

COSC 594/650

Empirical Risk Minimization

Dongsheng Ding

dongshed@utk.edu

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

Convergence results so far

■ UNCONSTRAINED

Method	Convex	Strongly convex
Gradient descent	$O(1/\epsilon)$	$O(\kappa \log(1/\epsilon))$
Nesterov acceleration	$O(1/\sqrt{\epsilon})$	$O(\sqrt{\kappa} \log(1/\epsilon))$

■ CONSTRAINED

Method	Work per iteration	Rate
Gradient projection	projection onto X	$(1 - 1/\kappa)^k$
Dual gradient ascent	exact inner solve	$O(1/k)$
Primal–dual	one gradient step	$O(1/\sqrt{k})$

Linear regression

Goal: predict Chloe's rating from her four friends' ratings

Ratings for Knoxville restaurants (1–5)

Input	Maya	Ethan	Sofia	Liam
x_1	4	3	5	4
x_2	5	5	4	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	5	4	5	3

$$\xrightarrow{y_i = \theta^\top x_i}$$

Output	Chloe
y_1	5
y_2	3
$\vdots$	$\vdots$
y_m	4

Linear model: $\theta \in \mathbb{R}^4$ are trust weights, one per friend

Linear regression as maximum likelihood

Noise model: $y_i = \theta^\top x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Consequence: $y_i | x_i \sim \mathcal{N}(\theta^\top x_i, \sigma^2)$

■ MAXIMUM LIKELIHOOD

$$\underset{\theta}{\text{maximize}} \quad \log \prod_{i=1}^m \mathbb{P}(y_i | x_i) \quad \Longleftrightarrow \quad \underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^m \left(\theta^\top x_i - y_i \right)^2$$

Least squares: $\ell((x, y), \theta) = (\theta^\top x - y)^2$

Logistic model

Same inputs: will Chloe like the restaurant? $y_i \in \{0, 1\}$

Logit model: $\log \frac{\mathbb{P}(y = 1 | x)}{\mathbb{P}(y = 0 | x)} = \theta^\top x$

■ A LINEAR SCORE BECOMES A PROBABILITY

$$p_i = \mathbb{P}(y_i = 1 | x_i) = \frac{e^{\theta^\top x_i}}{1 + e^{\theta^\top x_i}}, \quad \mathbb{P}(y_i = 0 | x_i) = 1 - p_i$$

★ $p_i > 1/2 \iff \theta^\top x_i > 0$ – predict the label 1

Classification: the same $\theta^\top x$, read as a probability

Logistic regression as maximum likelihood

One example: $\mathbb{P}(y_i | x_i) = (p_i)^{y_i} (1 - p_i)^{1-y_i}$

■ MAXIMUM LIKELIHOOD

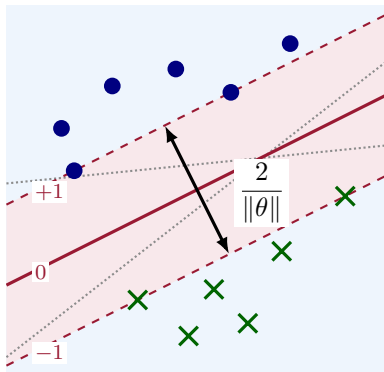
$$\underset{\theta}{\text{maximize}} \prod_{i=1}^m (p_i)^{y_i} (1 - p_i)^{1-y_i} \quad \Longleftrightarrow \quad \underset{\theta \in \mathbb{R}^d}{\text{minimize}} \sum_{i=1}^m \left[\log \left(1 + e^{\theta^\top x_i} \right) - y_i \theta^\top x_i \right]$$

- ★ $y_i = 1$ drives $\theta^\top x_i$ up
- ★ $y_i = 0$ drives $\theta^\top x_i$ down

Logistic loss: $\ell((x, y), \theta) = \log \left(1 + e^{\theta^\top x} \right) - y \theta^\top x$

Support vector machines

Labels: $y \in \{-1, +1\}$ (separable by $\theta^\top x + v = 0$)



values of $\theta^\top x + v$

Many separators: choose the one with the widest margin

Largest margin

Margin: the band between $\theta^\top x + v = \pm 1$ has width $\frac{2}{\|\theta\|}$

■ WIDEST BAND THAT STILL SEPARATES

$$\begin{array}{ll} \underset{\theta, v}{\text{maximize}} & \frac{2}{\|\theta\|} \\ \text{subject to} & y_i (\theta^\top x_i + v) \geq 1, \quad i = 1, \dots, m \end{array}$$

■ EQUIVALENT FORM

$$\begin{array}{ll} \underset{\theta, v}{\text{minimize}} & \|\theta\|^2 \\ \text{subject to} & y_i (\theta^\top x_i + v) \geq 1, \quad i = 1, \dots, m \end{array}$$

A convex quadratic program with linear constraints

Soft margin

Difficulty: no separating hyperplane may exist – infeasible

■ PAY FOR THE VIOLATIONS INSTEAD

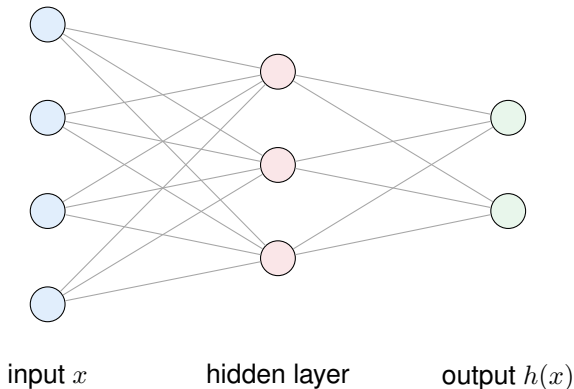
$$\underset{\theta, v}{\text{minimize}} \quad \lambda \|\theta\|^2 + \frac{1}{m} \sum_{i=1}^m \max \left\{ 0, 1 - y_i \left(\theta^\top x_i + v \right) \right\}$$

- ★ outside the margin – no penalty
- ★ inside the margin, or misclassified – linear penalty

Hinge loss: $\ell((x, y), \theta, v) = \max \{0, 1 - y(\theta^\top x + v)\}$

Neural networks

Layers: each layer takes a weighted combination of the previous one



A richer hypothesis class

Neural networks as an optimization problem

Hypothesis: $h \in \mathcal{H}$

■ THE PROBLEM

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

- ★ regression, logistic regression and SVM – one layer
- ★ $\mathcal{H}$ is no longer convex – the analysis is harder

The hypothesis class changed, and the shape of the problem did not

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

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 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 $L(h^\star)$

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

Quantity	Meaning	Available to us
$L(h)$	expected risk	no, $\mathbb{P}^*$ is unknown
$\hat{L}(h)$	empirical risk	yes, from the data

Questions?

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