18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Quan Li

Lecture 8
Oct. 5, 2015

Part II
Convexity

1. CONVEX RELAXATION OF THE EMPIRICAL RISK MINIMIZATION

In the previous lectures, we have proved upper bounds on the excess risk $R(\hat{h}^{\text{erm}}) - R(h^*)$ of the Empirical Risk Minimizer

$$\hat{h}^{\text{erm}} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq h(X_i)). \tag{1.1}$$

However due to the nonconvexity of the objective function, the optimization problem (1.1) in general can not be solved efficiently. For some choices of $\mathcal{H}$ and the classification error function (e.g. $\mathbb{1}(\cdot)$), the optimization problem can be NP-hard. However, the problem we deal with has some special features:

1. Since the upper bound we obtained on the excess risk is $O(\sqrt{\frac{d \log n}{n}})$, we only need to approximate the optimization problem with error up to $O(\sqrt{\frac{d \log n}{n}})$.
2. The optimization problem corresponds to the average case problem where the data $(X_i, Y_i) \overset{i.i.d}{\sim} P_{X,Y}$.
3. $\mathcal{H}$ can be chosen to be some 'natural' classifiers, e.g. $\mathcal{H} = \{\text{half spaces}\}$.

These special features might help us bypass the computational issue. Computational issue in machine learning have been studied for quite some time (see, e.g. [Kea90]), especially in the context of PAC learning. However, many of these problems are somewhat abstract and do not shed much light on the practical performance of machine learning algorithms.

To avoid the computational problem, the basic idea is to minimize a convex upper bound of the classification error function $\mathbb{1}(\cdot)$ in (1.1). For the purpose of computation, we shall also require that the function class $\mathcal{H}$ be a convex set. Hence the resulting minimization becomes a convex optimization problem which can be solved efficiently.

1.1 Convexity

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

Definition: A function $f : D \to \mathbb{R}$ on a convex domain D is convex if it satisfies

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y), \quad \forall x, y \in D, \text{ and } \lambda \in [0,1].$$

1.2 Convex relaxation

The convex relaxation takes three steps.

Step 1: Spinning.
Using a mapping $Y \mapsto 2Y - 1$, the i.i.d. data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is transformed to lie in $\mathcal{X} \times \{-1, 1\}$. These new labels are called *spinned* labels. Correspondingly, the task becomes to find a classifier $h : \mathcal{X} \mapsto \{-1, 1\}$. By the relation

$$h(X) \neq Y \Leftrightarrow -h(X)Y > 0,$$

we can rewrite the objective function in (1.1) by

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbb{1}}(-h(X_i)Y_i) \tag{1.2}$$

where $\varphi_{\mathbb{1}}(z) = \mathbb{1}(z > 0)$.

Step 2: Soft classifiers.
The set $\mathcal{H}$ of classifiers in (1.1) contains only functions taking values in $\{-1, 1\}$. As a result, it is non convex if it contains at least two distinct classifiers. Soft classifiers provide a way to remedy this nuisance.

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

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

- Linear functions:
$$\mathcal{F} := \{\langle a, x\rangle : a \in \mathcal{A}\}.$$
for some convex set $\mathcal{A} \in \mathbb{R}^d$. The associated hard classifier $h = \text{sign}(f)$ splits $\mathbb{R}^d$ into two half spaces.

- Majority votes: given weak classifiers $h_1, \ldots, h_M$,
$$\mathcal{F} := \Big\{ \sum_{j=1}^{M} \lambda_j h_j(x) : \lambda_j \ge 0, \sum_{j=1}^{M} \lambda_j = 1 \Big\}.$$

- Let φ_j, $j = 1, 2, \ldots$ a family of functions, e.g., Fourier basis or Wavelet basis. Define
$$\mathcal{F} := \{\sum_{j=1}^{\infty} \theta_j \varphi_j(x) : (\theta_1, \theta_2, \ldots) \in \Theta\},$$
where Θ is some convex set.

Step 3: Convex surrogate.
Given a convex set $\mathcal{F}$ of soft classifiers, using the rewriting in (1.2), we need to solve that minimizes the empirical classification error

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

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

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

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

- Hinge loss: $\varphi(z) = \max(1 + z, 0)$.
- Exponential loss: $\varphi(z) = \exp(z)$.
- Logistic loss: $\varphi(z) = \log_2(1 + \exp(z))$.

To bypass the nonconvexity of $\varphi_{\mathbb{1}}(\cdot)$, we may use a convex surrogate $\varphi(\cdot)$ in place of $\varphi_{\mathbb{1}}(\cdot)$ and consider the minimizing the *empirical φ-risk* $\hat{R}_{n,\varphi}$ defined by

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

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

$$R_\varphi(f) = \mathbb{E}[\varphi(-Y f(X))].$$

1.3 φ-risk minimization

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

Let

$$f^*_\varphi = \operatorname*{argmin}_{f\in\mathbb{R}^{\mathcal{X}}} E[\varphi(-Y f(X))]$$

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

To verify that minimizing the φ serves our purpose, we will first show that if the convex surrogate $\varphi(\cdot)$ is differentiable, then $\text{sign}(f^*_\varphi(X)) \geq 0$ is equivalent to $\eta(X) \geq 1/2$ where $\eta(X) = \mathbb{P}(Y = 1 \mid X)$. Conditional on $\{X = x\}$, we have

$$\mathbb{E}[\varphi(-Y f(X)) \mid X = x] = \eta(x)\varphi(-f(x)) + (1 - \eta(x))\varphi(f(x)).$$

Let

$$H_\eta(\alpha) = \eta(x)\varphi(-\alpha) + (1 - \eta(x))\varphi(\alpha) \tag{1.3}$$

so that

$$f_\varphi^*(x) = \operatorname*{argmin}_{\alpha\in\mathbb{R}} H_\eta(\alpha)\,, \qquad \text{and} \qquad R_\varphi^* = \min_{f\in\mathbb{R}^{\mathcal{X}}} R_\varphi(f) = \min_{\alpha\in\mathbb{R}} H_{\eta(x)}(\alpha)\,.$$

Since $\varphi(\cdot)$ is differentiable, setting the derivative of $H_\eta(\alpha)$ to zero gives $f_\varphi^*(x) = \bar{\alpha}$, where

$$H'_\eta(\bar{\alpha}) = -\eta(x)\varphi'(-\bar{\alpha}) + (1-\eta(x))\varphi'(\bar{\alpha}) = 0,$$

which gives

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

Since $\varphi(\cdot)$ is a convex function, its derivative $\varphi'(\cdot)$ is non-decreasing. Then from the equation above, we have the following equivalence relation

$$\eta(x) \geq \frac{1}{2} \Leftrightarrow \bar{\alpha} \geq 0 \Leftrightarrow \operatorname{sign}(f_\varphi^*(x)) \geq 0. \tag{1.4}$$

Since the equivalence relation holds for all $x \in \mathcal{X}$,

$$\eta(X) \geq \frac{1}{2} \Leftrightarrow \operatorname{sign}(f_\varphi^*(X)) \geq 0.$$

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

Lemma (Zhang's Lemma [Zha04]): Let $\varphi : \mathbb{R} \mapsto \mathbb{R}_+$ be a convex non-decreasing function such that $\varphi(0) = 1$. 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

$$|\eta - \frac{1}{2}| \leq c(1-\tau(\eta))^\gamma\,, \quad \forall \eta \in [0,1]\,, \tag{1.5}$$

then

$$R(\operatorname{sign}(f)) - R^* \leq 2c(R_\varphi(f) - R_\varphi^*)^\gamma$$

Proof. Note first that $\tau(\eta) \leq H_\eta(0) = \varphi(0) = 1$ so that condition (2.5) is well defined.

Next, let $h^* = \operatorname{argmin}_{h\in\{-1,1\}^{\mathcal{X}}} \mathbb{P}[h(X) \neq Y] = \operatorname{sign}(\eta - 1/2)$ denote the Bayes classifier, where $\eta = \mathbb{P}[Y = 1|X = x]$, . Then it is easy to verify that

$$\begin{aligned}
R(\operatorname{sign}(f)) - R^* &= \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(\operatorname{sign}(f(X)) \neq h^*(X))] \\
&= \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(f(X)(\eta(X) - 1/2) < 0)] \\
&\leq 2c\mathbb{E}[((1-\tau(\eta(X)))\mathbb{1}(f(X)(\eta(X) - 1/2) < 0))^\gamma] \\
&\leq 2c\left(\mathbb{E}[(1-\tau(\eta(X)))\mathbb{1}(f(X)(\eta(X) - 1/2) < 0)]\right)^\gamma\,,
\end{aligned}$$

where the last inequality above follows from Jensen's inequality.

We are going to show that for any $x \in \mathcal{X}$, it holds

$$(1 - \tau(\eta))\mathbb{1}(f(x)(\eta(x) - 1/2) < 0)] \leq \mathbb{E}[\varphi(-Yf(x)) \mid X = x] - R^*_\varphi \,. \tag{1.6}$$

This will clearly imply the result by integrating with respect to x.

Recall first that

$$\mathbb{E}[\varphi(-Yf(x)) \mid X = x] = H_{\eta(x)}(f(x)) \qquad \text{and} \qquad R^*_\varphi = \min_{\alpha \in \mathbb{R}} H_{\eta(x)}(\alpha) = \tau(\eta(x)) \,.$$

so that (2.6) is equivalent to

$$(1 - \tau(\eta))\mathbb{1}(f(x)(\eta(x) - 1/2) < 0)] \leq H_{\eta(x)}(\alpha) - \tau(\eta(x))$$

Since the right-hand side above is nonnegative, the case where $f(x)(\eta(x) - 1/2) \geq 0$ follows trivially. If $f(x)(\eta(x) - 1/2) < 0$, (2.6) follows if we prove that $H_{\eta(x)}(\alpha) \geq 1$. The convexity of $\varphi(\cdot)$ gives

$$\begin{aligned} H_{\eta(x)}(\alpha) &= \eta(x)\varphi(-f(x)) + (1 - \eta(x))\varphi(f(x)) \\ &\geq \varphi(-\eta(x)f(x) + (1 - \eta(x))f(x)) \\ &= \varphi((1 - 2\eta(x))f(x)) \\ &\geq \varphi(0) = 1 \,, \end{aligned}$$

where the last inequality follows from the fact that φ is non decreasing and $f(x)(\eta(x) - 1/2) < 0$. This completes the proof of (2.6) and thus of the 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 $\gamma = 1/2$.

References

[Kea90] Michael J Kearns. *The computational complexity of machine learning.* PhD thesis, Harvard University, 1990.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

18.657 Mathematics of Machine Learning
Fall 2015