CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 4**

Kernels and Constrained Optimization

Lecturer: Peter L. Bartlett *Scribe: Robert Gibbons*

1 Review of SVMs from last lecture

Last lecture we started to talk about SVMs. In review, the optimization problem involves finding a boundary such that every point is classified correctly and that maximizes the Euclidean distance γ from the boundary to the nearest point. That is:

$$\max_{\gamma,w}(\gamma) \text{ such that } \frac{y_i w^T x_i}{\|w\|} \geq \gamma$$

If we set $\|w\|\gamma = 1$, this optimization problem is equivalent to finding:

$$\min_w \frac{1}{2}\|w\|^2 \text{ such that } y_i w^T x_i \geq 1 \text{ for all } i = 1, 2, \ldots, n$$

We then come up with the Lagrangian:

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

which we want to minimize with respect to w and maximize with respect to $\alpha \geq 0$.

At the solution, we observed last lecture that $\frac{\partial}{\partial w} L(w, \alpha) = 0$, and $w^* = \sum \alpha_i y_i x_i$.

2 Dual optimization problem

Now let us go on to compute the dual α. Substituting into L, we have

$$L(w^*, \alpha) = \frac{1}{2}\|\sum_i \alpha_i y_i x_i\|^2 - \left(\sum_i \alpha_i y_i x_i^T\right)\left(\sum_j \alpha_j y_j x_j\right) + \sum_i \alpha_i = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

and we get the dual optimization problem, to find maximum

$$\max_{\alpha \in \mathbb{R}^n} \left(\sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j\right) \text{ such that } \alpha_i \geq 0$$

So as in the perceptron algorithm, it is only the inner products that determine the solution.

We can express the solution in terms of the dual variables α_i and write $f_n(x) = \text{sign}\langle w, x\rangle$ where so far we have been thinking of $\langle w, x\rangle = w^T x$, but we might have $f_n(x) = \text{sign}\left(\sum y_i \alpha_i \langle \Phi(x_i), \Phi(x)\rangle\right)$ where Φ might be some mapping to a higher dimensional space, and we can write this as $f_n(x) = \text{sign}\left(\sum \alpha_i y_i k(x_i, x)\right)$, so that α is the solution to:

$$\max_{\alpha \geq 0} \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

where k is a kernel function that corresponds to an inner product. Thus, everything can be done in terms of inner products.

3 Examples of kernels

Example: Polynomial kernel, for $\mathcal{X} \in \mathbb{R}^d$. For $u, v \in \mathbb{R}^d$, let $k_2(u, v) = (u^T v)^2$. Why is this an inner product? For $d = 2$,

$$(u^T v)^2 = \left(\sum_{i=1}^{d} u_i v_i\right)^2 = (u_1^2, \sqrt{2} u_1 u_2, u_2^2) \begin{pmatrix} v_1^2 \\ \sqrt{2} v_1 v_2 \\ v_2^2 \end{pmatrix}$$

That is, $k_2(u, v) = \Phi(u)^T \Phi(v)$ for

$$\Phi(u) = \begin{pmatrix} u_1^2 \\ \sqrt{2} u_1 u_2 \\ u_2^2 \end{pmatrix}$$

and you can do a similar thing with higher powers.

The space $\{x \mapsto w^T \Phi(x) : w \in \mathbb{R}^3\}$ is the set of linear combinations of monomials of degree 2. For degree up to 2, we can take the mapping $k_{\leq 2}(u, v) = (u^T v + 1)^2$. This is a simple example, in which we have gone up from $\mathbb{R}^2$ to $\mathbb{R}^3$.

One thing we wish to point out is that we do not have uniqueness in the inner product representation. We could also write $k_2(u, v) = \Psi(u)^T \Psi(v)$ where $\Psi(u) = (u_1^2, u_1 u_2, u_2 u_1, u_2^2)^T$. Here we are mapping to some manifold in $\mathbb{R}^4$. Note that $\{\Psi(u) : u \in \mathcal{X}\}$ is called the "feature space", which might not be unique.

Non-trivial example: Instead of mapping to a finite-dimensional space, we map to an infinite-dimensional space. Again, $\mathcal{X} = \mathbb{R}^2$. Let

$$k(u, v) = \exp\left(\frac{-\|u - v\|^2}{2\sigma^2}\right)$$

This is a Gaussian kernel with parameter σ. Here, u and v are similar in the case where they are close. We will see that we can express this as an inner product $k(u, v) = \langle \Phi(u), \Phi(v)\rangle$ with $\Phi : \mathcal{X} \mapsto \ell_2$, a map to an infinite-dimensional space.

For example, consider $\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x)\rangle = k(x, x) = 1$. Cauchy-Schwartz tells us that $0 \leq k(u, v) \leq 1$. Here, we are exploiting the fact that we can compute the inner product without explicit representation.

We are talking about kernels on $\mathbb{R}^d$, but we can imagine defining the inner products on other spaces, such as an inner product capturing a notion of similarity between graphs.

4 Constrained optimization

Now we want to look more carefully at ideas from optimization theory. This will be a review for those students who have taken an optimization course. Most of these ideas are from chapter 5 of *Convex Optimization* by Stephen Boyd and Lieven Vandenberghe.

Example of primal problem: find $\min_{x \in \mathbb{R}^d} f_0(x)$ such that $f_i(x) \leq 0$ for $i = 1, \ldots, m$. The solution is $f_0(x^*) = p^*$ where p^* is the solution to the primal problem.

In the SVM case, we have Lagrangian $L(x, \lambda) = f_0(x) + \sum \lambda_i f_i(x)$, where the λ_i terms are the Lagrange multipliers or "dual variables".

Define the Lagrange dual $g : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ where

$$g(\lambda) = \inf_x L(x, \lambda) = \inf_x f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

We have:

1. $g(\lambda)$ is concave, which means that $g(\alpha\lambda_1 + (1-\alpha)\lambda_2) \geq \alpha g(\lambda_1) + (1-\alpha) g(\lambda_2)$. This is true because the minimum over a set of linear functions is concave.
2. If $\lambda_i \geq 0$ for all i, and x is primal feasible (which means that all of the constraints are satisfied for x), then $g(\lambda) \leq f_0(x)$, because

   $$g(\lambda) = \inf_{x'} (f_0(x') + \sum \lambda_i f_i(x')) \leq f_0(x) + \sum \lambda_i f_i(x) \leq f_0(x)$$

   The last inequality follows because $\lambda_i \geq 0$ and $f_i(x) \leq 0$ for all i.

   In particular, for any $\lambda_i \geq 0$, $g(\lambda) \leq p^*$, so we get a lower bound on the optimal solution in the non-convex case.

To find the best lower bound, we can maximize the Lagrange dual:

$$\max_\lambda g(\lambda) \text{ such that } \lambda_i \geq 0 \ .$$

We will write the optimal value as $g(\lambda^*) = d^*$. This tells us that $d^* \leq p^*$, which is called "weak duality". The gap $p^* - d^*$ is called the "optimal duality gap".

The reason we like an optimization problem of this form is that, for a concave function, the maximizing set is convex. If a convex function $f : \mathbb{R}^n \to \mathbb{R}$ has a minimum $f(x^*)$ on a convex $S \subseteq \mathcal{X}$, then the set $\{x : f(x) = f(x^*)\}$ is convex. If f is strictly convex, then $\{x : f(x) = f(x^*)\} = x^*$ so we have a unique solution. The statement "f strictly convex" means in $\mathbb{R}^n$ that $f(\alpha x_1 + (1-\alpha)x_2) < \alpha f(x_1) + (1-\alpha) f(x_2)$ for $\alpha \in (0, 1)$.

5 Geometric interpretation of the dual

Define a set $G = \{(u, t) \in \mathbb{R}^{m+1} : \exists x \text{ such that } f_i(x) = u_i, f_0(x) = t\}$. Suppose $m = 1$. The primal feasible region is $\{(u, t) \in G : u = 0\}$ and for our primal optimization, the point that corresponds to our solution is

$p^* = \min_t$ such that (u, t) is in the primal feasible region.

So this is the primal, what does the dual look like? It is:

$$g(\lambda) = \inf_x f_0(x) + \sum_i \lambda_i f_i(x) = \inf_G \left\{ \begin{pmatrix} \lambda^T & 1 \end{pmatrix} \begin{pmatrix} u \\ t \end{pmatrix} : \begin{pmatrix} u \\ t \end{pmatrix} \in G \right\}$$

In the case of one constraint, the set:

$$\left\{ \begin{pmatrix} u \\ t \end{pmatrix} : \begin{pmatrix} \lambda^T & 1 \end{pmatrix} \begin{pmatrix} u \\ t \end{pmatrix} = c \right\}$$

is a line with slope $-\lambda$ and intercept c. In the higher-dimensional case, we have a plane.

Now we maximize over λ to find $\max_\lambda g(\lambda) = d^*$. In the case $p^* = d^*$, we call that "strong duality". For example, if G were convex, then $p^* = d^*$. A lot of algorithms for convex optimization look at the primal and dual problems simultaneously and stop when those two values get very close to one another. It iss important to understand when we have this strong duality. Convexity is not quite enough, we need convexity plus some constraint qualifications. The easiest one is "Slater's condition": the primal is strictly feasible, that is $\exists x : f_i(x) < 0$ for $i \in \{1, 2, \ldots, m\}$.