EECS 553: Machine Learning (ECE) **University of Michigan**

Empirical Risk Minimization

Winter 2023 **Clayton Scott**

In these notes we will see that several algorithms already discussed fall under a common framework.

1 Performance Measures for Supervised Learning

Consider a supervised learning problem (either classification or regression) with a jointly distributed $(\boldsymbol{X}, Y)$. Let $\mathcal{Y}$ denote the output space (regression: $\mathcal{Y} = \mathbb{R}$, binary classification: $\mathcal{Y} = \{-1, +1\}$).

A *loss function*, or *loss* for short, is a function $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. Let $f : \mathbb{R}^d \to \mathbb{R}$. The quantity

$$R_L(f) = \mathbb{E}_{\boldsymbol{X},Y}[L(Y, f(\boldsymbol{X}))]$$

is referred to as the *L-risk* of f, or simply the *risk* if the loss is understood.

In regression, we think of f as a candidate prediction function, and in binary classification, we think of f as a candidate *decision function*, which is the function we take the sign of to get a classifier.

Example 1. In regression, f is a candidate for the predictor. If,

$$L(y, t) = (y - t)^2$$

then R_L is the *mean squared error*, and we refer to L as the *squared error* loss.

Example 2. If,

$$L(y, t) = |y - t|$$

then R_L is the *mean absolute error* and L is the *absolute deviation* loss. This is a loss for robust regression.

Example 3. In binary classification, f is a decision function, which defines a classifier by

$$y \longmapsto \text{sign}(f(\boldsymbol{x}))$$

where

$$\text{sign}(t) = \begin{cases} 1 & t \geq 0, \\ -1 & t < 0. \end{cases}$$

For example, the decision function $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$ defines a linear classifier. If,

$$L(y, t) = \mathbf{1}_{\{y \neq \text{sign}(t)\}}$$

then R_L is the probability of error, and L is called the *0-1* loss.

2 Empirical Risk Minimization

Given training data $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$, a natural way to learn a good f is to solve,

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i L(y_i, f(\boldsymbol{x}_i)) + \lambda \Omega(f),$$

where $\mathcal{F}$ is the set of candidate f functions (for example linear functions) and $\Omega(f)$ is the regularizer (for example, $\Omega(f) = \|\boldsymbol{w}\|^2$ if $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$). This problem is called *(penalized/regularized) empirical risk minimization (ERM)*. The quantity

$$\widehat{R}_L(f) = \frac{1}{n}\sum_i L(y_i, f(\boldsymbol{x}_i))$$

is called the *empirical L-risk* of f, or simply the *empirical risk* if the loss is understood. We have already seen ERM in regression in the form of least squares regression and robust regression.

A loss L is said to be *convex* if, for each $y \in \mathcal{Y}$, $L(y,t)$ is a convex function of t. For example, the squared error and absolute deviation losses are convex. As an exercise, you are asked to show that if a loss is convex, and f is linear, then the empirical risk is a convex function of $\boldsymbol{\theta} = [b\ \boldsymbol{w}^T]^T$.

In binary classification, unfortunately, the situation is not so nice. The problem is that the 0/1 loss

$$L(y,t) = \mathbf{1}_{\{y \neq \text{sign}(t)\}}$$

is not convex in t. In fact, it's not even differentiable so we can't even apply gradient descent.

3 Surrogate Losses

A surrogate loss is a loss that takes the place of another loss, usually because it has nicer computational properties such as convexity or differentiability. Some common surrogate losses for binary classification are

$$L(y,t) = \log(1 + e^{-yt}) \qquad \text{(logistic loss)}$$
$$L(y,t) = \max\{0, 1 - yt\} \qquad \text{(hinge loss)}.$$

Notice that both of these surrogate losses depend on y and t only through the product yt, which is sometimes called the *algebraic margin*, as opposed to the *geometric margin* of a separating hyperplane. Any loss that depends only on yt is called a *margin loss*.

The 0/1 loss is essentially a margin loss. In particular,

$$\mathbf{1}_{\{y \neq \text{sign}(t)\}} \leq \mathbf{1}_{\{yt \leq 0\}}$$

for all y and t, with equality holding if $t \neq 0$. The loss on the right hand side is a margin loss and is often also referred to as the 0/1 loss. The advantage of this version of the 0/1 loss is that it allows for a simple comparison with other margin losses. See Figure 1.

Many classification algorithms can be viewed as ERM for a certain L, $\mathcal{F}$, and Ω. In fact, we have already seen two of them.

3.1 Logistic Regression

As an exercise in the section on logistic regression, you are asked to show that

$$-\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} L(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$$

where $\ell(\boldsymbol{\theta})$ is the logistic regression log-likelihood, L is the logistic loss, and $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T\tilde{\boldsymbol{x}}$. This implies that logistic regression can be re-derived as a form of empirical risk minimization, where the 0-1 loss is replaced by the logistic loss.

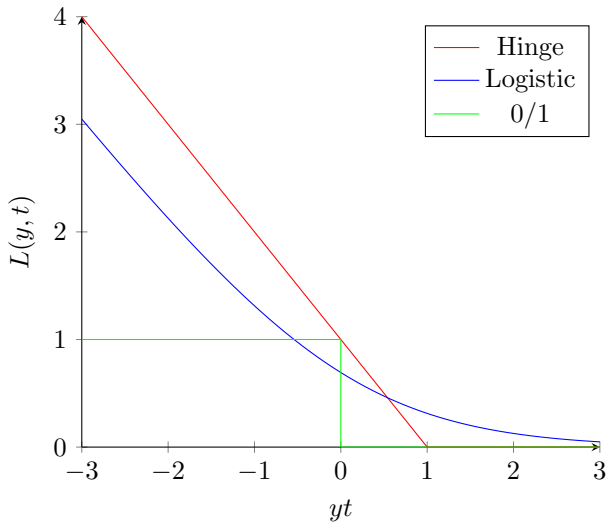

Figure 1: The 0-1 loss and two convex surrogates.

3.2 Optimal Soft-Margin Hyperplane

Recall the optimal soft-margin hyperplane solves:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \ & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i. \end{aligned} \tag{OSM}$$

If $\lambda = \frac{1}{C}$, then the solution $(\boldsymbol{w}^*, b^*)$ also solves

$$\min_{\boldsymbol{w},b} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) \tag{ERM-hinge}$$

which is regularized ERM with the hinge loss.

This can be seen by the following argument. First, scaling the objective function of (OSM) by $\frac{1}{C}$ doesn't change the solution. Next, the two constraints (for each i) can be merged into a single constraint :

$$\left.\begin{aligned} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \end{aligned}\right\} \iff \xi_i \geq \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)).$$

Thus far, we have shown that (OSM) reduces to

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \ & \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum \xi_i \\ \text{s.t.} \quad & \xi_i \geq \max\{0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\}. \end{aligned}$$

The solution to this last problem must satisfy

$$\xi_i = \max\{0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\} \quad \forall i,$$

otherwise we could reduce the objective without violating the constraints. This proves the claim.

4 Big Picture

Different choices of $L, \mathcal{F}, \Omega$ give rise to different methods. We will see several other examples. One advantage of the ERM framework is that it makes it easier to compare and contrast different methods. Another is that there are optimization strategies that can be used to solve large classes of ERM methods. Examples include (stochastic) gradient descent, (stochastic) subgradient methods, Newton-Raphson and quasi-Newton methods.

Exercises

1. Let $f : \mathbb{R} \to \mathbb{R}$ be convex.

   (a) (★) Show that the function
   $$g(\boldsymbol{\theta}) = f(\boldsymbol{r}^T\boldsymbol{\theta} + c)$$
   is convex in $\boldsymbol{\theta} \in \mathbb{R}^d$, where $\boldsymbol{r} \in \mathbb{R}^d$ and $c \in \mathbb{R}$.

   (b) (★) From the previous problem, deduce that if $L(y, t)$ is convex, then
   $$J(\boldsymbol{w}, b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \boldsymbol{w}^T\boldsymbol{x}_i + b)$$
   is convex in
   $$\boldsymbol{\theta} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}.$$

   (c) (★) Give a simple proof that the logistic regression empirical risk is convex in $\boldsymbol{\theta}$, without referencing the Hessian.

2. (★★) What loss is associated with the following quadratic program?
   $$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \ & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{aligned}$$