CS281B/Stat241B. Statistical Learning Theory. Lecture 19.

Peter Bartlett

Overview

- Optimal regret
  - Sequential Rademacher averages
- Kernel methods
  - Perceptron algorithm revisited
  - Inner products
  - Kernels
  - Reproducing kernel Hilbert spaces

Optimal Regret

We have:

- a set of actions $\mathcal{A}$,
- a set of loss functions $\mathcal{L}$.

At time t,

- Player chooses an action a_t from $\mathcal{A}$.
- Adversary chooses $\ell_t : \mathcal{A} \to \mathbb{R}$ from $\mathcal{L}$.
- Player incurs loss $\ell_t(a_t)$.

Regret is the value of the game:

$$V_n(\mathcal{A}, \mathcal{L}) = \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^{n} \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right).$$

Recall: Dual Game

Theorem: If $\mathcal{A}$ is compact and all ℓ_t are convex, continuous functions, then

$$V_n(\mathcal{A}, \mathcal{L}) = \sup_P \mathbf{E}\left(\sum_{t=1}^n \inf_{a_t \in \mathcal{A}} \mathbf{E}\left[\ell_t(a_t) | \ell_1, \ldots, \ell_{t-1}\right] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a)\right),$$

where the supremum is over joint distributions P over sequences $\ell_1, \ldots, \ell_n$ in $\mathcal{L}^n$.

Recall: Sequential Rademacher Averages

Theorem:

$$V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t \ell_t(a),$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher (uniform ± 1-valued) random variables.

- Rademacher averages in probabilistic setting:

$$\text{excess risk} \leq c \mathbf{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(Y_t, f(X_t)) \right|.$$

- Sequential Rademacher averages in adversarial setting:

$$V_n(\mathcal{A}, \mathcal{L}) \leq c \sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t \ell_t(a).$$

Sequential Rademacher Averages: Example

Consider step functions on $\mathbb{R}$:

$$\begin{aligned} f_a &: x \mapsto 1[x \geq a] \\ \ell_{x,y}(a) &= 1[f_a(x) \neq y] \\ \mathcal{L} &= \{a \mapsto 1[f_a(x) \neq y] : x \in \mathbb{R},\ y \in \{0,1\}\}. \end{aligned}$$

Fix a distribution on $\mathbb{R} \times \{\pm 1\}$, and consider the Rademacher averages,

$$\mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_{X_t,Y_t}(a).$$

Rademacher Averages: Example

For step functions on $\mathbb{R}$, Rademacher averages are:

$$
\begin{aligned}
&\mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_{X_t, Y_t}(a) \\
&= \mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t \ell_{X_t, 1}(a) \\
&\leq \sup_{x_t} \mathbf{E} \sup_{a \in \mathbb{R}} \sum_{t=1}^{n} \epsilon_t 1[x_t < a] \\
&= \mathbf{E} \max_{0 \leq i \leq n+1} \sum_{t=1}^{i} \epsilon_t \\
&= O(\sqrt{n}).
\end{aligned}
$$

Sequential Rademacher Averages: Example

Consider the sequential Rademacher averages:

$$\sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t \ell_t(a)$$

$$= \sup_{x_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{x_n} \mathbf{E}_{\epsilon_n} \sup_{a} \sum_{t=1}^{n} \epsilon_t 1[x_t < a].$$

- If $\epsilon_t = 1$, we'd like to choose a such that $x_t < a$.
- If $\epsilon_t = -1$, we'd like to choose a such that $x_t \geq a$.

Sequential Rademacher Averages: Example

We can choose $x_1 = 0$ and, for $t = 1, \ldots, n$,

$$x_t = \sum_{i=1}^{t-1} 2^{-i} \epsilon_i = x_{t-1} + 2^{-(t-1)} \epsilon_{t-1}.$$

Then if we set $a = x_n + 2^{-n} \epsilon_n$, we have

$$\epsilon_t 1[x_t < a] = \begin{cases} 1 & \text{if } \epsilon_t = 1, \\ 0 & \text{otherwise,} \end{cases}$$

which is maximal.

Sequential Rademacher Averages: Example

So the sequential Rademacher averages are

$$\sup_{\ell_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbf{E}_{\epsilon_n} \sup_a \sum_{t=1}^n \epsilon_t \ell_t(a) = \mathbf{E} \sum_{t=1}^n 1[\epsilon_t = 1] = \frac{n}{2}.$$

Compare with the Rademacher averages:

$$\mathbf{E} \sup_{a\in\mathbb{R}} \sum_{t=1}^n \epsilon_t \ell_a(Y_t, X_t) = O(\sqrt{n}).$$

Optimal Regret: Lower Bounds

- For the case of prediction with absolute loss:

$$\ell_t(a_t) = |y_t - a_t(x_t)|,$$

there are (almost) corresponding lower bounds:

$$\frac{c_1 R_n(\mathcal{A})}{\log^{3/2} n} \leq V_n \leq c_2 R_n(\mathcal{A}),$$

where

$$R_n(\mathcal{A}) = \sup_{x_1} \mathbf{E}_{\epsilon_1} \cdots \sup_{x_n} \mathbf{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^{n} \epsilon_t a(x_t).$$

Overview

- Optimal regret
  - Sequential Rademacher averages
- Kernel methods
  - Perceptron algorithm revisited
  - Inner products
  - Kernels
  - Reproducing kernel Hilbert spaces

Recall: Perceptron algorithm

Input: $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{\pm 1\}$

$\theta_0 = 0 \in \mathbb{R}^d$, $t = 0$

while some (x_i, y_i) is misclassified, i.e., $y_i \neq \mathrm{sign}(\theta_t^T x_i)$

    pick some misclassified (x_i, y_i)

    $\theta_{t+1} := \theta_t + y_i x_i$

    $t := t + 1$

Return θ_t.

Recall: Perceptron algorithm

Perceptron convergence theorem: Given *linearly separable data* ($y_i\theta^T x_i > 0$), the perceptron algorithm makes no more than $\frac{R^2}{\gamma^2}$ updates (R =radius, γ =margin).

Regret/mistake bound: For

$$\mathcal{A} = \{x \mapsto \mathrm{sign}(\theta^T x) : \theta \in \mathbb{R}^d\},$$

$$\mathcal{L}_t = \left\{a \mapsto 1[a(x_t) \neq y_t] : \{(x_s, y_s)\}_{s=1}^t \text{ radius } R\text{, margin } \gamma\right\},$$

the perceptron algorithm has regret no more than R^2/γ^2.

Risk bound: If $\theta' x\, y/\|\theta\| \geq \gamma$, then risk $\leq R^2/(n\gamma^2)$. (And this is optimal.)

Kernel methods

The perceptron algorithm (and its convergence proof) works in a more general *inner product space*:

- We can write θ_t in terms of the data:
  $\theta_t = \sum_i \alpha_i x_i$ with $\|\alpha\|_1 = \sum_i |\alpha_i| = t$.
- We can replace the inner product $\langle x, \theta \rangle = x^T \theta$ with an arbitrary inner product:
  predict: $\hat{y}_i = \text{sign}\left(\sum_j \alpha_j \langle x_j, x_i \rangle\right)$,
  update: if $\hat{y}_i \neq y_i$, set $\alpha_i^{(t+1)} := \alpha_i^{(t)} + y_i$.

Inner products: definition

An inner product on a vector space is:

Symmetric $\langle u, v\rangle = \langle v, u\rangle$.

Linear $\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$,
$\langle \alpha u, v\rangle = \alpha\langle u, v\rangle$.

Positive definite $\langle u, u\rangle \geq 0$,
$\langle u, u\rangle = 0 \Rightarrow u = 0$.

Inner products: examples

1. Dot product on $\mathbb{R}^d$: $\langle u, v\rangle = u'v$.
2. Arbitrary inner product on $\mathbb{R}^d$: $\langle u, v\rangle = u'Av$ for symmetric positive definite A.
   (The eigendecomposition of A shows that this is the regular dot product of a scaled—in d orthogonal directions—version of the u, v.)
3. Random variables, $\langle X, Y\rangle = \mathbf{E}(XY)$.
4. Continuous functions on $[a, b]$, $\langle f, g\rangle = \int_a^b f(x)g(x)\,dx$.
5. Symmetric matrices, $\langle A, B\rangle = \mathrm{tr}(AB)$.
6. Square summable sequences, $\langle u, v\rangle = \sum_{i=1}^{\infty} u_i v_i$, where $\|u\|^2 < \infty$.

Kernels

In these examples, we define the inner product on a particular vector space. But for the perceptron algorithm and analysis, all we needed was that there is an inner product on *some* vector space:

$$\hat{y} = \operatorname{sign}\left(\sum_j \alpha_j \langle \Phi(x_j), \Phi(x) \rangle\right),$$

$$\Phi : \mathcal{X} \mapsto \mathcal{V}.$$

We don't need to explicitly evaluate $\Phi(x)$, as long as we can evaluate the inner products.

Example: Polynomial kernels

$$
\begin{aligned}
k_2(u,v) = (u'v)^2 &= (u_1 v_1 + u_2 v_2)^2 \\
&= \begin{pmatrix} u_1^2 & \sqrt{2}u_1u_2 & u_2^2 \end{pmatrix} \begin{pmatrix} v_1^2 \\ \sqrt{2}v_1v_2 \\ v_2^2 \end{pmatrix} \\
&= \Phi_2(u)'\Phi_2(v).
\end{aligned}
$$

Here, $\Phi_2 : \mathbb{R}^2 \mapsto \mathbb{R}^3$.

Example: Polynomial kernels

- The function class $\{x \mapsto \theta' \Phi_2(x) : \theta \in \mathbb{R}^3\}$ gives all homogeneous degree 2 polynomials. Decision boundaries are solution sets for polynomial equations.
- Similarly, we can write $k_m(u, v) = (u'v)^m$, with a feature map $\Phi_m : \mathbb{R}^d \to \mathbb{R}^D$, and the function class $\{x \mapsto \theta' \Phi_m(x) : \theta \in \mathbb{R}^D\}$ gives all homogeneous degree m polynomials.
- The feature map $\Phi_m : \mathbb{R}^d \to \mathbb{R}^D$ has $D = \binom{d+m-1}{m}$ features, which grows exponentially with m. But for the perceptron algorithm, we only need to evaluate quantities involving $k(u, v) = \Phi_m(u)' \Phi_m(v)$, and we never need to explicitly compute the (huge) feature map.

What k correspond to inner product spaces?

Suppose we have a function $k : \mathcal{X}^2 \to \mathbb{R}$. Does it correspond to an inner product in *some* vector space?

i.e.: What properties should k have to ensure that there is some underlying inner product space $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ and feature map $\Phi : \mathcal{X} \to \mathcal{F}$ such that

$$k(u, v) = \langle \Phi(u), \Phi(v) \rangle?$$

What k correspond to inner product spaces?

Necessary conditions:

1. Because an inner product is symmetric, we must have **symmetry**: $k(u, v) = k(v, u)$.
2. Because an inner product is positive definite, we must have $k(u, u) \geq 0$.
   (But we might not have $k(u, u) = 0 \Rightarrow u = 0$.)
3. Cauchy-Schwarz implies $k(u, v)^2 \leq k(u, u)k(v, v)$.

What k correspond to inner product spaces?

In fact, 2 and 3 follow from k being positive semidefinite:

> **Definition:** $k : \mathcal{X}^2 \to \mathbb{R}$ is **positive semidefinite** if, for all n and all $x_1, \ldots, x_n \in \mathcal{X}$, the *Gram matrix* $K \in \mathbb{R}^{n\times n}$—defined by $K_{ij} = k(x_i, x_j)$—is positive semidefinite.

Notice that $k(u, v) = \langle \Phi(u), \Phi(v) \rangle$ is positive semidefinite:

$$
\begin{aligned}
v'Kv = \sum_{i,j} v_i v_j k(x_i, x_j) &= \sum_{i,j} v_i v_j \langle \Phi(x_i), \Phi(x_j) \rangle \\
&= \left\langle \sum_i v_i \Phi(x_i), \sum_j v_j \Phi(x_j) \right\rangle \geq 0.
\end{aligned}
$$

Also, $n = 1$ shows $k(u, u) \geq 0$.
And $n = 2$ shows $k(u, v)^2 \leq k(u, u)k(v, v)$.

What k correspond to inner product spaces?

These conditions are necessary *and sufficient*:

Definition: $k : \mathcal{X}^2 \to \mathbb{R}$ is a **kernel** if it is

1. Symmetric: $k(u, v) = k(v, u)$, and
2. Positive semidefinite: every Gram matrix $K_{ij} = k(x_i, x_j)$ is positive semidefinite.

Theorem: If k is a kernel, then there is an inner product space $\mathcal{F}$ and a feature map Φ such that $k(u, v) = \langle \Phi(u), \Phi(v) \rangle$.

Kernels and inner product spaces

Consider:

$$\Phi(x) = k(\cdot, x),$$

$$\mathcal{F} = \operatorname{span}\left\{\Phi(x) : x \in \mathcal{X}\right\},$$

$$\left\langle \sum_i \alpha_i \Phi(u_i), \sum_j \beta_j \Phi(v_j) \right\rangle = \sum_{i,j} \alpha_i \beta_j k(u_i, v_j).$$

Then it's easy to check: $\mathcal{F}$ is a linear space of functions, $\langle \cdot, \cdot \rangle$ is symmetric, linear, positive definite.