CS281B/Stat241B. Statistical Learning Theory. Lecture 10.

Peter Bartlett

- Review: Covering numbers
- Pseudodimension
- Convex losses for classification.

ERM and uniform laws of large numbers

Empirical risk minimization:

Choose $f_n \in F$ to minimize $\hat{R}$.

$$\begin{aligned} R(f_n) &\leq \inf_{f \in F} R(f) + \sup_{f \in F} \left| R(f) - \hat{R}(f) \right| + O(1/\sqrt{n}) \\ &= \inf_{f \in F} R(f) + O\left(\mathbb{E} \| R_n \|_F \right). \end{aligned}$$

Review: Covering numbers

Theorem: For $F \subseteq [-1,1]^{\mathcal{X}}$ and $x_1, \ldots, x_n \in \mathcal{X}$, consider the $L_2(P_n)$ pseudometric on F,

$$d_n(f,g)^2 = P_n(f-g)^2.$$

Then

$$\mathbb{E}\|R_n\|_F \leq \inf_{\alpha>0}\left(\mathbb{E}\sqrt{\frac{2\log(2\mathcal{N}(\alpha, F, d_n))}{n}} + \alpha\right).$$

Review: Chaining and Dudley's entropy integral

Theorem: For some universal constant c, if $F \subseteq [0,1]^{\mathcal{X}}$,

$$\mathbb{E}\|R_n\|_F \leq c\mathbb{E}\int_0^\infty \sqrt{\frac{\log \mathcal{N}(\alpha, F, d_n)}{n}}\, d\alpha.$$

Dudley's entropy integral versus the simple discretization

$$
\begin{aligned}
&\inf_{0\le\alpha\le 1}\left(\mathbb{E}\int_{\alpha}^{1}\sqrt{\frac{\log\mathcal{N}(\epsilon,F,d_n)}{n}}\,d\epsilon+\alpha\right)\\
&\le \inf_{0\le\alpha\le 1}\left(\mathbb{E}\int_{\alpha}^{1}\sqrt{\frac{\log\mathcal{N}(\alpha,F,d_n)}{n}}\,d\epsilon+\alpha\right)\\
&= \inf_{0\le\alpha\le 1}\left((1-\alpha)\mathbb{E}\sqrt{\frac{\log\mathcal{N}(\alpha,F,d_n)}{n}}+\alpha\right)\\
&\le \inf_{0\le\alpha\le 1}\left(\mathbb{E}\sqrt{\frac{\log\mathcal{N}(\alpha,F,d_n)}{n}}+\alpha\right).
\end{aligned}
$$

Review: Sudakov's Theorem

Theorem:

$$\mathbb{E}\|R_n\|_F \geq \frac{c}{\log n} \sup_{\alpha} \left(\alpha \mathbb{E} \sqrt{\frac{\log(\mathcal{N}(\alpha, F, d_n))}{n}} \right) .$$

Ignoring the $\log n$, this lower bound is the largest rectangle that we can fit under the graph of $\sqrt{\log(\mathcal{N}(\alpha, F, d_n))/n}$.

Review: Covering numbers

- There is a gap between the upper and lower bounds on $\mathbb{E}\|R_n\|_F$ in terms of covering numbers. This gap is essential.
- We have seen that $\mathbb{E}\|R_n\|_F$ gives tight bounds on $\|P - P_n\|_F$. Covering numbers do not.
- Covering numbers are convenient: it is often easy to bound them by piecing together approximations.

Overview

- Review: Covering numbers
- Pseudodimension
- Convex losses for classification.

Pseudodimension

Definition: Pollard's pseudodimension for a class $F \subseteq \mathbb{R}^{\mathcal{X}}$ is

$$d_P(F) = d_{VC}\left(\{(x,y) \mapsto \mathrm{sign}(f(x) - y) : f \in F\}\right).$$

- $\{(x,y) \mapsto \mathrm{sign}(f(x) - y) : f \in F\}$ is the set of decision rules for the epigraphs ($\{(x,y) : y \geq f(x)\}$) of functions $f \in F$.
- For $F \subseteq \{\pm 1\}^{\mathcal{X}}$, $d_P(F) = d_{VC}(F)$.

Pseudodimension

- For F a linear space of functions of dimension d, $d_P(F) = d$.
- For F a parameterized class $\{x \mapsto f(x, \theta) : \theta \in \mathbb{R}^d\}$, we have $d_P(F) = d_{VC}(G)$, where G is the parameterized class

$$G = \left\{ (x, y) \mapsto g((x, y), \theta) : \theta \in \mathbb{R}^d \right\},$$

  with $g((x, y), \theta) = \mathrm{sign}(f(x, \theta) - y)$. So all the tools for bounding VC-dimension in terms of arithmetic complexity are immediately applicable to pseudodimension.

Pseudodimension and covering numbers

Theorem: For $F \subseteq [0,1]^{\mathcal{X}}$ with $d_P(F) \leq d$,

$$\mathcal{M}(\epsilon, F, d_n) \leq \left(\frac{c}{\epsilon}\right)^{2d}.$$

Pseudodimension and covering numbers: proof

Fix $f, g : \mathcal{X} \to [0,1]$. Define $X \sim P_n$ and $Y \sim \mathcal{U}$ with $\mathcal{U} = \mathrm{Unif}[0,1]$.

$$\begin{aligned}
d_n(f,g)^2 &= P_n(f(X) - g(X))^2 \\
&= P_n(\Pr(Y \leq f(X)|X) - \Pr(Y \leq g(X)|X))^2 \\
&\leq (P_n \times \mathcal{U})(1[Y \leq f(X)] - 1[Y \leq g(X)])^2 \qquad \text{(Jensen)} \\
&= (P_n \times \mathcal{U}) \left|1[Y \leq f(X)] - 1[Y \leq g(X)]\right|.
\end{aligned}$$

Thus, for $G = \{(x,y) \mapsto 1[f(x) - y \geq 0] : f \in F\}$,

$$\begin{aligned}
\mathcal{M}(\epsilon, F, L_2(P_n)) &\leq \mathcal{M}(\epsilon^2, G, L_1(P_n \times \mathcal{U})) \\
&\leq \left(\frac{c}{\epsilon^2}\right)^{d_{VC}(G)} \qquad \text{(Haussler)} \\
&= \left(\frac{c'}{\epsilon}\right)^{2d_P(F)}.
\end{aligned}$$

Pseudodimension

Finiteness of pseudodimension is not necessary for covering numbers to give useful bounds on $\|P - P_n\|_F$:

Example: For the set F of non-decreasing functions,

$$\mathcal{N}(\epsilon, F, d_n) = n^{O(1/\epsilon)}.$$

But it is easy to see that $d_P(F) = \infty$.

Compare this to the case of $F \subseteq \{\pm 1\}^{\mathcal{X}}$, where the growth function $\Pi_F(n)$ is either 2^n or n^d.

Fat-shattering dimension

It turns out that there is a combinatorial dimension that characterizes $\mathbb{E}\|P - P_n\|_F \to 0$. It is a scale-sensitive version of the pseudo-dimension: the fat-shattering dimension $\mathrm{fat}_F(\epsilon)$.

Definition: A class $F \subseteq \mathbb{R}^{\mathcal{X}}$ ϵ-shatters $x_1, \ldots, x_n$ if there is a sequence $\gamma_1, \ldots, \gamma_n \in \mathbb{R}$ for which, for all $y \in \{\pm 1\}^n$ there is an $f \in F$ for which

$$y_i(f(x_i) - \gamma_i) \geq \epsilon.$$

And $\mathrm{fat}_F(\epsilon)$ is the size of the largest ϵ-shattered set.

Mendelson and Vershynin, improving on a result of Alon, Ben-David, Cesa-Bianchi and Haussler, showed that

$$\mathcal{N}(\epsilon, F, d_n) \leq \left(\frac{c}{\epsilon}\right)^{\mathrm{fat}_F(c\epsilon)}.$$

Overview

- Review: Covering numbers
- Pseudodimension
- Convex losses for classification.
  - Classification calibration.
  - Excess risk versus excess ϕ-risk.

Convex loss for classification

Up to this point, we have considered the performance of methods that choose f to minimize $\hat{R}(f)$. For classification, this corresponds to minimizing the number of misclassifications, which is typically a difficult combinatorial optimization problem.

(e.g., linear threshold functions.)

Convex loss for classification

Instead, there are many examples where *convex* loss functions are used for classification. While we might aim to choose a decision rule $f : \mathcal{X} \to \mathbb{R}$ to minimize

$$R(f) = \Pr(Y \neq \text{sign}(f(X))) = \mathbb{E}1[Yf(X) \leq 0],$$

we often work with f chosen to minimize a (regularized version of a) sample average of a convex loss function like:

$$\begin{aligned}
\phi_{svm}(yf(x)) &= (1 - yf(x))_+ \,, \\
\phi_{AdaBoost}(yf(x)) &= \exp(-yf(x)), \\
\phi_{logistic}(yf(x)) &= \log(1 + \exp(-yf(x))).
\end{aligned}$$

This allows the use of efficient convex optimization algorithms.

Convex loss for classification

What is the cost of this computational convenience?

We will ignore the issue of $\mathbb{E}\phi(Yf(X))$ versus $\hat{\mathbb{E}}\phi(Yf(X))$: suppose that we choose $f : \mathcal{X} \to \mathbb{R}$ to minimize $\mathbb{E}\phi(Yf(X))$. When does this lead to a good classifier (that is, with small risk)?

Convex loss for classification

Define

$$\begin{aligned} \ell(y, f(x)) &= 1[yf(x) \leq 0], \\ R(f) &= \mathbb{E}\ell(Y, f(X)), \\ R_\phi(f) &= \mathbb{E}\phi(Yf(X)). \\ \text{e.g., } \phi(yf(x)) &= (1 - yf(x))_+. \end{aligned}$$

First, we can observe that $\ell(y, f(x)) \leq c\phi(yf(x))$ implies that $R(f) \leq cR_\phi(f)$. So a small $R_\phi(f)$ gives small $R(f)$.

But this is a rather weak assurance if, for example, $\inf_f R_\phi(f) > 0$. When does minimizing R_ϕ lead to minimal R?

Convex loss for classification

Consider a *fixed* $x \in \mathcal{X}$.

$$
\begin{aligned}
\text{Define } \eta(x) &= \Pr(Y = 1 | X = x). \\
\text{Then } R_\phi(f) &= \mathbb{E}\phi(Yf(X)) \\
&= \mathbb{E}\mathbb{E}\left[\phi(Yf(X))|X\right], \\
\mathbb{E}\left[\phi(Yf(X))|X = x\right] &= \Pr(Y = 1|X = x)\phi(f(x)) \\
&\qquad + \Pr(Y = -1|X = x)\phi(-f(x)) \\
&= \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).
\end{aligned}
$$

Define the optimizer of this conditional expectation:

$$
\begin{aligned}
H(\eta) &:= \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right) \\
\alpha^*(\eta) &:= \arg \min_{\alpha \in \mathbb{R} \cup \{\pm\infty\}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).
\end{aligned}
$$

Examples

For $\phi(\alpha) = (1-\alpha)_+$,

$$\alpha^*(\eta) = \operatorname{sign}\left(\eta - \frac{1}{2}\right),$$
$$H(\eta) = 2\min(\eta, 1-\eta).$$

For $\phi(\alpha) = \exp(-\alpha)$,

$$\alpha^*(\eta) = \frac{1}{2}\log\left(\frac{\eta}{1-\eta}\right),$$
$$H(\eta) = 2\sqrt{\eta(1-\eta)}.$$

Classification calibration

The prediction $\hat{y}$ with minimal conditional risk is $\mathrm{sign}(2\eta(x) - 1)$. If the optimal conditional expectation $\mathbb{E}[\phi(Yf(X))|X = x]$ can be achieved with a value of α with the wrong sign, then minimizing R_ϕ is not useful for classification. So define

$$H^-(\eta) := \inf \left\{ \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha) : \alpha(2\eta - 1) \leq 0 \right\}.$$

Examples

For $\phi(\alpha) = (1-\alpha)_+$,

$$\begin{aligned}
\alpha^*(\eta) &= \operatorname{sign}\left(\eta - \frac{1}{2}\right), \\
H(\eta) &= 2\min(\eta, 1-\eta), \\
H^-(\eta) &= \phi(0) = 1, \\
\psi(\theta) &= 1 - 2\min\left(\frac{1+\theta}{2}, \frac{1-\theta}{2}\right) = \theta.
\end{aligned}$$

Examples

For $\phi(\alpha) = \exp(-\alpha)$,

$$\begin{aligned}
\alpha^*(\eta) &= \frac{1}{2}\log\left(\frac{\eta}{1-\eta}\right), \\
H(\eta) &= 2\sqrt{\eta(1-\eta)}, \\
H^-(\eta) &= \phi(0) = 1, \\
\psi(\theta) &= 1 - \sqrt{1-\theta^2}.
\end{aligned}$$

Classification calibration

$$H^-(\eta) := \inf\left\{\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) : \alpha(2\eta - 1) \le 0\right\}.$$

Definition: We say that ϕ is **classification-calibrated** if, for all $\eta \neq 1/2$, $H^-(\eta) > H(\eta)$.

Classification-calibration is clearly necessary for minimization of R_ϕ to lead to minimization of R. We shall see that it is also sufficient.

Classification calibration for convex ϕ

Theorem: For ϕ convex, ϕ is classification-calibrated iff

1. ϕ is differentiable at 0,
2. $\phi'(0) < 0$.

Proof: *If* is straightforward to check.

Only if: suppose that ϕ is not differentiable at 0. Then convexity implies that it lies above several tangent lines. But then for values of η near $1/2$, $\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$ is minimized by $\alpha = 0$, so ϕ is not classification-calibrated.

Also, $\phi'(0) \geq 0$ leads to $\mathrm{sign}(\alpha^*(\eta)) \neq \mathrm{sign}(\eta - 1/2)$.

Excess risk versus excess ϕ-risk

Theorem: For any nonnegative ϕ, measurable $f : \mathcal{X} \to \mathbb{R}$ and probability distribution P on $\mathcal{X} \times \{\pm 1\}$,

$$\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*,$$

where $R_\phi^* := \inf_f R_\phi(f)$, $R^* := \inf_f R(f)$, and, if ϕ is convex,

$$\psi(\theta) := H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$$

Furthermore, ϕ is classification calibrated iff

$$\psi(\theta_i) \to 0 \text{ iff } \theta_i \to 0.$$

And if ϕ is classification calibrated and convex, $\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right)$.

[When ϕ is classification calibrated, ψ is invertible.]

Excess risk versus excess ϕ-risk

If ϕ is not convex, the theorem holds with $\psi = \tilde{\psi}^{**}$, the Legendre biconjugate of

$$\tilde{\psi}(\theta) := H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right).$$

(The biconjugate g^{**} of g is the largest convex lower bound on $\tilde{\psi}$, defined by $\mathrm{epi}\, g^{**} = \overline{\mathrm{co}}\, \mathrm{epi}\, g$. So the definitions are equivalent if ϕ is convex.) [Recall that the epigraph is $\mathrm{epi}\, g = \{(x,t) : g(x) \le t\}$.]

Excess risk versus excess ϕ-risk: Proof

First, some observations about H and ψ:

1. $H(\eta) = H(1-\eta)$; $H^-(\eta) = H^-(1-\eta)$.
2. H is concave, ψ is convex.
3. $\psi(0) = 0$.
4. $\mathbb{E}H(\eta(X)) = R_\phi^*$.

Excess risk versus excess ϕ-risk: Proof

In Lecture 1, we saw that

$$R(f) - R^* = \mathbb{E}\left(1\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}\left(\eta(X) - \frac{1}{2}\right)\right]|2\eta(X) - 1|\right).$$

Since ψ is convex, Jensen's inequality implies

$$\begin{aligned}
\psi\left(R(f) - R^*\right) &\leq \mathbb{E}\psi\left(1\left[\cdots\right]|2\eta(X) - 1|\right) \\
&= \mathbb{E}1\left[\cdots\right]\psi\left(|2\eta(X) - 1|\right) && \text{(since } \psi(0) = 0\text{)} \\
&= \mathbb{E}1\left[\cdots\right]\left(H^-(\eta(X)) - H(\eta(X))\right) && \text{(def of } \psi\text{)}
\end{aligned}$$

Excess risk versus excess ϕ-risk: Proof

Now, $H^-(\eta(X))$ is the minimizer of $\mathbb{E}[\phi(Y\alpha)|X]$ when $\text{sign}(\alpha) \neq \text{sign}\,(\eta(X) - 1/2)$, so in particular, when $\text{sign}(f(X)) \neq \text{sign}\,(\eta(X) - 1/2)$, we have $H^-(\eta(X)) \leq \mathbb{E}[\phi(Yf(X))|X]$.

Also whether the sign condition is satisfied or not,

$$\mathbb{E}\left[\phi(Yf(X))|X\right] \geq H(\eta(X)).$$

Thus, considering either value of the indicator shows that

$$\begin{aligned}\psi\left(R(f) - R^*\right) &\leq \mathbb{E}\left[\phi(Yf(X)) - H(\eta(X))\right] \\ &= R_\phi(f) - R_\phi^*.\end{aligned}$$

Classification calibration for convex ϕ

Extensions:

- Every classification-calibrated ϕ is an upper bound on loss: there is a c such that $c\phi(\alpha) \geq 1[\alpha \leq 0]$.
- Flatter ϕ (smaller Bregman divergence at 0) gives a tighter bound on $R(f) - R^*$ in terms of $R_\phi(f) - R_\phi^*$.
- Under a low noise condition (that is, $\eta(X)$ is unlikely to be near $1/2$), the bound on excess risk in terms of excess ϕ-risk is improved.

Overview

- Review: Covering numbers
- Pseudodimension
- Convex losses for classification.