COMS 4773: Surrogate losses

Daniel Hsu

March 21, 2024

1 Empirical risk minimization for linear threshold functions

Recall the hypothesis class $\mathsf{LTF}_d = \{h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ of linear threshold functions $h_{w,b} \colon \mathbb{R}^d \to \{-1, 1\}$ on $\mathbb{R}^d$, where

$$h_{w,b}(x) = \mathrm{sign}(\langle x, w \rangle + b).$$

Finding a linear separator for a linearly separable dataset can be formulated as a (polynomial-size) linear program. This implies that there is a polynomial-time algorithm for PAC learning LTF_d.[1]

However, if the training dataset is not linearly separable, then it is not immediately obvious how to find a good linear separator. It turns out that finding a linear threshold function that minimizes the number of misclassified examples (i.e., ERM for LTF_d) is NP-hard (Johnson and Preparata, 1978).[2]

We prove the hardness of the corresponding decision problem: given a dataset $S \subset \{0,1\}^d \times \{-1,1\}$ and an integer k, is there a linear threshold function that misclassifies at most k examples in S? To show the hardness of this problem, we give an efficient reduction from the Vertex Cover problem, which is well-known to be NP-complete (Karp, 1972). The Vertex Cover problem is as follows: given an undirected graph $G = (V, E)$ and an integer k, is there a vertex cover $U \subseteq V$ of cardinality at most k? A vertex cover for a graph is a subset of the vertices that contains, for every edge $\{u, v\}$ in the graph, at least one of u and v.

Let $G = (V, E)$ and k be an instance of the Vertex Cover problem. Denote the vertices by $V = \{1, 2, \ldots, n\}$, and let m be the number of edges. We construct a dataset $S \subset \{0,1\}^d \times \{-1,1\}$ with $d = 2n$ and $|S| = n + 2m$ based on the graph G. Let $e_1, \ldots, e_{2n}$ be the standard coordinate basis vectors for $\mathbb{R}^{2n}$. For any distinct $i, j \in [2n]$, define the vector $x_{i,j} := e_i + e_j \in \mathbb{R}^{2n}$. We construct the dataset S as follows:

- For each $i \in V$, add the labeled example $(x_{i,n+i}, +1)$ to S. These are the positive examples, one per vertex.
- For each edge $\{i, j\} \in E$, add the labeled examples $(x_{i,j}, -1)$ and $(x_{n+i,n+j}, -1)$ to S. These are the negative examples, two per edge.

We show the following:

$$\begin{aligned} & \text{There is a linear threshold function } h_{w,b} \text{ that misclassifies at most } k \text{ examples in } S. \\ \Leftrightarrow\ & \text{There is a vertex cover for } G \text{ of cardinality at most } k. \end{aligned}$$

[1] One can view boosting as an algorithm for PAC learning LTF_d under the additional "margin" assumption (which is equivalent to the weak learning assumption).

[2] In fact, it is hard to even solve this problem approximately (Guruswami and Raghavendra, 2009).

($\Rightarrow$) Suppose there is a linear threshold function $h_{w,b}$ that misclassifies at most k examples in S. Let $U \subseteq V$ be defined as follows:

- For each vertex $i \in V$, if $h_{w,b}(x_{i,n+i}) = -1$ (a mistake), then add i to U.
- For each edge $\{i,j\} \in E$, if $h_{w,b}(x_{i,j}) = +1$ (a mistake) or $h_{w,b}(x_{n+i,n+j}) = +1$ (also a mistake), then add exactly one of i and j to U.

Since $h_{w,b}$ misclassifies at most k examples in S, this set U has cardinality at most k. We claim that U is a vertex cover for G. Consider any edge $\{i,j\} \in E$. If $h_{w,b}$ misclassifies $x_{i,n+i}$ or $x_{j,n+j}$, then $i \in U$ or $j \in U$. Now instead suppose $h_{w,b}$ correctly classifies both $x_{i,n+i}$ and $x_{j,n+j}$. Writing $w = (w_1, \ldots, w_{2n})$, we must have

$$\begin{aligned} & w_i + w_{n+i} + b \geq 0 \\ \text{and} \qquad & w_j + w_{n+j} + b \geq 0, \end{aligned}$$

which implies

$$(w_i + w_j + b) + (w_{n+i} + w_{n+j} + b) \geq 0.$$

This means that at least one of $x_{i,j}$ and $x_{n+i,n+j}$ is misclassified, so we include one of i and j in U. We conclude that U is a vertex cover for G.

($\Leftarrow$) Now instead suppose G has a vertex cover U of cardinality at most k. Define the linear threshold function $h_{w,b}$ as follows:

- Set $b := -1$.
- For each $i \in V$, set

$$w_i := w_{n+i} := \begin{cases} -1 & \text{if } i \in U, \\ +1 & \text{if } i \notin U. \end{cases}$$

We claim that $h_{w,b}$ misclassifies at most k examples. Consider any edge $\{i,j\} \in E$ and the corresponding negative examples $x_{i,j}$ and $x_{n+i,n+j}$. Then $\langle x_{i,j}, w \rangle = w_i + w_j \leq 0$ since at least one of i and j is in U. Therefore $h_{w,b}(x_{i,j}) = -1$. Similarly, $h_{w,b}(x_{n+i,n+j}) = -1$. So $h_{w,b}$ correctly classfies all negative examples. Now consider any $i \in V$ and the corresponding positive example $x_{i,n+i}$. Then $\langle x_{i,n+i}, w \rangle = w_i + w_{n+i} < 1$ iff $i \in U$. So $h_{w,b}(x_{i,n+i}) = -1$ iff $i \in U$. Since $|U| \leq k$, it follows that $h_{w,b}$ misclassifies at most k (positive) examples.

Coping with intractability. Similar forms of computational intractability hold for many other hypothesis classes. What can we do about this? Here are some possibilities (which may not be mutually exclusive nor individually sufficient):

- Make additional assumptions about the data distribution and/or the training data. Example: realizability.
- Use a different (and, typically, larger) hypothesis class $\mathcal{H}'$. Example: learn 3-term DNFs using the larger class of 3-CNFs.
- Use a different loss function that is easier to optimize. Example: exponential loss (as used in boosting).
- Change the learning model. Example: assume query access to target function.

2 Surrogate losses

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ be a dataset of labeled examples.[3] The training error rate of $\operatorname{sign} \circ g$ for $g \colon \mathcal{X} \to \mathbb{R}$ can be written as

$$\widehat{\operatorname{Risk}}(g) := \frac{1}{n} \sum_{i=1}^{n} \ell_{0/1}(y_i g(x_i)),$$

where $\ell_{0/1}(z) = \mathbb{1}\{z \leq 0\}$ is the zero-one loss function. If $\mathcal{X} = \mathbb{R}^d$ and we consider affine functions g (which have the form $g(x) = \langle x, w \rangle + b$ for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$), then minimizing $\widehat{\operatorname{Risk}}(g)$ is the NP-hard problem from the previous section.

However, sometimes we can replace $\ell_{0/1}$ with a different function that makes the optimization problem easier. For example, suppose $\ell_{0/1}(z)$ is replaced by $\phi(z) = |1 - z|$ (*absolute error*). Then the objective becomes an empirical average of absolute errors

$$\widehat{\operatorname{Risk}}_{\text{abs}}(g) = \frac{1}{n} \sum_{i=1}^{n} |1 - y_i g(x_i)| = \frac{1}{n} \sum_{i=1}^{n} |y_i - g(x_i)|.$$

If we restrict attention to affine functions g, the minimizing this objective can be formulated as a linear program and approximately solved in polynomial-time. The key is that $\phi(z) = |1 - z|$ is a convex function of z. Many other convex losses appear in the literature:

- $\phi(z) = (1 - z)^2$ (squared error, used in linear regression);
- $\phi(z) = \max\{0, 1 - z\}$ (hinge loss, used in soft-margin support vector machines);
- $\phi(z) = \exp(-z)$ (exponential loss, used in boosting);
- $\phi(z) = \log(1 + \exp(-z))$ (logistic loss, used in logistic regression).

These losses are called *convex surrogate losses*. Again, restricting attention to affine functions g, the minimization of the corresponding empirical risk with any of these loss functions is a convex optimization problem, which can be approximately solved using the Ellipsoid algorithm in polynomial-time. (In fact, it is also often sufficient to use simple algorithms like gradient descent!)

Using the theory of uniform convergence, we can relate $\widehat{\operatorname{Risk}}_{\text{abs}}(g)$ to the mean absolute error[4]

$$\operatorname{Risk}_{\text{abs}}(g) := \mathbb{E}[|1 - Y g(X)|]$$

when the training data is an iid sample from the distribution of (X, Y). This typically requires restricting the class $\mathcal{G}$ from which g is chosen and/or making assumptions on the data distribution.

Proposition 1. *Let (X, Y) be a random example taking values in $\mathbb{R}^d \times \{-1, 1\}$. Let $\mathcal{G} = \{x \mapsto \langle x, w \rangle : w \in \mathbb{R}^d, \|w\|_2 \leq 1\}$, and assume $\mathbb{E}[\|X\|_2^2] < \infty$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an iid sample from the distribution of (X, Y). Then*

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} \operatorname{Risk}_{\text{abs}}(g) - \widehat{\operatorname{Risk}}_{\text{abs}}(g)\right] \leq 2\sqrt{\frac{\mathbb{E}\left[\|X\|_2^2\right]}{n}}.$$

[3] In this and subsequent sections, it is important that we use $\{-1, 1\}$ as the label space instead of $\{0, 1\}$, because we will treat these labels as real numbers.

[4] For other surrogate losses ϕ, we generally call $\operatorname{Risk}_\phi(g) := \mathbb{E}[\phi(Y g(X))]$ the *surrogate risk* (or *ϕ-risk*) of g.

Proof. Let $Z_i := (X_i, Y_i)$ for all $i \in [n]$. We just need to compute the Rademacher complexity of the function class $\mathcal{F} := \{(x, y) \mapsto |1 - yg(x)| : g \in \mathcal{G}\}$. Since the function $t \mapsto |1 - yt|$ is 1-Lipschitz (for any $y \in \{-1, 1\}$), it follows by the Lipschitz contraction property of Rademacher averages that $\mathrm{Rad}_n(\mathcal{F}(Z_{1:n})) \le \mathrm{Rad}_n(\mathcal{G}(X_{1:n}))$. Moreover,

$$\mathrm{Rad}_n(\mathcal{G}(X_{1:n})) = \mathbb{E}_\sigma\left[\sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle X_i, w\rangle\right] = \mathbb{E}_\sigma\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2\right].$$

Therefore

$$\mathbb{E}\,\mathrm{Rad}_n(\mathcal{F}(Z_{1:n})) \le \mathbb{E}\,\mathbb{E}_\sigma\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2\right] \le \sqrt{\mathbb{E}\,\mathbb{E}_\sigma\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2^2\right]} = \sqrt{\frac{\mathbb{E}\left[\|X\|_2^2\right]}{n}}. \qquad \square$$

Using Proposition 1, we can show that $\hat{g} \in \arg\min_{g \in \mathcal{G}} \widehat{\mathrm{Risk}}_{\mathrm{abs}}(g)$ (for the function class $\mathcal{G}$ defined in Proposition 1) satisfies[5]

$$\mathbb{E}[\mathrm{Risk}_{\mathrm{abs}}(\hat{g})] \le \inf_{g \in \mathcal{G}} \mathrm{Risk}_{\mathrm{abs}}(g) + O\left(\sqrt{\frac{\mathbb{E}\|X\|_2^2}{n}}\right).$$

But why should we care about $\mathrm{Risk}_{\mathrm{abs}}(\hat{g})$ (or the risk corresponding to any other surrogate loss) if we ultimately care about classification? One reason is that small $\mathrm{Risk}_{\mathrm{abs}}(\hat{g})$ implies small $\mathrm{err}(\mathrm{sign} \circ \hat{g})$, because $\mathbb{1}\{z \le 0\} \le |1 - z|$.[6] In fact, we can improve this comparison by a factor of two if we allow thresholding $\hat{g}$ at a different value other than zero.[7]

Proposition 2. *For any $g \colon \mathcal{X} \to \mathbb{R}$, there exists $\theta \in \mathbb{R}$ such that*

$$\mathbb{E}[\mathbb{1}\{Y \,\mathrm{sign}(g(X) - \theta) \le 0\}] \le \frac{1}{2}\,\mathbb{E}[|Y - g(X)|].$$

Proof. WLOG assume $g(x) \in [-1, 1]$ for all $x \in \mathcal{X}$. Let $\theta \sim \mathrm{Uniform}([-1, 1])$. If $Y = +1$, then

$$\mathbb{E}_\theta[\mathbb{1}\{Y(g(X) - \theta) \le 0\}] = \Pr_\theta[\theta \ge g(X)] = \frac{1 - g(X)}{2} = \frac{|Y - g(X)|}{2}.$$

If $Y = -1$, then

$$\mathbb{E}_\theta[\mathbb{1}\{Y(g(X) - \theta) \le 0\}] = \Pr_\theta[\theta \le g(X)] = \frac{g(X) - (-1)}{2} = \frac{|g(X) - Y|}{2}.$$

So

$$\mathbb{E}_\theta[\mathbb{E}_{X,Y}[\mathbb{1}\{Y \,\mathrm{sign}(g(X) - \theta) \le 0\}]] = \mathbb{E}_{X,Y}\left[\frac{|Y - g(X)|}{2}\right]. \qquad \square$$

However, what is still not clear is why it is meaningful to compare $\mathrm{Risk}_{\mathrm{abs}}(\hat{g})$ to $\inf_{g \in \mathcal{G}} \mathrm{Risk}_{\mathrm{abs}}(g)$.

[5] Getting a corresponding "high probability" result is also easy if we assume the distribution of X is bounded.

[6] A similar comparison holds for (scalings of) the other surrogate losses mentioned above.

[7] Notice that if $g(x) \in \{-1, 1\}$ for all $x \in \mathcal{X}$, then $\mathrm{Risk}_{\mathrm{abs}}(g) = 2\,\mathrm{err}(\mathrm{sign} \circ g)$.

3 Excess risks

In statistical learning, we typically compare the error rate of the learned classifier to the best achievable error rate using hypotheses from a hypothesis class $\mathcal{H}$. So far in our development, we are comparing to the best achievable Risk_{abs} using functions from some function class $\mathcal{G}$. Can these comparisons be reconciled?

There is a general theory for relating such *excess risks* in the extreme case where $\mathcal{H}$ is all classifiers $h\colon \mathcal{X} \to \{-1, 1\}$, and $\mathcal{G}$ is all functions $g\colon \mathcal{X} \to \mathbb{R}$ (Zhang, 2004; Bartlett et al., 2006). Below, we instantiate it for the special case of absolute error.

What is the best possible error rate achievable by any classifier? Define

$$\eta(x) := \Pr(Y = 1 \mid X = x) \quad \text{for all } x \in \mathcal{X}.$$

Then, for any $x \in \mathcal{X}$ and $h(x) \in \{-1, 1\}$, we have

$$\mathbb{E}[\mathbb{1}\{h(X) \neq Y\} \mid X = x] = \eta(x)\mathbb{1}\{h(x) = -1\} + (1 - \eta(x))\mathbb{1}\{h(x) = +1\},$$

and this is minimized by

$$h(x) = \text{sign}(2\eta(x) - 1).$$

The classifier with this property is called the *Bayes (optimal) classifier*, and its error rate is called the *Bayes error rate* or *Bayes (classification) risk*.

What is the best possible mean absolute error achievable by any function? For any $x \in \mathcal{X}$ and $g(x) \in [-1, 1]$, we have

$$\begin{aligned}
\mathbb{E}[|Y - g(X)| \mid X = x] &= \eta(x)|1 - g(x)| + (1 - \eta(x))|-1 - g(x)| \\
&= \eta(x)(1 - g(x)) + (1 - \eta(x))(1 + g(x)) \\
&= g(x)(1 - 2\eta(x)),
\end{aligned}$$

which is minimized by

$$g(x) = \text{sign}(2\eta(x) - 1).$$

So the Bayes classifier also achieves the smallest possible mean absolute error, and its mean absolute error is exactly twice the Bayes error rate. (The absolute error is somewhat special in this regard.)

We can now relate the excess error rate to the excess mean absolute error. For any $g\colon \mathcal{X} \to \mathbb{R}$, there exists $\theta \in \mathbb{R}$ such that

$$\Pr[\text{sign}(g(X) - \theta) \neq Y] - \inf_{h^\star} \Pr[h^\star(X) \neq Y] \leq \frac{1}{2}\left(\mathbb{E}[|Y - g(X)|] - \inf_{g^\star} \mathbb{E}[|Y - g^\star(X)|]\right). \tag{1}$$

Similar relationships can be established for other surrogate losses. For example, for squared error, we have

$$\Pr[\text{sign}(g(X)) \neq Y] - \inf_{h^\star} \Pr[h^\star(X) \neq Y] \leq \sqrt{\mathbb{E}[(Y - g(X))^2] - \inf_{g^\star} \mathbb{E}[(Y - g^\star(X))^2]}, \tag{2}$$

where the minimum mean squared error is achieved by the conditional mean function

$$x \mapsto \mathbb{E}[Y \mid X = x] = 2\eta(x) - 1.$$

The main deficiency of this theory is that it is difficult to guarantee small excess surrogate risk without assuming that the $g^\star$ achieving the infimum in (1) or (2) is contained in (or well-approximated by some function in) the function class $\mathcal{G}$ used by the learner. It does not directly justify the use of surrogate loss functions if one only seeks to compare to the best error rate achievable within a given hypothesis class $\mathcal{H}$.

References

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

David S Johnson and Franco P Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

Richard M Karp. Reducibility among combinatorial problems. In R.E. Miller and J.W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.