Empirical Risk Minimization

Overview

- Several methods studied so far, and others still to come, can be cast in a common framework.
- This general framework makes it possible to understand several different methods at once

Announcements

- HW2 due today, HW3 assigned

Outline

- Loss and Risk
- Empirical Risk Minimization
- Surrogate Losses

Loss and Risk

- Consider a supervised learning problem with jointly distributed $(\boldsymbol{X}, Y)$.
- Let $\mathcal{Y}$ denote the output space
  - Regression: $\mathbb{R}$
  - Binary classification: $\{-1, 1\}$
- A *loss* is a function $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$.

$L(y, t) =$ cost of predicting t when true output is y

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function. The *risk* of f is defined to be

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

Loss and Risk: Regression

- For regression problems, $f : \mathbb{R}^d \to \mathbb{R}$.
- If L is the *squared error* loss

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

  and

  $$R_L(f) = \mathbb{E}\left[(y - f(x))^2\right]$$

  is the mean squared error
- If L is the *absolute deviation* loss

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

  and

  $$R_L(f) = \mathbb{E}\left[|y - f(x)|\right]$$

  is the mean absolute error

Loss and Risk: Binary Classification

- For binary classification problems, f is called a *decision function* or *discriminant function.* The predicted label is

$$\text{sign}(f(x))$$

- For example, a linear classifier has $f(\boldsymbol{x}) =$ $w^T x + b$

- If L is the *0-1* loss

$$L(y,t) = \begin{cases} 1 & \text{if } y \neq \text{sign}(t) \\ 0 & \text{ow} \end{cases} = \mathbb{1}_{\{y \neq \text{sign}(t)\}}$$

then

$$R_L(f) = \mathbb{E}\left[\mathbb{1}_{\{y \neq \text{sign}(f(x))\}}\right]$$

is the

$$\Pr(y \overset{?}{\neq} \text{sign}(f(x)))$$

Poll

- Consider the following loss function for binary classification, where $\alpha \in (0, 1)$:

$$L_\alpha(y, t) := \begin{cases} \alpha, & y = 1, t < 0 \\ 1 - \alpha, & y = -1, t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- Consider a medical diagnosis application where it is more important to avoid a false negative (disease present ($Y = 1$) but classifier says it's not) than a false positive (disease not present ($Y = -1$) but classifier says it is).

- For such an application, the value of α should be chosen such that

  (A) $\alpha < 1/2$

  (B) $\alpha = 1/2$

  (C) $\alpha > 1/2$ ✓

  (D) $\alpha = \Pr(Y = 1)$

Empirical Risk Minimization

- Given: training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ for regression *or* binary classification.

- The quantity

$$\widehat{R}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) \approx \mathbb{E}\left[L(Y, f(X))\right]$$

  is called the *empirical risk* of f.

- *(Regularized) empirical risk minimization* learns f by solving

$$\min_{f \in \mathcal{F}} \hat{R}(f) + \lambda \Omega(f)$$

  where

  - $\mathcal{F}$ is the set of candidate f functions. *Example:* $f(x) = w^T x + b$
  - $\Omega(f)$ is the regularizer. *Example:* $\Omega(f) = \|w\|^2$
  - $\lambda \geq 0$, user-specified

ERM Examples: Regression

- Squared error loss

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \Omega(f)$$

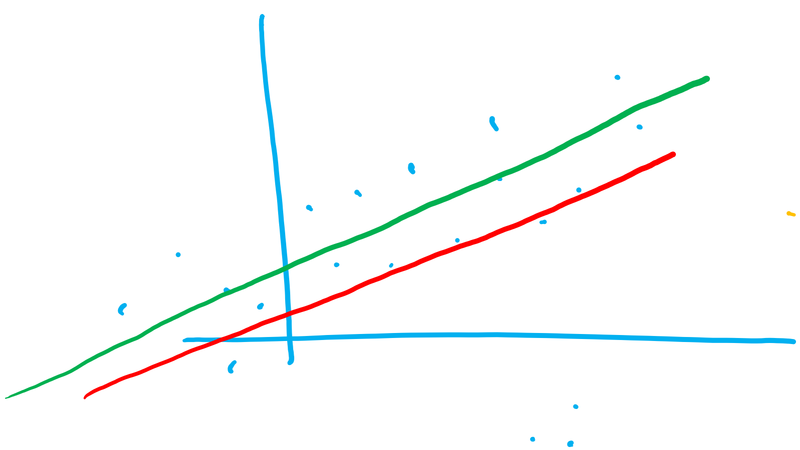
Linear reg: $\mathcal{F} = \{ f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R} \}$

$\Omega(f) = \|w\|^2, \quad \Omega(f) = \|w\|_1$

- Absolute deviation loss

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

$$\min \frac{1}{n} \sum |y_i - f(x_i)| + \lambda \Omega(f)$$



ERM Examples: Binary Classification

- 0-1 loss

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq \text{sign}(f(x_i))\}}$$

$$= \text{training error}$$

- Unfortunately, even for linear classifiers, this problems is *intractable*
- This motivates the use of *surrogate losses*

Surrogate Losses

- A surrogate loss is a loss that takes the place of another, usually because of nicer computational properties (convexity, differentiability).
- Some common surrogate losses for binary classification are
  - Logistic loss

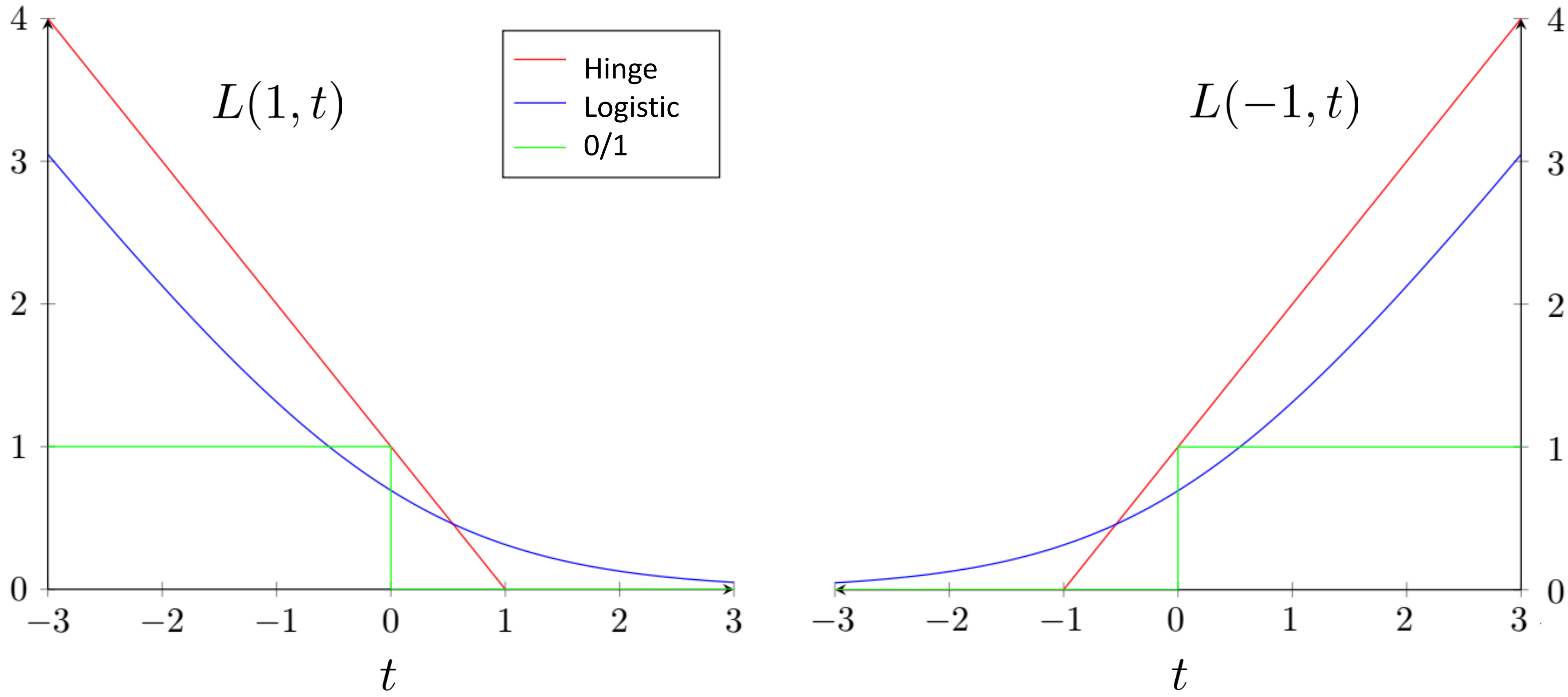
$$L(y,t) = \log(1 + \exp(-yt))$$

  - Hinge loss

$$L(y,t) = \max(0, 1 - yt)$$

Surrogate Losses

- These losses can be related graphically



Exercise

We say that a loss L is *convex* if, for each fixed y, $L(y,t)$ is a convex function of t.

1. Show that the logistic loss is convex. $L(y,t) = \log(1+\exp(-yt))$
2. Show that if L is a convex loss, then

$$\widehat{R}(\boldsymbol{w}, b) = \frac{1}{n} \sum_i L(y_i, \boldsymbol{w}^T \boldsymbol{x}_i + b)$$

is a convex function of $\boldsymbol{\theta} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}$.

Exercise

1. $\frac{\partial}{\partial t} L(y,t) = \frac{\partial}{\partial t} \log(1+e^{-yt})$

$$= \frac{-ye^{-yt}}{1+e^{-yt}}$$

$$\frac{\partial}{\partial t^2} L(y,t) = \frac{y^2e^{-yt} + (y^2-y)e^{-2yt}}{(1+e^{-yt})^2} > 0$$

Exercise

2. $\theta = \begin{bmatrix} b \\ w \end{bmatrix}, \tilde{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$ $\theta^T \tilde{x}_i = w^T x_i + b$

$\hat{R}(\theta) = \frac{1}{n} \sum L(y_i, \theta^T \tilde{x}_i)$

Want to show: $\forall \theta_1, \theta_2, \quad \forall \alpha \in [0,1]$

$\hat{R}(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha \hat{R}(\theta_1) + (1-\alpha)\hat{R}(\theta_2)$

Exercise

$$\hat{R}(\alpha\theta_1 + (1-\alpha)\theta_2) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, (\alpha\theta_1 + (1-\alpha)\theta_2)^T \tilde{x}_i\right)$$

$$= \frac{1}{n}\sum L\Big(y_i, \alpha \underbrace{\theta_1^T \tilde{x}_i}_{t_1} + (1-\alpha)\underbrace{\theta_2^T \tilde{x}_i}_{t_2}\Big)$$

$$\leq \frac{1}{n}\sum \left[\alpha L(y_i, \theta_1^T \tilde{x}_i) + (1-\alpha) L(y_i, \theta_2^T \tilde{x}_i)\right]$$

$$= \alpha \hat{R}(\theta_1) + (1-\alpha)\hat{R}(\theta_2).$$

Logistic Regression

- As an exercise it can be shown that

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

  where $\ell(\boldsymbol{\theta})$ is the logistic regression log-likelihood, L is the logistic loss, $y_i \in \{-1, 1\}$, and $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i$.

- This fact, combined with the previous excercise, give a second proof that the LR objective function is convex

- Take home message: Logistic regression can be derived from two different perspectives: maximum likelihood and ERM with logistic loss.

Optimal Soft-Margin Hyperplane

$\mathrm{sign}(w^T x + b)$

- Recall the optimal soft margin hyperplane solves:

$$\min_{\boldsymbol{w},b,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \qquad \text{(OSM)}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

$\left.\right\} \xi_i \geq \max(0, 1 - y_i(w^T x_i + b))$

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

$$\min_{w,b} \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\max(0, 1 - y_i(w^T x_i + b))}_{L(y_i, w^T x_i + b)}$$

- Proof: next slide

$L(y,t) = \max(0, 1 - yt)$

- Conclusion: The OSM hyperplane corresponds to regularized ERM with hinge loss

Optimal Soft-Margin Hyperplane

- The statement on the previous slide can be seen by scaling the objective function of (OSM) by $\frac{1}{C}$, which doesn't change the solution, and merging the constraints into a single constraint (for each i):

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

So (OSM) reduces to

$$\min_{w,b,\xi} \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum \xi_i$$

$$\xi_i \geq \max(0, 1 - y_i(w^T x_i + b))$$

Clearly the solution must satisfy

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b))$$

(otherwise we could decrease the objective), which reduces the problem to ERM with hinge loss. Can now eliminate ξ_i

Poll

- Consider the following loss functions:

  1. absolute deviation: $L(y,t) = |y - t|$
  2. sigmoid: $L(y,t) = \frac{1}{1+e^{yt}}$
  3. exponential: $L(y,t) = e^{-yt}$

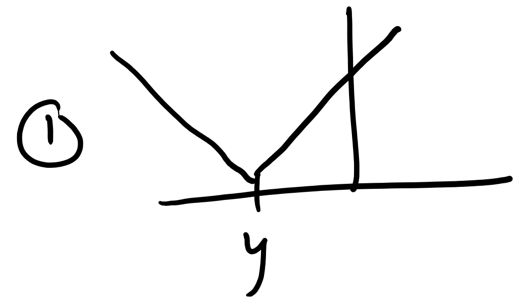
- Which of these loss functions are convex?

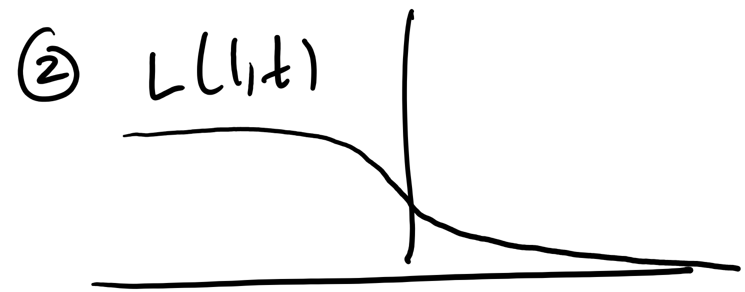
  (A) 1 and 2

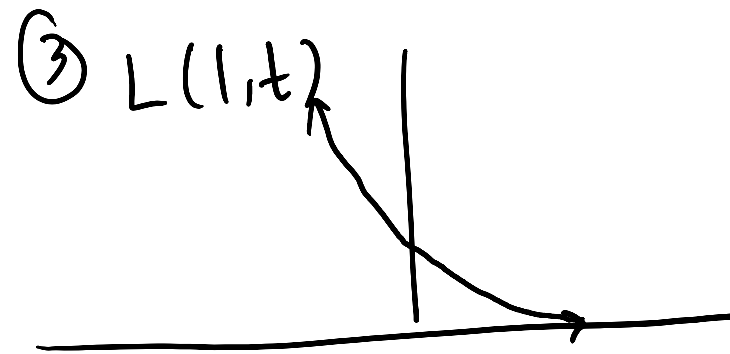
  (B) 1 and 3

  (C) 2 and 3

  (D) all of them







Big Picture

- *(Regularized) empirical risk minimization* learns f by solving

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

- Different choices of $L, \mathcal{F}, \Omega$ give rise to different methods.
- We will see several other examples including support vector machines, boosting, decision trees, neural networks, and sparse linear regression
- One advantage of this 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. Some will be covered in a future lecture.
- Also facilitates theoretical analysis