CS281B/Stat241B. Statistical Learning Theory. Lecture 3.

Peter Bartlett

1. Review: Linear threshold functions, perceptron algorithm.
2. Lower bounds (d/n) on minimax risk for linear threshold functions.
3. Upper and lower bounds ($R^2/n\gamma^2$) on minimax risk for perceptron algorithm.
4. Risk bounds, uniform convergence, concentration.

Review: Linear threshold functions on $\mathbb{R}^d$

$$F = \left\{ x \mapsto \mathrm{sign}(\theta' x) : \theta \in \mathbb{R}^d \right\}.$$

Empirical risk minimization:

Choose f from F to minimize the *empirical risk,*

$$\hat{R}(f) = \hat{\mathrm{E}}\ell(f(X), Y) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Review: 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 $\theta_t$.
```

Review: Perceptron algorithm

Theorem: For any $\theta \in \mathbb{R}^d$ such that for all i, $y_i \theta^T x_i > 0$ (linearly separable data), for any choices made at the update step, the perceptron algorithm terminates (with empirical risk zero) after no more than $\frac{R^2}{\gamma^2}$ updates, where

$$R = \max_i \|x_i\|, \qquad \text{(radius of data)}$$

$$\gamma = \min_i \frac{\theta^T x_i y_i}{\|\theta\|}. \qquad \text{(margin)}$$

A glimpse of kernel methods

Notes:

- 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: $\hat{y}_i = \operatorname{sign}\left(\sum_j \alpha_j \langle x_j, x_i \rangle\right)$,

  update: if $\hat{y}_i \neq y_i$, set $\alpha_i^{(t+1)} := \alpha_i^{(t)} + y_i$.

So the perceptron algorithm (and its convergence proof) works in a more general *inner product space*.

Minimax risk

Consider the minimax risk,

$$\min \max_{P} \mathrm{E} R(f_n),$$

where the max is over all P for which some $f \in F$ has zero risk, and the min is over all methods that use data to choose a prediction rule f_n (perhaps in F, perhaps not).

If $n \leq d$, then we should expect the minimax risk to be large. For instance, if $x_1, \ldots, x_n$ are linearly independent, then for any $y_1, \ldots, y_n$, we can find $\theta \in \mathbb{R}^d$ such that

$$\theta' \left[x_1 | x_2 | \cdots | x_n\right] = \left[y_1, y_2, \ldots, y_n\right],$$

and hence $\mathrm{sign}(\theta' x_i) = y_i$.

So we can fit *any* labels, and we should not expect the predictions for subsequent points to be accurate.

Minimax risk lower bound

Theorem: For any $n \geq 1$ and any mapping $f_n : \mathbb{R}^d \times \left(\mathbb{R}^d \times \{\pm 1\}\right)^n \to \{\pm 1\}$, there is a probability distribution P on $\mathbb{R}^d \times \{\pm 1\}$ for which some linear threshold function $f \in F$ has $R(f) = 0$ but

$$\mathrm{E}R(f_n) \geq \frac{\min(n, d) - 1}{2n} \left(1 - \frac{1}{n}\right)^n .$$

Notes:

1. f_n need not use prediction rules from the class F of linear threshold functions.
2. P can depend on n. That is, the theorem does not show that for some P the risk decreases at least as slowly as d/n. Rather, it shows that there is no uniform upper bound on risk that's better than d/n.

Minimax risk lower bound: proof

Uses the probabilistic method: choose P randomly from some class $\mathcal{P}$, and show that the expectation of $R(f_n)$ under this random choice is large. This implies that for *some* distribution in the class, $R(f_n)$ is large.

(NB: not constructive. The P must depend on the algorithm. But every algorithm must fail.)

We'd like the distributions in $\mathcal{P}$ to satisfy:

1. For some $f \in F$, $R(f) = 0$.

2. A sample of size n contains limited information about this f.

For (1), we restrict the marginal distribution on $\mathbb{R}^d$ to have support on a linearly independent set $\{v_1, \ldots, v_d\} \subset \mathbb{R}^d$. So for any $b = (b_1, \ldots, b_d) \in \{\pm 1\}^d$, there is an $f_b \in F$ with, for all i, $f_b(v_i) = b_i$.

Minimax risk lower bound: proof

For (2), we concentrate the probability on a single point, say v_d, and make the other points unlikely:

$$P_b(x, y) = \begin{cases} \frac{\epsilon}{d-1} & \text{if } (x, y) = (v_i, b_i) \text{ for } i = 1, \ldots, d-1, \\ 1 - \epsilon & \text{if } (x, y) = (v_d, b_d). \end{cases}$$

The idea is that many points will not be seen in the sample (and hence their label cannot be predicted), but they will have enough mass that these mistakes matter.

Define $U = \{v_1, \ldots, v_{d-1}\} - \{X_1, \ldots, X_n\}$ as the set of *unseen* 'light' elements of S. Let $N = |U|$.

Minimax risk lower bound: proof

Choose b uniformly at random from $\{\pm 1\}^d$ (hence P u.a.r. from $\mathcal{P}$).

$$\mathrm{E}R(f_n) = \sum_{k=0}^{d-1} \mathrm{E}\left[R(f_n)|N=k\right] \Pr(N=k)$$

and $$\mathrm{E}\left[R(f_n)|N=k\right] \geq \frac{1}{2}k\frac{\epsilon}{d-1}.$$

This is because, for the N unseen points in U, the corresponding b_i can be chosen afterwards (the bits are independent). So on those points, the decision rule can do no better than tossing a coin.

Minimax risk lower bound: proof

So

$$\begin{aligned} \mathrm{E}R(f_n) &= \sum_{k=1}^{d} \mathrm{E}\left[R(f_n) | N = k\right] \Pr(N = k) \\ &\geq \frac{\epsilon}{2(d-1)} \sum_{k=1}^{d} k \Pr(N = k) \\ &= \frac{\epsilon}{2(d-1)} \mathrm{E}N. \end{aligned}$$

But the expected number of unseen light elements is

$$\begin{aligned} \mathrm{E}N &= \sum_{i=1}^{d-1} \Pr\left(v_i \notin \{X_1, \ldots, X_n\}\right) \\ &= (d-1)\left(1 - \frac{\epsilon}{d-1}\right)^n. \end{aligned}$$

Minimax risk lower bound: proof

Thus,

$$\mathrm{E}R(f_n) \geq \frac{\epsilon}{2}\left(1 - \frac{\epsilon}{d-1}\right)^n .$$

Then choose ϵ to optimize the bound:

For $n \geq d-1$, choose $\epsilon = (d-1)/n$. Then

$$\mathrm{E}R(f_n) \geq \frac{d-1}{2n}\left(1 - \frac{1}{n}\right)^n .$$

Otherwise (if $n < d-1$), choose $\epsilon = (n-1)/n (< (d-1)/n)$. Then

$$\mathrm{E}R(f_n) \geq \frac{n-1}{2n}\left(1 - \frac{n-1}{(d-1)n}\right)^n \geq \frac{n-1}{2n}\left(1 - \frac{1}{n}\right)^n .$$

Minimax risk lower bound

Theorem: For any $n \geq 1$ and any mapping $f_n : \mathbb{R}^d \times \left(\mathbb{R}^d \times \{\pm 1\}\right)^n \to \{\pm 1\}$, there is a probability distribution P on $\mathbb{R}^d \times \{\pm 1\}$ for which some linear threshold function $f \in F$ has $R(f) = 0$ but

$$\mathbf{E} R(f_n) \geq \frac{\min(n, d) - 1}{2n} \left(1 - \frac{1}{n}\right)^n .$$

So for any method, if d/n is large, some probability distribution will cause a large excess risk.

Perceptron algorithm

We'll see that, if d/n is small, then small empirical risk over linear threshold functions ensures small risk.

The perceptron algorithm converges quickly if P allows a large margin solution. In that case, its solution incorporates few (approximately R^2/γ^2) (X_i, Y_i) pairs. The data is *compressed* in some sense. This is enough to ensure good performance.

Perceptron algorithm

Theorem: Suppose P is such that, for some $\theta \in \mathbb{R}^d$ and $\gamma > 0$,

$$\|X\| \leq R, \qquad \text{and} \qquad \frac{\theta' XY}{\|\theta\|} \geq \gamma. \qquad \text{a.s.}$$

Define f_n as the function returned by the perceptron algorithm with input $(X_1, Y_1), \ldots, (X_n, Y_n)$, and $\tilde{f}_n$ as the function returned by the perceptron algorithm with input $(X_1, Y_1), \ldots, (X_M, Y_M)$, where M is chosen uniformly from $\{1, \ldots, n\}$. Then

$$\mathrm{E} R(\tilde{f}_n) \leq \frac{R^2}{n\gamma^2}.$$

Perceptron algorithm: Proof

Define $D^m = ((X_1, Y_1), \ldots, (X_m, Y_m))$

$$\begin{aligned} \mathrm{E} R(\tilde{f}_n) &= \frac{1}{n} \sum_{m=1}^{n} \mathrm{E} \ell(f_m(X; D^m), Y) \\ &= \mathrm{E} \frac{1}{n} \sum_{m=1}^{n} \ell(f_m(X_{m+1}; D^m), Y_{m+1}), \end{aligned}$$

because (X, Y) and (X_{m+1}, Y_{m+1}) are iid. But the perceptron convergence theorem shows that

$$\sum_{m=1}^{n} \ell(f_m(X_{m+1}; D^m), Y_{m+1}) \leq \frac{R^2}{\gamma^2},$$

Perceptron algorithm: Lower bound

Idea: If algorithm makes no more than k mistakes, then expected proportion of mistakes is no more than k/n.

And this is the best we can hope for under these conditions.

Theorem: For any f_n, γ, R, d, n, there is a P on $\mathbb{R}^d \times \{\pm 1\}$ s.t. some $\theta \in \mathbb{R}^d$ has

$$\frac{\theta' XY}{\|\theta\|} \geq \gamma \qquad \text{and} \qquad \|X\| \leq R \qquad \text{a.s.,}$$

but

$$\mathrm{E}R(f_n) \geq \frac{\min(R^2/\gamma^2, n, d) - 1}{2n} \left(1 - \frac{1}{n}\right)^n.$$

Risk bounds and uniform convergence

For empirical risk minimization strategies, which choose $f_n \in F$ to minimize

$$\hat{R}(f) = \hat{\mathrm{E}}\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i),$$

how does the risk $R(f_n) = \mathrm{E}\ell(f_n(X), Y)$ behave?
Does $R(f_n) \to \inf_{f \in F} R(f)$?
How rapidly?

Risk bounds and uniform convergence

If we consider a single prediction rule f, we can appeal to the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \to \mathrm{E}\ell(f(X), Y).$$

And with some assumptions (e.g., on the moments of $\ell(f(X), Y)$), we can obtain rates. For instance, ℓ bounded implies $\Pr(|\hat{R}(f) - R(f)| > \epsilon)$ decreases exponentially in n.

For this, we'll study *concentration inequalities*, which bound the probability of deviations of random variables from their expectations. But because we use data to choose f_n, we need something stronger than a law of large numbers.

Risk bounds and uniform convergence

Example:

For pattern classification ($\mathcal{Y} = \{0, 1\}$), consider $F = F_+ \cup F_-$ with

$$\begin{aligned} F_+ &= \{1[S] : |S| < \infty\}, \\ F_- &= \{1[S] : |\mathcal{X} - S| < \infty\} \end{aligned}$$

Then for a continuous distribution on $\mathcal{X}$ with $P(Y = 1|X) = 0.9$,

$$R(f) = \begin{cases} 0.1 & \text{for } f \in F_-, \\ 0.9 & \text{for } f \in F_+. \end{cases}$$

But for any sample, there is an empirical risk minimizer $f_n \in F_+$ with $\hat{R}(f) = 0$.

Overview

1. Review: Linear threshold functions, perceptron algorithm.
2. Lower bounds (d/n) on minimax risk for linear threshold functions.
3. Upper and lower bounds ($R^2/n\gamma^2$) on minimax risk for perceptron algorithm.
4. Risk bounds, uniform convergence, concentration.