CS369M: Algorithms for Modern Massive Data Set Analysis **Lecture 7 - 10/14/2009**

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

Lecturer: Michael Mahoney *Scribes: Mark Wagner and Weidong Shao*

**Unedited Notes*

1 Support Vector Machine (continued)

Given $(\vec{x_i}, y_i) \in \mathbb{R}^n \times \{-1, +1\}$, we want to find a good classification hyperplane.

Assumptions: data are linearly separable, i.e., there exists a hyperplane such that $\begin{cases} \langle w, x \rangle + b \geq 1 & y = 1 \\ \langle w, x \rangle + b \leq -1 & y = -1 \end{cases}$,

or

$$y_i \left(\langle w, x_i \rangle + b \right) \geq 1$$

To decide on a hyperplane, we want to maximize margin.

1.1 Problem statement

(Primal)

$$\begin{aligned} \min_{w,b} \quad & \tfrac{1}{2} ||w||_2^2 \\ \text{s.t.} \quad & y_i \left(\langle w, x_i \rangle + b \right) \geq 1 \end{aligned}$$

The Lagrangian for the above problem

$$L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{n} \alpha_i y_i \left(\left(\langle w, x_i \rangle + b \right) - 1 \right)$$

We can view this as a 2-player game, i.e.

$$\min_{w,b} \max_{\alpha \geq 0} L(w, b, \alpha)$$

where player A chooses w and b while player B chooses an α

In particular, if A chooses an *infeasible* point (i.e. constraint violated), B can make the expression as large as possible. If A chooses a *feasible* point, then

$$\forall i \text{ such that } \left(y_i \langle w, x_i \rangle + b \right) - 1 > 0$$

We must have $\rightarrow \alpha_i = 0$

If feasible, then optimizing $\frac{1}{2} ||w||^2$

Alternatively: Consider the "Dual game"

$$\max_{\alpha_i > 0} \min_{w,b} L(w, b, \alpha) \leq \min_{w,b} \max_{\alpha} L(w, b, \alpha)$$

which is the weak duality. But for a wide class of objectives the *equality* holds (i.e., no duality gap–minimax).

LET

$$\begin{aligned}(w^*, b^*) &= \arg\min_{w,b}\max_{\alpha} L(w, b, \alpha) \quad (*\mathbf{A})\\ \alpha^* &= \arg\max_{\alpha}\min_{w,b} L(w, b, \alpha) \quad (*\mathbf{B})\end{aligned}$$

$$\begin{aligned}L(w^*, b^*, \alpha^*) &\leq \max_{\alpha} L(w^*, b^*, \alpha)\\ &= \min_{w,b}\max_{\alpha} L(w, b, \alpha)\end{aligned}$$

by $(*\mathbf{A})$

and then by minimax we can switch the order from above (not proved)

$$\begin{aligned}&= \max_{\alpha}\min_{w,b} L(w, b, \alpha)\\ &\leq \min_{w,b} L(w, b, \alpha^*)\end{aligned}$$

by $(*\mathbf{B})$

$$\leq L(w^*, b^*, \alpha^*)$$

Therefore all the above inequalities are equalities

Since $L(\cdot, \alpha)$ is convex with respect to w, b for a fixed α, we can find the optimum by the First Order Condition (fix α).

$$\begin{aligned}\frac{\partial L}{\partial b} = 0 \quad &\rightarrow \quad \sum_i \alpha_i y_i = 0\\ w_i \quad &= \quad \sum_i \alpha_i y_i x_i\\ \frac{\partial L}{\partial w} = 0 \quad &\rightarrow \quad w^* = \sum_i \alpha_i y_i x_i\end{aligned}$$

ie the optimal solution can be written in terms of the data points

$$\vec{w} = \sum_i \alpha_i y_i \vec{x_i}$$

1.2 Dual problem

(DUAL)

$$\begin{aligned}&\max \sum_i \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i y_i \alpha_j y_j \langle x_i, x_j \rangle\\ st \quad &\alpha_i \geq 0\\ &\sum_i \alpha_i y_i = 0\end{aligned}$$

where $\langle x_i, x_j \rangle$ is the kernel or Gram matrix $k(x_i, x_j)$

1.3 Generalizations

What if there are a few outliers? The data might not be separable, or might be separable but might have noise.

Problem statement (PRIMAL). Define slack variable ζ. Define regularization parameter η.

$$\begin{array}{rc} \min\limits_{w,b,\zeta} & \frac{1}{2}||w||_2^2 + \eta||\zeta|| \\ st & y_i(\langle w, x_i \rangle + b) \geq 1 - \zeta_i \\ & \zeta \geq 0 \end{array}$$

where ζ_i measures the degree of misclassification of the x_i. To remove constrains define Lagrangian over parameters α

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i y_i ((\langle w, x_i \rangle + b) - 1)$$

In the dual:

$$\begin{array}{c} \max \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i y_i \alpha_j y_j \langle x_i, x_j \rangle \\ \eta \geq \alpha_i \geq 0 \\ \sum_i \alpha_i y_i = 0 \end{array}$$

IDEA:

$k(x_i, x_j) \sim$ correlation matrix based on dot products

$$\Phi : x \to \Phi(x) \in \mathcal{F}$$

where $\mathcal{F}$ is the feature space, which may be high dimensional. Work with $(\Phi(x_i), y_i)$ in $\mathcal{F}$. But since $\mathcal{F}$ is higher dimensional it will be worse algorithmically and statistically.

Good news:

- hyperplanes are particularly nice - regularized heavily, vector space computations (eigenvalues, convex optimization)
- for certain $\mathcal{F}$ this works

Note: $k(x, y)$ may be very inexpensive to calculate, even though $\Phi(x)$ itself may be very expensive to calculate (e.g., in high dimensions). Examples,

$$\begin{array}{rcl} k(x, y) & \sim & \exp\left(-\beta ||x - y||^2\right) \\ & \sim & (\langle x, y \rangle + 1)^{\beta} \\ & \sim & \tanh(\alpha \langle x, y \rangle + \beta) \end{array}$$

k can also be defined "operationally" from data-defined graphs.

2 Reproducing Kernel Hilbert Spaces

2.1 Hilbert Space

DEFINE A vector space is a space with things (vectors) such that addition and scalar multiplication (over a field) are defined. e.g. $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{\mathbb{R}}-$functions from $\mathbb{R} \to \mathbb{R}$. $\mathbb{R}^X-$set of functions from $X \to \mathbb{R}$ (where X might be $\mathbb{R}^n$ or some subset of it).

DEFINE A BANACH SPACE is a vector space with a norm, i.e., elements have some "size" measure.

e.g, consider $\mathbb{R}^n$, fix a number $p \geq 1$, then $||x||_p = \left(\sum_i |x_i|^p\right)^{1/p}$ (the $p-$norm)

Define $\left\{L_p = f : \mathbb{R}^n \to \mathbb{R} : \int |f|^p \, dx < \infty\right\}$ with norm $||f||_{L_p} = \int |f|^p \, dx$.

A HILBERT SPACE is a Banach space that is complete with respect to the norm induced by the inner product. (Note: A metric space M is complete if every Cauchy sequence in M converges in M).

Examples:

$\mathbb{R}^n$ where $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$

l_2 where $\langle x, y \rangle_{l_2} = \sum_{i=1}^\infty x_i y_i$

L_2 where $\langle x, y \rangle_{L_2} = \int_{-\infty}^\infty x(t)\, y(t)\, dt$

Intuitively, L_2 is an infinite-dimensional version of $\mathbb{R}^n$ but

- it's too big to get tractable algorithms, too big for good generalization properties
- too many "weird" or "pathological" functions

So, consider a subset of it (RKHS):

2.2 RKHS

DEFINE for a compact subset of $\mathbb{R}^n$ and some Hilbert space H of functions from $X \to \mathbb{R}$. H is a REPRODUCING KERNEL HILBERT SPACE if $\exists$ some kernel $k : X \times X \to \mathbb{R}$ such that

1. k has the *reproducing property*: $\langle k(., x), f \rangle = f(x)$
2. k spans H, i.e. span $\{k(\bullet, x) : x \in X\} = H$

Technical point

Reisz Representer theorem.

If Φ is a bounded functional on H then $\exists$ unique $u \in H$ such that $\Phi(f) = \langle f, u \rangle_H \, \forall f \in H$

Define a function/operator k is positive-definite if $\forall$functions $\int f(x)\, k(x, x')\, f(x')\, dx dx' > 0$

The high level idea:

1. start with kernel k
2. define universe V of H ie a set of functions and define a dot product on $V \times V$
3. This dot product gives a norm, which makes a reproducing kernel hilbert space

Given a positive-definite kernel $k(x, x')$ AND $x_1, ..x_n$, define a Gram matrix K such that

$$K_{ij} = k(x_i, x_j)$$

Note Cauchy Schwarz holds ie $k(x_i, x_j)^2 \leq k(x_i, x_i) k(x_j, x_j)$

Define reproducing property

$$\Phi : x \to k(\cdot, x)$$

i.e., represent each x by its behavior with respect to every other point.

Construct a vector space by linear combinations of $k(\cdot, x)$

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

- vector space, reproducing kernel hilbert space

Define dot product:

$$g(\cdot) = \sum_{j} B_j k(\cdot, x_j)$$

Then $\langle f(\cdot), g(\cdot) \rangle = \sum_{ij} \alpha_i \beta_i k(\cdot, x_i) k(\cdot, x_j) = \sum_{ij} \alpha_i \beta_j k(x_i, x_j)$

claim: this is an inner product

$$\langle k(\cdot, x), f \rangle = \sum \alpha_i k(x_i, x)$$

, i.e. k is the "representer" of the evaluation (analog of the delta function).

In particular, one possible f could be the kernel $k(\cdot, x)$ in which case the dot product:

$$\langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$$

this is reproducing

2.3 Mercer Theorem

If k is a positive definite kernel, then $\exists$ continuous $\{\Phi_i\}_{i=1}^{\infty}$ $\{\lambda_i\}_{i=1}^{\infty}$ such that $k(x, x') = \sum_{i=1}^{n} \lambda_i \Phi(x) \Phi(x')^T$

will show

- that we can represent data as a finite set of points
- solutions to optimization problems can be written in terms of data points

Basis for any algorithm that depends on the data just in terms of dot products can be represented by $k(x, x')$

- construction of data dependent kernels (isomap, lle, laplacian eigenmaps)

3 References

1. Cortes and Vapnik, "Support-Vector Networks", *Machine Learning*, p 273-297, 1995
2. Scholkopf, Smola, and Muller, "Nonlinear component analysis as a kernel eigenvalue problem", 1998
3. Scholkopf, Herbrich, Smola, and Williamson, "A Generalized Representer Theorem"