Statistical Learning Theory

Section 5. Convex surrogate loss

Yongdai Kim

December 6, 2021

Convex relaxation of the ERM

- In the previous sections, we have proved upper bounds on the excess risk $R(\hat{h}^{erm}) - R(h^*)$ of the empirical risk minimizer

$$\hat{h}^{erm} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Y_i \neq h(X_i)).$$

- However, the objective function is nonconvex so that the optimization problem cannot be solved in general.
- To avoid the computational problem, the basic idea is to minimize a convex upper bound of the classification error function $\mathbb{I}(\cdot)$. For the purpose, we shall also require that the function class $\mathcal{H}$ be a convex set.

Convexity

Convex set

- We say a set C is convex if for all $x, y \in C$ and $\lambda \in [0, 1]$, $\lambda x + (1 - \lambda)y \in C$.

Convex function

- We say a function $f : D \to \mathbb{R}$ is convex if it satisfies

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \forall x, y, \in D, \lambda \in [0, 1].$$

Convex relaxation

The convex relaxation takes three steps.

(Step 1): Spinning

By the relaxation $h(X) \neq Y \iff -h(X)Y > 0$, $(Y \in \{-1, 1\})$ we rewrite the objective function by

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(h(X_i \neq Y_i)) = \frac{1}{n}\sum_{i=1}^{n} \phi_1(-h(X_i)Y_i)$$

where $\phi_1(z) = \mathbb{I}(z > 0)$.

(Step 2): Soft classifier

A soft classifier is any measurable function $f : \mathcal{X} \to [-1, 1]$. The hard classifier associated to a soft classifier f is given by $h = sign(f)$.

Convex relaxation

(Step 2): Soft classifier (cont.)

Let $\mathcal{F} \in \mathbb{R}^{\mathcal{X}}$ be a convex set soft classifiers. Several popular choices of $\mathcal{F}$ are:

- Linear functions: $\mathcal{F} := \{a^\top x : a \in \mathcal{A}\}$ for some convex set $\mathcal{A} \in \mathbb{R}^d$. The associated hard classifier h splits $\mathbb{R}^d$ into two half spaces.
- Majority votes: given weak classifiers $h_1, \ldots, h_M$, $\mathcal{F} = \{\sum_{j=1}^{M} \lambda_j h_j(x) : \lambda_j \geq 0, \sum \lambda_j = 1\}$.

Convex relaxation

(Step 3): Convex surrogate

Given a convex set $\mathcal{F}$ of soft classifiers, we need to solve

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi_1(-f(X_i)Y_i).$$

However, while we are working with a convex constraint, the above objective is still not convex: we need a surrogate for the classification error.

Convex surrogate

A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called a *convex surrogate* if it is a convex non-decreasing function such that $\phi(0) = 1$ and $\phi(z) \geq \phi_1(z)$ for all $z \in \mathbb{R}$.

The following is a list of convex surrogates of loss functions.

- Hinge loss: $\phi(z) = \max(1 + z, 0)$.
- Exponential loss: $\phi(z) = e^z$.
- Logistic loss: $\phi(z) = \log(1 + \exp(z))$.

Convex relaxation

We may use a convex surrogate ϕ in place of ϕ_1 and consider minimizing the *empirical ϕ-risk* defined by

$$\hat{R}_{n,\phi}(f) = \frac{1}{n}\sum_{i=1}^{n}\phi(-Y_i f(X_i)).$$

It is the empirical counterpart of the ϕ-risk R_ϕ defined by

$$R_\phi(f) = \mathrm{E}(\phi(-Yf(X))).$$

ϕ-risk minimization

In this section, we derive the relation between the ϕ-risk $R_\phi(f)$ of a soft classifier f and the classification error $R(h) = \mathsf{P}(h(X) \neq Y)$ of its associated hard classifier $h = \text{sign}(f)$. Firstly let

$$f_\phi^* = \underset{f \in \mathbb{R}^{\mathcal{X}}}{\text{argmin}}\, \mathsf{E}(\phi(-Yf(X)))$$

where the infimum is taken over all measurable functions $f : \mathcal{X} \to \mathbb{R}$.

ϕ-risk minimization

- We will first show that if $\phi(\cdot)$ is differentiable, then $\text{sign}(f_\phi^*(X)) \geq 0$ is equivalent to $\eta(X) \geq 1/2$ where $\eta(X) = \text{P}(Y = 1|X)$. Conditional on $\{X = x\}$, we have

$$\text{E}(\phi(-Yf(X))|X = x) = \eta(x)\phi(-f(x)) + (1 - \eta(x))\phi(f(x)).$$

- Now let $H_\eta(\alpha) = \eta(x)\phi(-\alpha) + (1 - \eta(x))\phi(\alpha)$ so that

$$f_\phi^*(x) = \underset{\alpha \in \mathbb{R}}{\text{argmin}}\, H_\eta(\alpha), \text{ and } R_\phi^* = \min_f R_\phi(f) = \min_{\alpha \in \mathbb{R}} H_{\eta(x)}(\alpha).$$

ϕ-risk minization

- Since ϕ is differentiable, setting the derivative of $H_\eta(\alpha)$ to zero gives $f_\phi^*(x) = \bar{\alpha}$, where $H_\eta'(\bar{\alpha}) = -\eta(x)\phi'(-\bar{\alpha}) + (1-\eta(x))\phi'(\bar{\alpha}) = 0$, which gives

$$\frac{\eta(x)}{1-\eta(x)} = \frac{\phi'(\alpha)}{\phi'(-\bar{\alpha})}.$$

- Since ϕ is convex, its derivative ϕ' is non-decreasing. Then we have $\eta(x) \geq 1/2 \iff \bar{\alpha} \geq 0 \iff \text{sign}(f_\phi^*(x)) \geq 0$.
- Since the equivalence relation holds for all $x \in \mathcal{X}$,

$$\eta(X) \geq 1/2 \iff \text{sign}(f_\phi^*(X)) \geq 0.$$

Zhang's Lemma

The following lemma shows that if the excess ϕ-risk $R_\phi(f) - R_\phi^*$ of a soft classifier f is small, then the excess-risk of its associated hard classifier $\mathrm{sign}(f)$ is also small.

Zhang's Lemma:

Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a convex non-decreasing function such that $\phi(0) = 1$. Define for any $\eta \in [0, 1]$, $\tau(\eta) := \inf_{\alpha \in \mathbb{R}} H_\eta(\alpha)$. If there exists $c > 0$ and $\gamma \in [0, 1]$ such that

$$|\eta - \frac{1}{2}| \leq c(1 - \tau(\eta))^\gamma, \forall \eta \in [0, 1],$$

then

$$R(\mathrm{sign}(f)) - R^* \leq 2c(R_\phi(f) - R_\phi^*)^\gamma.$$

Zhang's Lemma

It is not hard to check the following values for the quantities $\tau(\eta), c$ and γ for the three losses introduced above:

- Hinge loss: $\tau(\eta) = 1 - |1 - 2\eta|$ with $c = 1/2$ and $\gamma = 1$.
- Exponential loss: $\tau(\eta) = 2\sqrt{\eta(1-\eta)}$ with $c = 1/\sqrt{2}$ and $\gamma = 1/2$.
- Logistic loss: $\tau(\eta) = -\eta \log \eta - (1-\eta)\log(1-\eta)$ with $c = 1/\sqrt{2}$ and $\gamma = 1/2$.

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

Recall that

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(X_i) \neq Y_i).$$

By considering soft classifiers (i.e., whose output is in $[-1, 1]$ rather than in $\{0, 1\}$) and convex surrogates of the loss function (e.g., hinge, exponential, logistic), we can write:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_{\phi,n}(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(-Y_i f(X_i)),$$

and $\hat{h} = \text{sign}(\hat{f})$ will be used as the corresponding hard classifier.

Now, we want to bound the quantity $R_\phi(\hat{f}) - R_\phi(\bar{f})$, where $\bar{f} = \operatorname{argmin}_{f \in \mathcal{F}} R_\phi(f)$. It can be derived by the several following steps.

- $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_{\phi,n}(f)$, thus

$$\begin{aligned} R_\phi(\hat{f}) &= R_\phi(\bar{f}) + \hat{R}_{\phi,n}(\bar{f}) - \hat{R}_{\phi,n}(\bar{f}) + \hat{R}_{\phi,n}(\hat{f}) - \hat{R}_{\phi,n}(\hat{f}) \\ &\quad + \hat{R}_\phi(\hat{f}) - \hat{R}_\phi(\bar{f}) \\ &\le R_\phi(\bar{f}) + \hat{R}_{\phi,n}(\bar{f}) - \hat{R}_{\phi,n}(\hat{f}) + \hat{R}_\phi(\hat{f}) - \hat{R}_\phi(\bar{f}) \\ &\le R_\phi(\bar{f}) + 2 \sup_{f \in \mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)|. \end{aligned} \tag{1}$$

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

- Let us focus on $E(\sup_{f\in\mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)|)$. Using the symmetrization trick as before, we know it is upper-bounded by $2\mathcal{R}_n(\phi\circ\mathcal{F})$, where the Rademacher complexity is written as

$$\mathcal{R}_n(\phi\circ\mathcal{F}) = \sup_{X_1,\dots,X_n,Y_1,\dots,Y_n} E(\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi(-Y_i f(X_i))|).$$

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

- One thing to notice is that $\phi(0) = 1$ for the loss functions we consider, but in order to apply contraction inequality later, we require $\phi(0) = 0$. Let us define $\psi(\cdot) = \phi(\cdot) - 1$. Clearly $\psi(0) = 0$, and

$$\begin{aligned}
&\mathrm{E}(\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} (\phi\,(-Y_i f(X_i)) - \mathrm{E}\,(\phi(-Y_i f(X_i)))\,|) \\
&= \mathrm{E}(\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} (\psi\,(-Y_i f(X_i)) - \mathrm{E}\,(\psi(-Y_i f(X_i)))\,|) \\
&\leq 2\mathcal{R}_n(\psi \circ \mathcal{F}).
\end{aligned} \tag{2}$$

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

- The Rademacher complexity of $\psi \circ \mathcal{F}$ is still difficult to deal with. Let us assume that ϕ is L-Lipschitz, (as a result, ψ is also L-Lipschitz), apply the contraction inequality, we have

$$\mathcal{R}_n(\psi \circ \mathcal{F}) \leq 2L\mathcal{R}_n(\mathcal{F}). \tag{3}$$

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

- Let $Z_i = (X_i, Y_i), i = 1, 2, \ldots, n$ and

$$g(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)|$$

$$= \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} (\phi(-Y_i f(X_i)) - \mathrm{E}(\phi(-Y_i f(X_i))))|$$

- Since $\phi(\cdot)$ is monotonically increasing, it is not difficult to verify that $\forall Z_1, \ldots, Z_n, Z_i'$,

$$|g(Z_1, \ldots, Z_i, \ldots, Z_n) - g(Z_1, \ldots, Z_i', \ldots, Z_n)|$$

$$\leq \frac{1}{n}(\phi(1) - \phi(-1)) \leq \frac{2L}{n}.$$

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

- The last inequality holds since g is L-Lipschitz. By applying the Bounded Difference Inequality, we have

$$P(|\sup_{f\in\mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)| - E(\sup_{f\in\mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)|)| > t)$$

$$\leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(2L/n)^2}\right).$$

- Set the RHS of above equation to δ, we get:

$$\sup_{f\in\mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)| \leq E(\sup_{f\in\mathcal{F}} |\hat{R}_{\phi,n}(f) - R_\phi(f)|) + 2L\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{4}$$

Bounding $R_\phi(\hat{f}) - R_\phi(\bar{f})$

- Now, the above steps allow us to compute the bound of $R_\phi(\hat{f}) - R_\phi(\bar{f})$.
- That is, combining equations (1) to (4), we have

$$R_\phi(\hat{f}) \leq R_\phi(\bar{f}) + 8L\mathcal{R}_n(\mathcal{F}) + 2L\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

Boosting

- We will specialize the above analysis to a particular learning model: **Boosting**. The basic idea of Boosting is to convert a set of weak learners (i.e., classifiers that do better than random, but have high error probability) into a strong one by using the weighted average of weak learners' opinions.
- More precisely, we consider the following function class

$$\mathcal{F} = \{ \sum_{j=1}^{M} \theta_j h_j(\cdot) : |\theta|_1 \leq 1, \\ h_j : \mathcal{X} \to [-1, 1], j \in \{1, \ldots, M\} \text{ are (weak) classifiers}\} \tag{5}$$

Boosting

- We want to compute the upper bound $\mathcal{R}_n(\mathcal{F})$ for this choice of $\mathcal{F}$.

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &= \sup_{Z_1,\dots,Z_n} \mathrm{E}\left(\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i Y_i f(X_i)|\right) \\ &= \frac{1}{n}\sup_{Z_1,\dots,Z_n} \mathrm{E}\left(\sup_{|\theta|\le 1} |\sum_{j=1}^{M}\theta_j \sum_{i=1}^{n} Y_i\sigma_i h_j(X_i)|\right) \end{aligned} \tag{6}$$

- It turns out that (HW)

$$\mathcal{R}_n(\mathcal{F}) \le \sqrt{\frac{2\log(4M)}{n}}.$$

- Thus for Boosting,

$$R_\phi(f) \le R_\phi(\bar{f}) + 8L\sqrt{\frac{2\log(4M)}{n}} + 2L\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1-\delta$.

To get some ideas of what values of Lipschitz constant L usually takes, consider the following examples:

- for hinge loss, i.e., $\phi(x) = (1 + x)_+$, $L = 1$.
- for exponential loss, i.e., $\phi(x) = e^x$, $L = e$.
- for logistic loss, i.e., $\phi(x) = \log(1 + e^x)$, $L = e \log_2(e)/(1 + e) \approx 2.43$.

Excess risk for Boosting

- Now we have bounded $R_\phi(\hat{f}) - R_\phi(f)$, but this is not yet the **excess risk**.
- Recall that the excess risk of $\hat{f}$ is defined as $R(\hat{f}) - R(f^*)$, where $f^* = \text{argmin}_{f \in \mathcal{F}} R_\phi(f)$.
- The following theorem in the next page provides a bound for excess risk for Boosting.

Excess risk for Boosting

Theorem

Let $\mathcal{F} = \{\sum_{j=1}^{M} \theta_j h_j : ||\theta||_1 \leq 1, h_j s$ are weak classifiers$\}$ and ϕ is an L-Lipschitz convex surrogate. Define $\hat{f} = \text{argmin}_{f \in \mathcal{F}} R_{\phi,n}(f)$ and $\hat{h} = \text{sign}(\hat{f})$. Then,

$$R(\hat{h}) - R^* \leq 2c \left(\inf_{f \in \mathcal{F}} R_\phi(f) - R_\phi(f^*) \right)^\gamma + 2c \left(8L \sqrt{\frac{2\log(4M)}{n}} \right)^\gamma + 2c \left(2L \sqrt{\frac{\log(2/\delta)}{2n}} \right)^\gamma \tag{7}$$

with probability $1 - \delta$.

Ending comments

- $O(\sqrt{1/n})$ upper bound of the excess risk is not tight.
- Under certaint conditions, it can be shown that the tight upper bound of excess risk is in between $O(1/n)$ and $(\sqrt{1/n})$.
- The proof of deriving the optimal bound is very technically involved.
- The optimal upper bound of the excess risk depends heavily on the choice of $\mathcal{H}$ or $\mathcal{F}$.
- We do not cover how to calculate the complexity (i.e. VC dimension or covering number) of a given $\mathcal{H}$ or $\mathcal{F}$ (e.g. Boosting, RKHS, Deep neural networks,...), which is also very technically involved.