Consistency of convex ERM for classification problems, part 2.

Administrative/Meta:

- Key reference for today, one of my favorite papers of all time: Zhang (2004).

Overview

Let's refresh on some definitions from last time. First, recall the notions of convex and regular risk from last time. As one point of departure, today's lecture will consider risk over a distribution and not a finite sample.

$$\begin{aligned} X &\qquad& \text{input random variable;} \\ Y &\qquad& \text{output random variable;} \\ \mathcal{R}(f) := \mathbb{E}(\ell(-Yf(X))) &\qquad& \text{convex risk;} \\ \mathcal{R}_{\text{z}}(f) := \Pr[\text{sgn}(f(X)) \neq Y] &\qquad& \text{risk (misclassification rate.} \end{aligned}$$

We can construct an optimal classifier: for any x, we should predict $+1$ when $\Pr[Y = 1|X = x] > 1/2$ and -1 when $\Pr[Y = 1|X = x] < 1/2$ (and the case $\Pr[Y = 1|x = x] = 1/2$ is irrelevant).

$$\begin{aligned} \eta(x) &:= \Pr[Y = 1|X = x] &\quad& \text{"regression function";} \\ \bar{g}(x) &:= \text{sgn}(2\eta(x) - 1) &\quad& \text{"Bayes decision rule".} \end{aligned}$$

Key question (just like last time)**:** under what conditions will a function f with small $\mathcal{R}(f)$ also achieve small $\mathcal{R}_{\text{z}}(f)$, where the former is something we can optimize tractably, but the latter is what we actually care about? Last time we gave a scenario where this is impossible; this time we'll circumvent the situation by allowing the function class to grow.

Remark. As a technical note, one needs some conditions on the space to break a joint distribution on (X, Y) into a marginal distribution on X and a conditional distribution of Y given X. In machine learning and statistics, it's usually safe to ignore this technicality. Anyone interested in more details can see the book by Kallenberg (search for "disintegration") or just ping me. *[future matus: give a proper reference.]*

Circumventing the negative scenario

The negative scenario from last lecture worked with linear predictors $x \mapsto \langle w, x\rangle$ over $\mathbb{R}^1$. These predictors have a severe limitation: either $w = 0$, or the sign of the prediction over $\mathbb{R}_{++}$ is the opposite of the sign of the prediction on $\mathbb{R}_{--}$.

Said another way, the function class is not *expressive*: given a pair of points (x, x') and desired labels (y, y'), we can't in general expect our function class to contain a function f with $\text{sgn}(f(x)) = y$ and $\text{sgn}(f(x')) = y'$.

Remark.

- Notice the similarity to material from the early lectures on representation; for instance, the invocation of Stone-Weierstrass required the mappings to "separate" points, which meant the ability to map different input points to different values.

- The requirement to make different predictions on different points will be at odds with statistical issues, namely the third part of the course, “generalization”. In machine learning, we train over a finite sample from a distribution, but we aim to predict well on some underlying distribution. As we consider function classes which are more and more expressive, there will be a bigger difference between error on the finite draw and error on the distribution.

The preceeding requirement of expressiveness is sufficient to defeat the negative scenario: namely, that example had two point masses which should both be labeled positive, so any class of functions that can assign distinct labels to pairs of points will suffice.

There is one more issue to resolve, after which we can attack the positive scenario in full. Suppose many points have the same input x, but differ in their choice of label y. The optimal classifier $\bar{g}$ will agree with majority vote on x.

For this, we will need a condition on the loss function: we need the loss function to *agree with majority vote.* Let x be arbitrary, and consider two cases.

- If $\eta(x) = 1/2$, it doesn’t matter what ℓ prefers.
- Suppose $\eta(x) \neq 1/2$. We want to avoid the situation that x maps to some α with $\operatorname{sgn}(\alpha) \neq \bar{g}(x)$. This is a clumsy expression, which we will relax to say: we want to avoid $(2\eta(x) - 1)\alpha \leq 0$ (which implies the preceding). In terms of the loss, we want the loss to be better outside of the preceding case, meaning we want
  $$\inf_{\alpha\in\mathbb{R}} \eta(x)\ell(-\alpha) + (1-\eta(x))\ell(\alpha) < \inf_{\substack{\alpha\in\mathbb{R}\\ (2\eta(x)-1)\alpha\leq 0}} \eta(x)\ell(-\alpha) + (1-\eta(x))\ell(\alpha).$$
  This is one way of formalizing “the loss agrees with majority vote”.

Let’s see how this helps us with the following completely heuristic derivation (it has numerous technical issues which will be discussed momentarily). Suppose $\mathcal{F}$ contains all possible functions. Then

$$\bar{f} := \text{“argmin}_{f\in\mathcal{F}}\text{”}\mathcal{R}(f) = \text{“argmin}_{f\in\mathcal{F}}\text{”}\mathbb{E}(\ell(-Yf(X))) = \text{“argmin}_{f\in\mathcal{F}}\text{”}\mathbb{E}(\mathbb{E}(\ell(-Yf(X) \mid X)))$$

can be evaluated pointwise, meaning for every x

$$\bar{f}(x) = \text{“argmin}_{\alpha\in\mathbb{R}}\text{”}\mathbb{E}(\ell(-Y\alpha) \mid |X = x) = \text{“argmin}_{\alpha\in\mathbb{R}}\text{”}(\eta(x)\ell(-\alpha) + (1-\eta(x))\ell(\alpha));$$

the “expressiveness” property allowed us to consider optimization over each x independently, without choice on some x constraining the choice on another x'. Furthermore, on any fixed x, we know by the preceeding “agrees with majority vote” property that either $\eta(x) = 1/2$, or $\operatorname{sgn}(\bar{f}(x))$ agrees with $\bar{g}(x)$.

Remark.

- There are two reasons argmax is in quotes. First, consider $\ell = \exp$ (and similar losses which asymptote towards 0 without attaining it): If $\eta(x) \in \{0, 1\}$, then the coordinate-wise argmin is off at $-\infty$ or $+\infty$, in particular there is no minimum in $\mathbb{R}$. Second, due to measure-theoretic restrictions, we can’t quite minimize over all possible functions, but only all measurable functions, and specifically can’t control f over every point; but we can get pretty close, and the above heuristic sketch is basically fine.
- The above inuition of “loss function should agree with majority vote” also holds in the multiclass case, and indeed this was also handle by Tong Zhang in 2004 *[future matus: citation please!].* In detail, suppose we now have k possible labels $\{1, \ldots, k\}$, and we want to find a good function $f : \mathbb{R}^d \to \mathbb{R}^k$, and we will predict according to
  $$x \mapsto \operatorname{argmax}_i f(x)_i.$$
  Similarly, the Bayes decision rule becomes
  $$x \mapsto \operatorname{argmax}_i \Pr[Y = i | X = x].$$
  Notice that everything so far agrees with the preceding binary development when $k = 2$; All we need is to translate the “agrees with majority vote” expression for what is now a *vector* of loss functions

$(\ell_i)_{i=1}^k$. For this, fix any x and any label j. If $\Pr[Y=j|X=x] < \max_i \Pr[Y=i|X=x]$ (meaning we have thrown out the earlier $\eta(x) = 1/2$ case, and moreover take j to be some suboptimal label), then we want

$$\inf_{\alpha\in\mathbb{R}^k} \sum_i \Pr[Y=i|X=x]\ell_i(\alpha_i) < \inf\left\{\sum_i \Pr[Y=i|X=x] : \alpha \in \mathbb{R}^k, \alpha_j = \max_i \alpha_i\right\}.$$

This generalization is quite powerful, and suggests further generalizations to cases beyond multiclass (e.g., structured output prediction). Moreover, if we apply some normalization, we can treat the vector $(\ell_i(\alpha_i))_{i=1}^k$ as a conditional probability model; this idea can be found in the two papers by Tong Zhang *[future matus: specific refs, please.]*

Positive scenario: a proper theorem with rates

So far we have not given a real theorem statement. Also, we have only given an outlandish condition (you need to fit all possible functions!) under which we agree with $\bar{g}$: what we really want is a way to translate between suboptimality in $\mathcal{R}$ and suboptimality in $\mathcal{R}_z$, meaning some Φ with

$$\mathcal{R}_z(f) - \inf_{g \text{measurable}} \mathcal{R}_z(g) \le \Phi\left(\mathcal{R}(f) - \inf_{g \text{measurable}} \mathcal{R}(g)\right)$$

Remark. Here we will give a simplified version of the analysis by Zhang (2004). While this analysis may seem more restrictive than the analysis due to Bartlett, Jordan, and McAuliffe (2006) (which basically gives Φ as above in general settings), one needs an explicit form of Φ (or an upper bound on it) to get a true rate, which is exactly what the analysis of Zhang (2004) gives us. (Note the historical record is a little complicated, as Bartlett, Jordan, and McAuliffe (2006) had a tech report in 2003.)

To see how we could get a rate, let's try to probe the left hand side a little more closely. Note that

$$\begin{aligned}
\mathcal{R}_z(f) - \mathcal{R}_z(\bar{g}) &= \mathbb{E}\left(\Pr[\operatorname{sgn}(f(X)) \ne Y|X] - \Pr[\operatorname{sgn}(\bar{g}(X) \ne Y|X]\right) \\
&= \mathbb{E}\left(\mathbf{1}[\bar{g}(X) = \operatorname{sgn}(f(X))]\cdot 0 + \mathbf{1}[\bar{g}(X) \ne \operatorname{sgn}(f(X))]\left(\max\{\eta(X), 1-\eta(X)\} - \min\{\eta(X), 1-\eta(X)\}\right)\right) \\
&= \mathbb{E}\left(\mathbf{1}[\bar{g}(X) \ne \operatorname{sgn}(f(X))]|2\eta(X)-1|\right).
\end{aligned}$$

This is saying that not just $\operatorname{sgn}(2\eta(x)-1) = \operatorname{sgn}(\bar{g}(x))$ is important, but moreover the magnitude $|2\eta(x)-1|$. This is precisely the idea behind the analysis due to Zhang (2004): we control the gap between $\mathcal{R}_z$ and $\mathcal{R}$ by noticing how ℓ scales along with $|2\eta(x)-1|$.

Theorem (Zhang 2004)**.** Suppose $\ell : \mathbb{R} \to \mathbb{R}_+$ is convex, $\partial\ell(0) \cap \mathbb{R}_+ \ne \emptyset$, and there exist constants $c \ge 0$ and $r \ge 1$ so that, for every x,

$$|2\eta(x) - 1| \le c(\phi(0,x) - \inf_{\alpha\in\mathbb{R}} \phi(\alpha, x))^{1/r}$$

where $\phi(\alpha, x) = \eta(x)\ell(-\alpha) + (1-\eta(x))\ell(\alpha)$. Then

$$\mathcal{R}_z(f) - \inf_{g \text{measurable}} \mathcal{R}_z(g) \le c\left(\mathcal{R}(f) - \inf_{g \text{measurable}} \mathcal{R}(g)\right)^{1/r}.$$

Remark.

- For the hinge loss, $c = r = 1$; for all other losses we discussed last time, $r = 2$ and $c \in \{1, \sqrt{2}\}$. Regardless, the left hand side goes to zero as the right hand side goes to zero, meaning we minimize $\mathcal{R}_z$ if we minimize $\mathcal{R}$ over all measurable functions. (See Zhang (2004) for details, including a very nice comparison of the loss functions.)

- We can restate the condition without invoking $\eta(x) = \Pr[Y = 1|X = x]$ (which means we must discuss measures and not just properties of ℓ, simply by saying it must hold for all x and all *scalars* $\eta \in [0, 1]$.

Proof. Starting from the previous derivation, and using the various conditions and also Jensen's inequality,

$$\begin{aligned}
\mathcal{R}_z(f) - \mathcal{R}_z(\bar{g}) &= \mathbb{E}\left(\mathbf{1}[\bar{g}(X) \neq \operatorname{sgn}(f(X))]|2\eta(X) - 1|\right) \\
&\leq \mathbb{E}\left(\mathbf{1}[f(X)(2\eta(X) - 1) \leq 0]|2\eta(X) - 1|\right) \\
&\leq c\mathbb{E}\left(\mathbf{1}[f(X)(2\eta(X) - 1) \leq 0]^{1/r}(\phi(0, x) - \inf_\alpha \phi(\alpha, x))^{1/r}\right) \\
&\leq c\left(\mathbb{E}\left(\mathbf{1}[f(X)(2\eta(X) - 1) \leq 0](\phi(0, x) - \inf_\alpha \phi(\alpha, x))\right)\right)^{1/r}.
\end{aligned}$$

If we can prove that $f(x)(2\eta(x) - 1) \leq 0$ implies $\phi(0, x) \leq \phi(f(x), x)$, then the proof is complete. To this end, let $s \in \partial\ell(0) > 0$ be arbitrary, and note by two applications of the definition of subgradient that

$$\begin{aligned}
\ell(-\alpha) &\geq \ell(0) + s(-\alpha - 0), \\
\ell(\alpha) &\geq \ell(0) + s(\alpha - 0).
\end{aligned}$$

Respectively scaling these by $\eta(x)$ and $1 - \eta(x)$ and then summing the result,

$$\begin{aligned}
\phi(\alpha, x) &= \eta(x)\ell(-\alpha) + (1 - \eta(x))\ell(\alpha) \\
&\geq \eta(x)\ell(0) + (1 - \eta(x))\ell(0) + s(-\eta(x)\alpha + (1 - \eta(x))\alpha) \\
&= \phi(0, x) + s\alpha(1 - 2\eta(X)) \\
&\geq \phi(0, x).
\end{aligned}$$

□

References

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. 2006. "Convexity, Classification, and Risk Bounds." *Journal of the American Statistical Association* 101 (473): 138–56.

Zhang, Tong. 2004. "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization." *The Annals of Statistics* 32: 56–85.