Lecture 12: Surrogate Losses and Calibration

Akshay Krishnamurthy
akshay@cs.umass.edu

November 1, 2017

1 Recap

Last time we wrapped up by discussing the soft margin SVM formulation

$$\min_{w,\xi} \lambda\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^n \xi_i \text{ s.t. } \forall i \in [n], y_i\langle w, x_i\rangle \geq 1-\xi, \xi_i \geq 0$$

We saw that by solving for ξ_is we can write this as a hinge loss minimization problem with ℓ_2^2 regularizer

$$\min_w \lambda\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^n \max(0, 1-y_i\langle w, x_i\rangle) = \min_w \lambda\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^n \ell_{\text{hinge}}(y_i, \langle w, x_i\rangle).$$

We also know how to prove margin-based generalization bounds of the form

$$\mathbb{P}[yf(x) < 0] \leq \mathbb{P}[yf(x) \leq \gamma] + O\left(\sqrt{\frac{1}{n\gamma^2}} + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

which can be used to understand the misclassification error of a procedure like SVM. Note that this does not actually relate the 0/1 loss to the hinge loss. It instead relates the 0/1 loss to the margin distribution. The goal for today is to understand how to relate the 0/1 loss to a *surrogate loss* like the hinge loss.

In more detail, suppose we are ultimately interested in minimizing the risk associated with the 0/1-loss $\ell_{0/1}$. For computational reasons this can be quite challenging and instead we may want to minimize the risk associated with some other loss function ℓ, a *surrogate loss*. Is this a good thing to do? More specifically, when can we have some assurance that minimizing something like the hinge loss will give us a low misclassification rate?

2 Surrogate Losses

First of all, observe that the risk associated with the 0/1 loss is non-convex, which suggests that minimizing it might be computationally challenging. The idea for using a surrogate loss function is to use a *convex* loss function that is amenable to numerical optimization techniques. For example the hinge loss with a linear model is convex in the parameter w so we can use various methods to compute the ERM for the hinge-risk. We'll discuss convex optimization really briefly in the next lecture. Today we'll think more about the statistical consequences of using a convex loss function.

To set things up let $\mathcal{F}$ be the function class we wish to optimize over and let $\mathcal{D}$ be some distribution. Let $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbf{1}\{yf(x) \leq 0\}$ be the 0/1 risk functional. Let $\phi : \mathbb{R} \to \mathbb{R}$ be some other function and define $R_\phi(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\phi(yf(x))$ to be the surrogate risk functional. Let $R^\star = \inf_f R(f)$ denote the Bayes 0/1 risk, so the infimum is over all measurable functions, and analogously let $R^\star_\phi = \inf_f R_\phi(f)$ denote the Bayes ϕ-risk. Note that these are not the best-in-class risks.

Our first goal is to relate the 0/1 risk to the ϕ-risk in a generic way. To do so, we need to better understand the optimum of the ϕ-risk. Define $\eta(x) = \mathbb{P}[Y = 1|X = x]$ to be the regression function and let's think about what happens when we predict $f(x)$ on some point x

$$\mathbb{E}[\phi(Yf(X))|X = x] = \eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x))$$

Think of this right hand side as a function of both $\eta(x)$ and $f(x)$. If we are to optimize the R_ϕ, we should choose $f(x)$ to pointwise optimize this function. Or more formally, let us define

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \eta\phi(\alpha) + (1-\eta)\phi(-\alpha).$$

With this definition, the Bayes ϕ-risk is just $R_\phi^\star = \mathbb{E}H(\eta(X))$. Observe that the optimal $f^\star$ is to predict the minimizing α at each point $x \in \mathcal{X}$.

We will also need to understand what the ϕ-risk looks like when you make a mistake. Here we will consider the same minimization over α, except with a constraint that α disagrees with the bayes optimal predictor for the 0/1 loss (which is just $\operatorname{sign}(2\eta(x)-1)$).

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

Definition 1. *The ψ-transform of a loss function ϕ is the convexified version of*

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

For now you should just think of $\tilde{\psi}$ as a convex function, which it will be in the examples of the general theory.

Theorem 2. *For any measurable f, any non-negative loss function ϕ, and any probability distribution $\mathcal{D}$*

$$\psi(R(f) - R^\star) \le R_\phi(f) - R_\phi^\star$$

The point here is that we will obtain a quantitative relationship between the ϕ-excess risk and the 0/1 excess risk. However note that we are comparing to the Bayes error. In general it is hard to get this excess risk to be very small, and typically we will require realizability type assumptions to do so.

When is the ψ function well behaved? A basic requirement we could ask for is that if we find a sequence of f_i such that $R_\phi(f_i) \to R_\phi^\star$ then we also get convergence to the 0/1 risk. This requirement is guaranteed by a condition called classification calibration.

Definition 3. *ϕ is classification calibrated if for any $\eta \neq 1/2$*

$$H^-(\eta) > H(\eta)$$

Recall that H^- is the smallest ϕ-risk if you predict the wrong thing and H is the smallest ϕ risk overall. This natural sanity check is necessary and sufficient for ϕ-risk convergence to ensure 0/1-risk convergence.

Theorem 4. *ϕ is classification calibrated if and only if for every sequence of functions f_i and every distribution $\mathcal{D}$*

$$R_\phi(f_i) \to R_\phi^\star \textit{ implies } R(f_i) \to R^\star$$

2.1 Hinge Loss

The hinge loss is $\phi(\alpha) = \max\{1-\alpha, 0\}$. We will first see that

$$\operatorname{sign}(\eta - 1/2) = \operatorname*{argmin}_\alpha \eta\phi(\alpha) + (1-\eta)\phi(-\alpha).$$

To see why, first for $\eta = 0$ the first term vanishes and any $\alpha \le -1$ makes the whole expression vanish. Similarly for $\eta = 1$, $\alpha \ge 1$ makes the whole expression vanish. For $\eta \in (0,1)$ when $\alpha \le -1$ the second term is zero, so the function is strictly decreasing and similarly when $\alpha \ge 1$ the first term is zero so the function is strictly increasing. Thus the minimum must be in $[-1,1]$, but on this interval the function is linear since the "zero" part of both maxes kick in. By inspection the minimum is attained at 1 for $\eta > 1/2$ and -1 for $\eta < 1/2$, which proves what we want.

Let us now calculate $H(\eta)$.

$$\begin{aligned} H(\eta) &= \min_\alpha \eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \\ &= \eta \max\{1 - \operatorname{sign}(\eta - 1/2), 0\} + (1-\eta)\max\{1 + \operatorname{sign}(\eta-1/2), 0\} = 2\min\{\eta, 1-\eta\} \end{aligned}$$

As for H^-, we have

$$H^-(\eta) = \min_{\alpha:\alpha(2\eta-1)\leq 0} \eta \max\{1-\alpha, 0\} + (1-\eta)\max\{1+\alpha, 0\} = 1.$$

Thus,

$$\psi(\theta) = \tilde{\psi}(\theta) = 1 - 2\min\left\{\frac{1+\theta}{2}, \frac{1-\theta}{2}\right\} = \theta.$$

So that we get $R(f) - R^\star \leq R_{\text{hinge}}(f) - R^\star_{\text{hinge}}$. It is easy to see that the hinge loss is calibrated.

2.2 Square Loss

Note that the square loss doesn't fall into the framework above since it cannot be expressed as $\phi(yf(x))$ of just the product. However, there is an elementary way to show that the square loss $(f(x)-y)^2$ is calibrated. Let $f^\star = \eta$ be the function achieving the minimal squared error.

$$f^\star = \operatorname*{argmin}_f \mathbb{E}_{(x,y)\sim\mathcal{D}}(f(x)-y)^2$$

This is clearly the regression function, and we have seen this argument before.

Theorem 5. *Let $f : \mathcal{X} \to [0,1]$, $\mathcal{D}$ be some distribution, and define $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbf{1}\{sgn(f(x)-1/2) \neq y\}$ to be the 0/1 risk. Then*

$$R(f) - R(f^\star) \leq 2\sqrt{R_{sq}(f) - R_{sq}(f^\star)}$$

Proof. The first thing to observe is that the excess squared risk can be expressed in a compact way

$$\begin{aligned} R_{\text{sq}}(f) - R_{\text{sq}}(f^\star) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}(f(x)-y)^2 - (f^\star(x)-y)^2 \\ &= \mathbb{E}_x\mathbb{E}_{y|x} f(x)^2 - 2y(f(x) - f^\star(x)) - f^\star(x)^2 \\ &= \mathbb{E}_x(f(x) - f^\star(x))^2 \end{aligned}$$

The important point there is that $\mathbb{E}y|x = f^\star(x)$. Now, for the 0/1 error

$$\begin{aligned} \mathbb{E}_{x,y}[\text{sign}(f(x)) \neq y] - \mathbb{E}_{x,y}[\text{sign}(f^\star(x)) \neq y] &= \mathbb{E}_{x,y}[\mathbf{1}\{\text{sign}(f(x)) \neq \text{sign}(f^\star(x))\}(\mathbf{1}\{y = \text{sign}(f(x))\} - \mathbf{1}\{y = \text{sign}(f^\star(x))\})] \\ &= \mathbb{E}_x[\mathbf{1}\{\text{sign}(f(x)) \neq \text{sign}(f^\star(x))\}|2f^\star(x) - 1|] \end{aligned}$$

This last step pushes the expectation over y inside and uses the fact that $\mathbb{P}[y = 1|x] = f^\star(x)$. Suppose that $f^\star(x) \geq 1/2$, which, based on the indicator implies that $f(x) \leq 1/2$. We then get

$$\begin{aligned} &2\mathbb{E}_x\mathbf{1}\{f^\star(x) \geq 1/2, f(x) < 1/2\}(f^\star(x) - 1/2) + \mathbf{1}\{f^\star(x) \leq 1/2, f(x) > 1/2\}(1/2 - f^\star(x)) \\ &\leq 2\mathbb{E}_x\mathbf{1}\{f^\star(x) \geq 1/2, f(x) < 1/2\}(f^\star(x) - f(x)) + \mathbf{1}\{f^\star(x) \leq 1/2, f(x) > 1/2\}(f(x) - f^\star(x)) \\ &\leq 2\mathbb{E}_x|f(x) - f^\star(x)|. \end{aligned}$$

Finally by Jensen's inequality

$$R(f) - R(f^\star) \leq 2\sqrt{\mathbb{E}(f(x) - f^\star(x))^2} = 2\sqrt{R_{\text{sq}}(f) - R_{\text{sq}}(f^\star)}. \qquad \square$$

2.3 Comments

1. Calibration is a basic property that we would want for any surrogate loss minimization procedure. If you come up with a surrogate loss function for some problem, you should try to check if it actually is related to the loss function you are actually trying to optimize. The theory here has been extended to other problems like ranking, multiclass classification, and structured prediction.

2. Unfortunately the calibration statements do not really say too much about what would happen if the problem is not realizable. For example, in the square loss case, if you do square loss ERM over some class $\mathcal{F}$ that does not contain $f^\star$ then you might not get a very strong guarantee on the 0/1 performance. You can prove

$$R_{\mathrm{sq}}(\hat{f}_n) - \min_{f \in \mathcal{F}} R_{\mathrm{sq}}(f) \le O\left(\frac{d \log(1/\delta)}{n}\right)$$

where d is some statistical complexity measure kind of like the VC-dimension. However this does not say anything about the approximation error term

$$\min_{f \in \mathcal{F}} R_{\mathrm{sq}}(f) - R_{\mathrm{sq}}(f^\star)$$

which is also relevant in the calibration statement. For this reason it is fairly common to assume some form of realizability for computational tractability. With realizability the approximation error term is zero.

3. A lot of loss functions that you have probably seen are calibrated. Examples include the exponential loss that we saw before, the sigmoid loss $\phi(\alpha) = 1 - \tanh(\alpha)$, which is non-convex but also calibrated.

3 Proofs

Proof of Theorem 2. Expanding the definitions

$$R(f) - R^\star = \mathbb{E}\mathbf{1}\{\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\}|2\eta(X) - 1|.$$

Using the fact that ψ is convex by definition, and also that $\psi(0) = 0$ (since $H^-(1/2) = H(1/2)$ and ψ is non-negative), we can apply Jensen's inequality

$$\begin{aligned}
\psi(R(f) - R^\star) &\le \mathbb{E}\psi\left(\mathbf{1}\{\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\}|2\eta(X) - 1|\right) \\
&= \mathbb{E}\mathbf{1}\{\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\}\psi(|2\eta(X) - 1|) \\
&\le \mathbb{E}\mathbf{1}\{\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\}\tilde{\psi}(|2\eta(X) - 1|) \\
&= \mathbb{E}\mathbf{1}\{\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\}\left(H^-(\eta(X)) - H(\eta(X))\right) \\
&= \mathbb{E}\mathbf{1}\{\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\}\left(\inf_{\alpha : \alpha(2\eta(X)-1) \le 0} \eta(X)\phi(\alpha) + (1 - \eta(X))\phi(-\alpha) - H(\eta(X))\right) \\
&\le \mathbb{E}\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X)) - H(\eta(X)) \\
&= R_\phi(f) - R_\phi^\star
\end{aligned}$$

Most of the steps are straightforward. The only tricky one is where we use the fact that $\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)$ to argue that $f(X)$ is in the domain of the minimization over α. Otherwise we are just working through definitions. □

Proof sketch of Theorem 4. For one direction the intuition is pretty straightforward but the details are a bit messy. Suppose ϕ is not classification calibrated. Then there exists some value η such that $H(\eta)^- \le H(\eta)$. Ignoring limiting behavior, this means that there is some α with $\alpha(2\eta - 1) \le 0$ but where $\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)$ is as small as possible, among *all* choices for α. In particular, if we think of a distribution with a single point x and with noise η on that point, then $f(x) = \alpha$ has zero excess ϕ risk but has non-zero 0/1 error, since $\alpha(2\eta - 1) \le 0$.

For the other direction, observe that due to the pointwise inequality, we get that

$$\lim_{i \to \infty} \psi(R(f_i) - R^\star) \le \lim_{i \to \infty} R_\phi(f_i) - R_\phi^\star = 0$$

Thus we need to prove that when ϕ is classification calibrated, $\psi(\theta_i) \to 0$ implies $\theta_i \to 0$. The basic fact here is that for calibrated ϕ, we must have $\psi(\theta) > 0$ for $\theta \in (0, 1]$, since $H^-(\eta) > H(\eta)$ except for at $\eta = 1/2$, which corresponds to $\theta = 0$. Doing this carefully with limits proves the result. □