EECS 281B / STAT 241B: Advanced Topics in Statistical Learning **Spring 2009**

Lecture 11 — February 25

Lecturer: Michael I. Jordan *Scribe: Kevin Canini*

Note: These lecture notes are still rough, and have only have been mildly proofread.

11.1 Surrogate loss functions

The 0-1 loss function has nice properties that we would like to take advantage of for many problems. However, because it is not convex, it is difficult to optimize using the 0-1 loss function, so we often turn to convex surrogate loss functions. In this lecture, we explore ways of quantifying the price we will pay by substituting a surrogate loss function for 0-1 loss.

11.1.1 Risk and ϕ-risk

Let $R(f)$ denote the risk of a classifier f based on 0-1 loss, i.e.,

$$R(f) = \mathbb{E}[\mathbb{I}[Y \neq f(X)]] \tag{11.1}$$
$$= \mathbb{P}(Y \neq f(X)), \tag{11.2}$$

and let R^* denote the Bayes risk, i.e.,

$$R^* = \inf_f R(f). \tag{11.3}$$

Given a loss function $\phi(t)$ (e.g., exponential loss, hinge loss, or logistic loss), where $t = yf(x)$ is the margin for data point (x, y), we define the "ϕ-risk" as

$$R_\phi(f) = \mathbb{E}[\phi(Yf(X))] \tag{11.4}$$
$$= \mathbb{E}[\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X))], \tag{11.5}$$

where $\eta(x) = \mathbb{P}(Y = 1|X = x)$. Similarly, we define the "optimal ϕ-risk" as

$$R^*_\phi = \inf_f R^*_\phi(f) \tag{11.6}$$
$$= \inf_f \mathbb{E}[\phi(Yf(X))] \tag{11.7}$$
$$= \inf_f \mathbb{E}[\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X))]. \tag{11.8}$$

The excess risk for a classifier f is given by $R(f) - R^*$, and the "excess ϕ-risk" is $R_\phi(f) - R^*_\phi$.

11.1.2 Classification-calibration

Looking at (11.8), for a fixed value of X, the minimum of the expectation is given by

$$H_\phi(\eta) = \inf_\alpha(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)), \tag{11.9}$$

so we can write

$$R^*_\phi = \mathbb{E}[H_\phi(\eta(X))]. \tag{11.10}$$

For a good classifier, we want $\text{sign}(f(x)) = \text{sign}(\alpha) = \text{sign}(\eta - \frac{1}{2})$, i.e., $\alpha(\eta - \frac{1}{2}) \geq 0$. So we define a quantity similar to (11.9) but optimized only where α is not a good classifier:

$$H^-_\phi(\eta) = \inf_{\{\alpha:\alpha(\eta-\frac{1}{2})\leq 0\}}(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)). \tag{11.11}$$

We define a loss function ϕ to be "classification-calibrated" if $H^-_\phi(\eta) > H_\phi(\eta)$ for all $\eta \neq \frac{1}{2}$. Intuitively, this means that the loss function strictly penalizes a classifier f for not classifying in accordance with $\eta(x)$.

Example: hinge loss is classification-calibrated

As an example, let's show that the hinge loss is classification-calibrated. The hinge loss is given by

$$\phi(t) = \begin{cases} 1-t & \text{if } t \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{11.12}$$

$$= (1-t)_+, \tag{11.13}$$

so we have

$$H_\phi(\eta) = \inf_\alpha(\eta(1-\alpha)_+ + (1-\eta)(1+\alpha)_+) \tag{11.14}$$

$$\leq \inf_{\alpha\in[-1,1]}(\eta(1-\alpha)_+ + (1-\eta)(1+\alpha)_+) \tag{11.15}$$

$$= \inf_{\alpha\in[-1,1]}(\alpha(1-2\eta)+1) \tag{11.16}$$

$$= \begin{cases} 2(1-\eta) & \text{if } \eta \leq \frac{1}{2} \\ 2\eta & \text{otherwise} \end{cases} \tag{11.17}$$

$$= 2\min(\eta, 1-\eta), \tag{11.18}$$

and

$$H^-_\phi(\eta) = \inf_{\{\alpha:\alpha(\eta-\frac{1}{2})\leq 0\}}(\eta(1-\alpha)_+ + (1-\eta)(1+\alpha)_+) \tag{11.19}$$

$$= 1, \tag{11.20}$$

so it follows that $H^-_\phi(\eta) > 2\min(\eta, 1-\eta) \geq H_\phi(\eta)$ for all $\eta \neq \frac{1}{2}$, so the hinge loss is classification-calibrated.

Theorem 11.1. *If ϕ is convex, then it is classification-calibrated if and only if it is differentiable at the origin and $\phi'(0) < 0$.*

See Bartlett et al. (posted on webpage) for the proof of this claim.

11.1.3 Excess risk bounds

Theorem 11.2. *Let ϕ be convex and classification-calibrated. Then for all f, $\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$, where $\psi(u) = H_\phi^-(\frac{1+u}{2}) - H_\phi(\frac{1+u}{2})$.*

For example, with hinge loss, we have $\psi(u) = 1 - \min(\frac{1+u}{2}, \frac{1-u}{2}) = |u|$, so Theorem 11.2 implies that $|R(f) - R^*| \leq R_\phi(f) - R_\phi^*$, i.e., that $R(f) - R^* \leq R_\phi(f) - R_\phi^*$. This means that the excess ϕ-risk for the hinge loss gives us an upper bound on the excess risk.

Proof:

$$R(f) - R^* = \mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(X)) \neq \text{sign}\left(\eta(X) - \frac{1}{2}\right)\right]|2\eta(X) - 1|\right] \tag{11.21}$$

$$= \mathbb{E}[g(X)], \tag{11.22}$$

where we define $g(x) = \mathbb{I}[\text{sign}(f(x)) \neq \text{sign}(\eta(x) - \frac{1}{2})]|2\eta(x) - 1|$. By Jensen's inequality, we have that if ψ is convex, then

$$\psi(R(f) - R^*) = \psi(\mathbb{E}[g(X)]) \tag{11.23}$$

$$\leq \mathbb{E}[\psi(g(X))]. \tag{11.24}$$

Let us verify that ψ is convex. First, because ϕ is convex and $\phi'(0) < 0$, we have

$$H_\phi^-(\eta) = \inf_{\{\alpha:\alpha(\eta-\frac{1}{2})\leq 0\}} (\eta(\phi(\alpha)) + (1-\eta)\phi(-\alpha)) \tag{11.25}$$

$$\geq \inf_{\{\alpha:\alpha(\eta-\frac{1}{2})\leq 0\}} \phi(\eta\alpha + (1-\eta)(-\alpha)) \tag{11.26}$$

$$= \inf_{\{\alpha:\alpha(\eta-\frac{1}{2})\leq 0\}} \phi(\alpha(2\eta - 1)) \tag{11.27}$$

$$\geq \phi(0). \tag{11.28}$$

But equality is achieved when $\alpha = 0$, so that $H_\phi^-(\eta) = \phi(0)$. So it follows that

$$\psi(u) = \phi(0) - H_\phi\left(\frac{1+u}{2}\right) \tag{11.29}$$

and since H_ϕ is concave (it is a pointwise minimum over linear functions), we conclude that ψ is convex, so (11.24) holds. So we have

$$\psi(R(f) - R^*) \leq \mathbb{E}\left[\psi\left(\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right]|2\eta(x) - 1|\right)\right] \tag{11.30}$$

$$= \mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right]\psi\left(|2\eta(x) - 1|\right)\right], \tag{11.31}$$

since $\psi(0) = 0$, and thus

$$\psi(R(f) - R^*) \leq \mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right](H_\phi^-(\eta(X)) - H_\phi(\eta(X)))\right]. \quad (11.32)$$

Note that if $\text{sign}(f(x)) \neq \text{sign}(\eta(x) - \frac{1}{2})$, then $\mathbb{E}[\phi(Yf(X))|X = x] \geq H_\phi^-(\eta(X))$; otherwise $\mathbb{E}[\phi(Yf(X))|X = x] \geq H_\phi(\eta(X))$. So

$$\begin{aligned}\psi(R(f) - R^*) &\leq \mathbb{E}[\phi(Yf(X)) - H_\phi(\eta(X))] &(11.33)\\ &= R_\phi(f) - R_\phi^*. &(11.34)\end{aligned}$$

□

11.2 Glivenko-Cantelli theorem

Given i.i.d. samples $x_1, \ldots, x_n$ of a random variable X generated by some cumulative distribution function (CDF) $F(x) = \mathbb{P}(X \leq x)$, we can construct an "empirical CDF"

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[x_i \leq x]. \quad (11.35)$$

For practical purposes, we want $\hat{F}$ to converge to F in some useful way. Pointwise convergence is usually not sufficient for applications, since the estimator can always be arbitrarily bad at some point, as long as it's not always at the same point. For some loss function ℓ, we define the "empirical risk" as

$$\begin{aligned}\hat{R}(f) &= \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) &(11.36)\\ &= \hat{\mathbb{E}}[\ell(f(X), Y)]. &(11.37)\end{aligned}$$

To find a regularized estimator, we take

$$\min_{f\in\mathcal{F}}(\hat{R}(f) + \lambda_n\Omega(f)) \quad (11.38)$$

with

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}}(\hat{R}(f) + \lambda_n\Omega(f)). \quad (11.39)$$

We would like it to be the case that $\hat{R}(f) \to \inf_{f\in\mathcal{F}} R(f)$.

Example

Let $\mathcal{F} = \{f\}$. Then

$$\hat{R}(f) - R(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) - \mathbb{E}[\ell(f(X), Y)] \tag{11.40}$$

$$\stackrel{a.s.}{\to} 0 \tag{11.41}$$

by the strong law of large numbers.

11.2.1 Uniform laws of large numbers

A class of functions $\mathcal{G}$ satisfies a uniform law of large numbers (ULLN) if

$$\lim_{n\to\infty} \mathbb{P}^n \left(\sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} g(z^{(i)}) - \mathbb{E}[g(Z)] \right| > \epsilon \right) \to 0. \tag{11.42}$$

Such classes $\mathcal{G}$ are called Glivenko-Cantelli.

Example

$F(x) = \mathbb{P}(X \le x) = \mathbb{E}[\mathbb{I}[X \le x]]$, with $\mathcal{G} = \{(-\infty, x] : x \in \mathbb{R}\}$.

Theorem 11.3. *(Glivenko-Cantelli). If $z^{(i)} \sim^{i.i.d.} \mathbb{P}$ and $F(t) = \mathbb{P}(Z \le t)$, then the empirical CDF $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[z^{(i)} \le t]$ satisfies*

$$\mathbb{P}\left(\sup_t \left| \hat{F}_n(t) - F(t) \right| \ge \epsilon \right) \le 8(n+1)e^{-n\epsilon/3}. \tag{11.43}$$