CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 9**

Convex loss vs. 0–1 loss

Lecturer: Peter Bartlett *Scribe: Shaunak Chatterjee and Norm Aleks*

1 Marginalized kernel, continued

Following up on the prior lecture, we verify that the marginalized kernel is indeed a kernel. First define a kernel on x: $k((x,h),(x',h'))$. (For example, x could be a DNA sequence and h, the hidden variable, could be the role it plays in the genome.) Using $\Pr(x,h)$ and $\Pr(h|x)$,

$$\begin{aligned}
k_m(x,x') &= \sum_{h,h'} k((x,h),(x',h'))\Pr(h|x)\Pr(h'|x') \\
&= \sum_{h,h'} k((x,h),(x',h'))k_1((h,x),(h',x')) \quad (k_1 \text{ is a kernel on } (x,h) \text{ pairs}) \\
&= \sum_{h,h'} k_2((h,x),(h',x')) \quad (\text{The product of two kernels is also a kernel}) \\
&= \sum_{h,h'} \Phi(h,x)^T\Phi(h',x') \\
&= \left[\sum_h \Phi(h,x)\right]^T \left[\sum_{h'} \Phi(h',x')\right] \\
&= \tilde{\Phi}(x)^T\tilde{\Phi}(x') \\
&= \tilde{k}(x,x')
\end{aligned}$$

2 0–1 loss vs. convex loss

Or, what's the impact of using a computationally convenient loss function?

The basic optimization problem we are working with is:

$$\min_w \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^n \phi(y_i w^T x_i)$$

where ϕ is any convex function; so far, we have been using the hinge function, $\phi(\alpha) = (1-\alpha)_+$ (where $x_+ = \max(0,x)$), as our example. In the optimization, the first addend increases with the complexity of the function, and the second decreases with a better fit to the data. This is an example of a *regularized empirical risk criterion*, which in general has the form

$$\text{Regularized empirical risk} = (\text{loss of } f \text{ on sample}) + (\text{complexity penalty on } f)$$

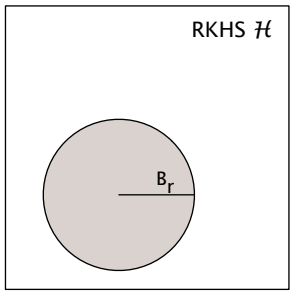

Figure 1: *We are working with functions within this ball, of radius B_r, in the Hilbert space.*

Think of the regularized empirical risk as equivalent to

$$\min_f (\text{loss of } f \text{ on sample}) \quad \text{s.t. complexity } (f) \leq B$$

There is a trade-off between the approximation and estimation errors: as we choose a more complex F (for example, larger radius B), we have a richer function class, and so can obtain smaller approximation error. However, at the same time, the estimation error—that is, the difference between the performance of the function that we choose and the performance of the best function in the class—becomes larger. This is illustrated in Figure 2.

For today we are focusing only on the difference in the expectation of the losses, and considering what happens when we replace the 0-1 loss, which is of interest in pattern classification, with a convex loss. That is, we might wish to choose $f \in F$, where $F = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\}$, to minimize the risk,

$$\Pr(Y \neq f(X)) = \mathbb{E}1[y_i \neq f(x_i)]$$

For computational convenience we work with a convex loss function rather than the indicator/step loss function. That is, we consider:

$$\min_{f \in F} \mathbb{E}\ \phi(Yf(X)) \qquad \text{vs.} \qquad \min_{f \in F} \mathbb{E}\ 1[Y \neq \text{sign}(f(X))]$$

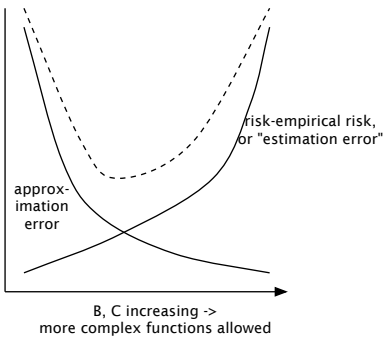

Figure 2: *A look at the trade-off between approximation error and estimation error for increasing values of B_r and C.*

This naturally leads to an interesting question: when does minimization of $R_\phi(f)$ (which equals $\mathbb{E}\ \phi(Yf(x))$) lead to small $R(f)$ (which equals $\mathbb{E}\ 1[Y \neq \text{sign}(f(X))$)?

Observation: If $\phi(\alpha) \geq 1[\alpha \leq 0]$ (that is, the loss according to ϕ is always at least the true loss), then $R(f) \leq R_\phi(f)$. (This is a weak observation if $R^* = \inf R(f) > 0$.)

If it is the case that $R^* > 0$, what more can we say? When does $R_\phi(f) = R_\phi^*(f)$ $(= \inf R_\phi(f))$ imply that $R(f) = R^*$?

Let's consider a fixed $x \in X$. Define $\eta(x) = \Pr(Y = 1|X = x)$.

$$\begin{aligned} R_\phi(f) &= \mathbb{E}\ \phi(Y(f(x))) \\ &= \mathbb{E}\ \mathbb{E}\ [\phi(Yf(x))|X] \end{aligned}$$

and

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

Define the optimal value of this criterion as

$$\begin{aligned} &\mathcal{H} : [0, 1] \to \mathbb{R} \\ &\mathcal{H}(\eta) = \inf_{\alpha \in \mathbb{R}} (\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)) \end{aligned}$$

(For example, substituting in the hinge function for ϕ, we get $\mathcal{H}(\eta) = 2\min(\eta, 1 - \eta)$). Now let us define:

$$\alpha^*(\eta) = \operatorname*{arg\,min}_{\alpha \in \mathbb{R} \cup \{\pm\infty\}} (\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha))$$

For example, using the hinge function for ϕ, we get:

$$\alpha^*(\eta) = \text{sign}(\eta - \frac{1}{2}).$$

Choice of the wrong sign of α, that is, different from $\text{sign}(\eta - 1/2)$, must result in a value of $\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)$ that is larger than $H(\eta)$ (otherwise minimization won't yield the correct answer), so we first define an optimizer with the "wrong sign":

$$\mathcal{H}^-(\eta) = \inf_{\alpha \text{ s.t. } \alpha(\eta - 1/2) \leq 0} (\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha))$$

Definition. ϕ is "classification-calibrated" if $\eta \neq \frac{1}{2} \Rightarrow \mathcal{H}^-(\eta) > \mathcal{H}(\eta)$

Theorem 2.1. For ϕ convex and classification-calibrated,

$$\forall f, \Psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^* \text{ , where}$$

$$\begin{aligned} \Psi(\theta) &= \mathcal{H}^- \left(\frac{1 + \theta}{2} \right) - \mathcal{H} \left(\frac{1 + \theta}{2} \right) \\ &= \phi(0) - \mathcal{H} \left(\frac{1 + \theta}{2} \right) \end{aligned}$$

Also, $\Psi(0) > 0$ iff $\phi > 0$.

For example, with the hinge function,

$$\begin{aligned}\Psi(\theta) &= 1 - \mathcal{H}\left(\frac{1+\theta}{2}\right) \\ &= 1 - 2\min\left(\frac{1+\theta}{2}, \frac{1-\theta}{2}\right) \\ &= 1 - 2\left(\frac{1}{2} + \frac{1}{2}\min(\theta, -\theta)\right) \\ &= |\theta|\end{aligned}$$

PROOF. Recall from Lecture 1 that

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

By Jensen's inequality,

$$\begin{aligned}\Psi(R(f) - R^*) &\leq \mathbb{E}\,\Psi\left(1\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right]\ |2\eta(x) - 1|\right) \quad \text{because } \phi(0) = 0 \\ &= \mathbb{E}\left[1\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right]\ \Psi(|2\eta(x) - 1|)\right] \quad \text{by } \phi\text{'s def. and symmetry of } \mathcal{H},\ \mathcal{H}^- \text{ around } \frac{1}{2} \\ &= \mathbb{E}\left[1\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right]\ (\mathcal{H}^-(\eta(x)) - \mathcal{H}(\eta(x)))\right] \\ &\leq \mathbb{E}\left[\phi(Yf(x)) - \mathcal{H}(\eta(x))\right] \\ &= R_\phi(f) - R_\phi^* \quad \text{by definitions of } R_\phi(f) \text{ and } \mathcal{H}\end{aligned}$$

The final inequality follows because:

- if $\text{sign}(f) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)$, then $\mathbb{E}\left[\phi(Yf(x))|x\right] \geq \mathcal{H}^-(\eta(x))$;
- otherwise, $\mathbb{E}\left[\phi(Yf(x))|x\right] \geq \mathcal{H}(\eta(x))$.

□

Lemma 2.2. For ϕ convex, ϕ is classification-calibrated iff

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

PROOF. If conditions (1) and (2) are true, it is easy to verify that the convex function ϕ is classification-calibrated.

Now for the "only if" part:

Suppose ϕ is not differentiable at zero, then it can be shown that even small perturbations of η around $\eta = 0.5$ will not move the minimum away from $\phi(0)$. This is illustrated in the figure when $\eta = 0.4$ and $\eta = 0.6$, $\mathcal{H}(\eta) = \mathcal{H}^-(\eta) = \phi(0)$, which obviously violates our definition of ϕ being classification calibrated.

□

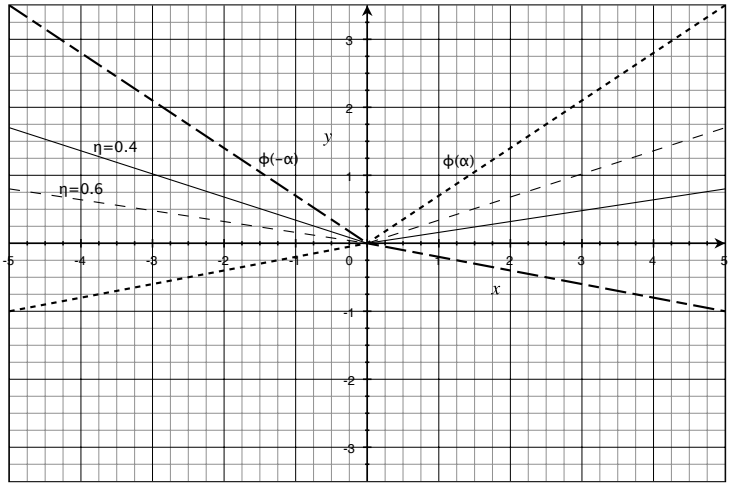

Figure 3: *Convex ϕ where ϕ is not differentiable at 0.*

It is worthwhile to consider the possibility of making the RHS of Theorem 2.1 closer to zero, since it is the upper bound on the function of the excess risk.

Notice that $R_\phi(f) - R_\phi^*$ will decrease monotonically as we increase the radius of the ball in Hilbert space, allowing more and more functions to be considered.

$$R_\phi(f_n) - R_\phi^* = (R_\phi(f_n) - \inf_{f \in B_r} R_\phi(f)) + (\inf_{f \in B_r} R_\phi(f) - R_\phi^*)$$