 - \delta)$, $\forall n > h$

Note The above bound represents a bias/variance trade-off. It doesn't not depend on $P(x, y)$. The main reason to use this bound is that we cannot compute the LHS, but given any h, we can compute the RHS.

3.4 VC dimension

Definition A *dichotomy* of set S is a partition of S into two disjoint pieces.

Definition A set of points S is *shattered* by a hypothesis space $\mathcal{H}$ if for all dichotomy of S, $\exists$ a hypothesis $h \in \mathcal{H}$ consistent with the dichotomy.

Definition The *VC dimension* of $\mathcal{H}$ over given set of points S is the size of the largest subset of S shattered by $\mathcal{H}$.

The point is that the complexity of a classifier does not depend on the size of $\mathcal{H}$, but on how it performs on S.

Note To show that the VC dimension of $\mathcal{H}$ is $\geq d$. View it as a game:

(1) I choose d points.
(2) The adversary chooses labels from $\{-1, 1\}$.
(3) I produce a hypothesis $h \in \mathcal{H}$.

The VC dimension is the maximum of such d.

Note The VC dimension is powerful to bound certain things, but

(1) it can be hard to work with.
(2) it is suboptimal bound.
(3) it is a distribution-independent bound.

3.5 Hyperplane

Definition *Hyperplane* is a set of $\mathcal{H}$ in $\mathbb{R}^n$ that is "nice" and has the following form: $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$. The decision boundary of a hyperplane is sign $(\langle \mathbf{w}, \mathbf{x} \rangle + b)$.

Claim In $\mathbb{R}^2$, I can find 3 points such that I can shatter with a hyperplane, but I can not find 4. The general result is that given m points in $\mathbb{R}^n$, they can be shattered by oriented hyperplane if and only if the points we have are linearly independent.

Claim The VC dimension of oriented hyperplane in $\mathbb{R}^n$ is $n + 1$.

Definition We say the data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are *linearly separable* if $\exists\{\mathbf{w}, b\}$ such that $\langle \mathbf{w}, \mathbf{x}\rangle + b = 0$ separates the data, i.e.

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq 1 \quad \text{if} \quad y_i = 1$$
$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{if} \quad y_i = -1$$

Definition Let d_+ (d_-) be the shortest distance from the separating hyperplane to a data point with $+$ ($-$) label. The *margin* of the separating hyperplane w.r.t. the data is $d_+ + d_-$. If $\mathbf{w}$ is the weight vector, then $d_+ = d_- = 1/||\mathbf{w}||$, so the margin $\gamma = 2/||\mathbf{w}||$.

Fact Let $\mathcal{H}$ be the set of linear classifiers, and $\mathcal{H}_\gamma$ be the set of linear classifiers with margin γ. Intuitively, $\mathcal{H}_\gamma$ is smaller than $\mathcal{H}$. Let R be the radius of the smallest inclosing ball of the data, then

$$VC(\mathcal{H}_\gamma) \leq R^2 \mathbf{w} \cdot \mathbf{w} + 1, \tag{1}$$

independent of dimension.

3.6 Support Vector Machines (SVM)

The fact in (1) suggests optimizing the margin (SVM). Given $\{\mathbf{x}_i, y_i\} \in \mathbb{R}^N \times \{-1, 1\}$, find a good classification hyperplane given by the following optimization problem:

$$\begin{aligned} \min{}_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||_2^2 \\ \text{Subject to} \quad & y_i\left(\langle \mathbf{w}, \mathbf{x}\rangle + b\right) \geq 1. \end{aligned} \tag{2}$$

We can write (2) as an unconstrained problem with the Lagrange multipliers. Define

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}||_2^2 - \sum_i \alpha_i \left(y_i(\langle \mathbf{w}, \mathbf{x}\rangle + b) - 1\right),$$

then (2) becomes

$$\min{}_{\mathbf{w},b}\max{}_{\alpha>0} \qquad L(\mathbf{w}, b, \alpha) \tag{3}$$

We can view (3) as a two player game

(1) If player A violates the constraint in (2), then player B can choose α such that the maximum goes to ∞.

(2) If player A satisfies the constraint in (2), then player B chooses $\alpha_i = 0$.