

COSC 594/650

Generalization

Dongsheng Ding

dongshed@utk.edu

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Optimization models across machine learning

Model	Prediction	Loss on one example
Linear regression	$\theta^\top x$	$(\theta^\top x - y)^2$
Logistic regression	$\text{sign}(\theta^\top x)$	$\log(1 + e^{\theta^\top x}) - y \theta^\top x$
Support vector machine	$\text{sign}(\theta^\top x + v)$	$\max\{0, 1 - y(\theta^\top x + v)\}$
Neural network	$h(x)$	$\ell((x, y), h)$

Only two columns differ: the prediction and the loss

One template

$$\underset{h \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h)$$

■ TWO MODELLING CHOICES

- ★ the hypothesis class $\mathcal{H}$ – linear, logistic, or a network
- ★ the loss ℓ – squared, logistic, or hinge

■ ONE COMPUTATIONAL PROBLEM

- ★ an average over the data

One template: a hypothesis class, a loss, and the average over the data

Expected risk / Test error

Data: $(x, y) \sim \mathbb{P}^*$, unknown $\mathbb{P}^*$

Loss: $\ell((x, y), h)$ for $h \in \mathcal{H}$

■ THE ERROR WE CARE ABOUT

$$L(h) = \mathbb{E}_{\mathbb{P}^*} [\ell((x, y), h)], \quad h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L(h)$$

★ $\mathbb{P}^*$ is unknown – $L(h)$ cannot be evaluated, for any h

What we want: the best hypothesis under the true distribution

Empirical risk / Training error

Training data: $(x_1, y_1), \dots, (x_m, y_m)$, drawn from $\mathbb{P}^\star$

■ THE ERROR WE CAN MEASURE

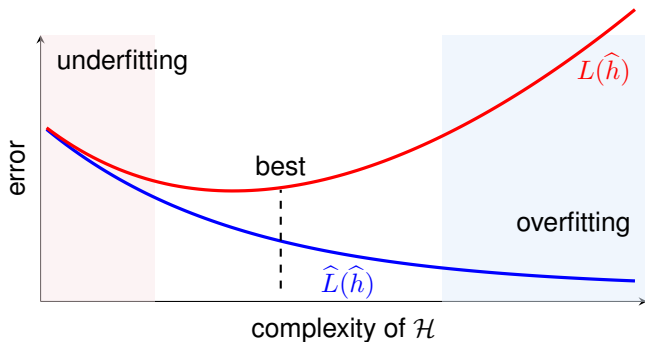
$$\widehat{L}(h) = \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h), \quad \widehat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{L}(h)$$

★ $\widehat{h}$ depends on the sample – $L(\widehat{h})$ is random

Empirical risk minimization: minimize $\widehat{L}$, and hope $L(\widehat{h})$ is close to $L(h^\star)$

Which hypothesis class?

Trade-off: training error against test error



The test error is U-shaped: more capacity is not better (classical pattern)

Underfitting versus overfitting

■ A RICHER $\mathcal{H}$ ALWAYS FITS THE DATA BETTER

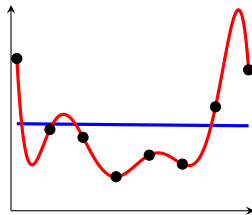
★ $\hat{L}(\hat{h})$ decreases with complexity – always

■ SIMPLE $\mathcal{H}$

★ $\hat{L}(\hat{h})$ and $L(\hat{h})$ are both large (underfitting)

■ RICH $\mathcal{H}$

★ $\hat{L}(\hat{h})$ is small and $L(\hat{h})$ is large (overfitting)



— Underfitting
— Overfitting

Choosing $\mathcal{H}$ is a trade-off, and not a maximization

A finite-class generalization bound

Finite class: $|\mathcal{H}| < \infty$

Zero-one loss: $\ell((x, y), h) = \mathbf{1}\{y \neq h(x)\}$

Realizable: $L(h^*) = 0$ – some $h \in \mathcal{H}$ is perfect

■ THEOREM

$$L(\hat{h}) - L(h^*) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{m}, \quad \text{w. p. } 1 - \delta$$

The gap is controlled by the sample size m and the size of $\mathcal{H}$

What the bound says

Bound: $L(\hat{h}) - L(h^*) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{m}$

■ SAMPLE COMPLEXITY

$$m \gtrsim \frac{\log |\mathcal{H}| + \log(1/\delta)}{\epsilon} \quad \text{samples suffice for accuracy } \epsilon$$

The sample size scales with $\log |\mathcal{H}|$, and not with $|\mathcal{H}|$

Proof #1: one fixed hypothesis

Sample: $(x_1, y_1), \dots, (x_m, y_m)$ i.i.d.

Fix $h \in \mathcal{H}$: before seeing the sample, with $L(h) > \epsilon > 0$

■ ONE CORRECT PREDICTION

$$\mathbb{P}(h(x_i) = y_i) = 1 - L(h)$$

■ ZERO TRAINING ERROR

$$\begin{aligned}\mathbb{P}(\hat{L}(h) = 0) &= \mathbb{P}(h(x_1) = y_1, \dots, h(x_m) = y_m) \\ &= \prod_{i=1}^m \mathbb{P}(h(x_i) = y_i) \\ &= (1 - L(h))^m \\ &\leq e^{-mL(h)} \leq e^{-m\epsilon}\end{aligned}$$

Proof #2: all hypotheses at once

One fixed hypothesis: $L(h) > \epsilon \implies \mathbb{P}(\hat{L}(h) = 0) \leq e^{-m\epsilon}$

■ UNION BOUND

$$\begin{aligned}\mathbb{P}(\exists h \in \mathcal{H} \mid L(h) > \epsilon, \hat{L}(h) = 0) &\leq \sum_{\substack{h \in \mathcal{H} \\ L(h) > \epsilon}} \mathbb{P}(\hat{L}(h) = 0) \\ &\leq |\mathcal{H}| e^{-m\epsilon}\end{aligned}$$

★ at most $|\mathcal{H}|$ terms, each at most $e^{-m\epsilon}$

$$\forall h \in \mathcal{H}, \quad \hat{L}(h) = 0 \implies L(h) \leq \epsilon, \quad \text{w. p. } 1 - |\mathcal{H}|e^{-m\epsilon}$$

Proof #3: apply the bound to ERM

■ REALIZABILITY AND EMPIRICAL MINIMIZATION

$$\begin{aligned} L(h^*) = 0 &\implies \hat{L}(h^*) = 0 \text{ almost surely} \\ \implies 0 &\leq \hat{L}(\hat{h}) = \min_{h \in \mathcal{H}} \hat{L}(h) \leq \hat{L}(h^*) = 0 \end{aligned}$$

■ APPLY SLIDE 10 TO $\hat{h} \in \mathcal{H}$

$$\mathbb{P}(L(\hat{h}) > \epsilon) \leq |\mathcal{H}|e^{-m\epsilon}$$

■ SET THE FAILURE PROBABILITY TO $\delta \in (0, 1)$

$$\epsilon = \frac{\log |\mathcal{H}| + \log(1/\delta)}{m} \implies |\mathcal{H}|e^{-m\epsilon} = \delta$$

$$L(\hat{h}) - L(h^*) = L(\hat{h}) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{m} \quad \text{w. p. } 1 - \delta$$

Regularization: an example

Least squares: $\mathcal{H} = \{ \theta^\top x \mid \theta \in \mathbb{R}^d \}$

■ SHRINK THE CLASS

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \sum_{i=1}^m (\theta^\top x_i - y_i)^2 \\ \text{subject to} & \|\theta\|^2 \leq \beta \end{array}$$

■ MOVE THE CONSTRAINT INTO THE OBJECTIVE

$$\underset{\theta}{\text{minimize}} \quad \sum_{i=1}^m (\theta^\top x_i - y_i)^2 + \lambda \|\theta\|^2$$

★ λ is a multiplier – one knob in place of β

Regularizer $\lambda \|\theta\|^2$ shrinks the class

Regularized empirical risk minimization

■ EMPIRICAL RISK MINIMIZATION

$$\underset{h \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h)$$

■ REGULARIZED EMPIRICAL RISK MINIMIZATION

$$\underset{h \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h) + \lambda \mathcal{R}(h)$$

- ★ Ridge, total variation, ℓ_1 – each prefers a different h
- ★ λ large – a very small class, and underfitting

Two knobs: the regularizer $\mathcal{R}$ and its weight λ

QUESTION: Which class, which regularizer, and how much?

Cross validation

Goal: avoid over-fitting and under-fitting

Strategy: keep part of the data out of the fit, and score $\hat{h}$ on it

■ THE TRAINING ERROR IS OPTIMISTIC

★ $\hat{h}$ was chosen to minimize $\hat{L} - \hat{L}(\hat{h})$ understates $L(\hat{h})$

■ HELD-OUT DATA IS HONEST

★ the validation samples took no part in choosing $\hat{h}$

Cross validation: an estimate of $L(\hat{h})$ from data the fit never saw

Two cross validation strategies

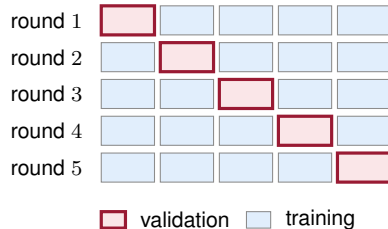
Partition: split the m samples into K parts of equal size

■ K -FOLD

- ★ train on $K - 1$ parts, score the last
- ★ repeat K times and average

■ LEAVE- p -OUT

- ★ validate on p of the m samples
- ★ $p = 1$ – leave-one-out



Slightly biased, and still very useful in practice

Computational structure of ERM

The optimization problem: $\underset{h \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h)$

■ A FINITE SUM

- ★ one gradient of $\hat{L}$ costs m gradients of ℓ
- ★ m can be millions – one iteration is already expensive

A finite sum is the structure we will exploit in optimization

Empirical risk minimization

$$\underset{h \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h)$$

- MODELS DIFFER ONLY IN THE HYPOTHESIS CLASS AND THE LOSS
- THE EMPIRICAL RISK IS A PROXY FOR THE EXPECTED RISK
- THE PRICE OF THE PROXY GROWS LIKE $\log |\mathcal{H}|$

Quantity	Meaning	Available to us
$L(h)$	expected risk	no, $\mathbb{P}^*$ is unknown
$\hat{L}(h)$	empirical risk	yes, from the data
$L(\hat{h}) - L(h^*)$	price of the proxy	$(\log \mathcal{H} + \log(1/\delta)) / m$

Questions?

(**Email: dongshed@utk.edu**)