Regression

Lecture 02 — CS 577 Deep Learning

Instructor: Yutong Wang

Computer Science
Illinois Institute of Technology

August 28, 2024

Recap

- Last time, samples, labels, and losses (among other things)
- This time, their origin story

Notations

Let $i = 1, \ldots, N$ (the sample index)

- Training samples $\mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$

Notations

Let $i = 1, \dots, N$ (the sample index)

- Training samples $\mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$
- One-dimensional samples (unbolded) are denoted like this

$$x^{(1)}, \dots, x^{(N)}$$

Notations

Let $i = 1, \ldots, N$ (the sample index)

- Training samples $\mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$
- One-dimensional samples (unbolded) are denoted like this

$$x^{(1)}, \ldots, x^{(N)}$$

- d-dimensional samples are denoted like this

$$\mathbf{x}^{(1)} = \begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_d^{(1)} \end{bmatrix}, \ldots, \mathbf{x}^{(N)} = \begin{bmatrix} x_1^{(N)} \\ \vdots \\ x_d^{(N)} \end{bmatrix}$$

Notations

Let $i = 1, \ldots, N$ (the sample index)

- Training samples $\mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$
- One-dimensional samples (unbolded) are denoted like this

$$x^{(1)}, \ldots, x^{(N)}$$

- d-dimensional samples are denoted like this

$$\mathbf{x}^{(1)} = \begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_d^{(1)} \end{bmatrix}, \ldots, \mathbf{x}^{(N)} = \begin{bmatrix} x_1^{(N)} \\ \vdots \\ x_d^{(N)} \end{bmatrix}$$

- Labels $y^{(i)} \in \mathcal{Y} = \mathbb{R}$ for regression

Notations (a bit different from Lec 1)

- Training samples $(\mathbf{x}^{(i)}, y^{(i)})$
- A test sample $(\mathbf{x}^{(\text{test})}, y^{(\text{test})})$
- A generic sample $(\mathbf{x}, y)$
  (A place holder. Can substitute with either a train or a test sample.)

Probability

- Where does the samples and label come from?

Probability

- Where does the samples and label come from?
- *Joint probability*: $p_{\text{data}}(\mathbf{x}, y)$ joint probability of $\mathbf{x}$ and y

Probability

- Where does the samples and label come from?
- *Joint probability*: $p_{\text{data}}(\mathbf{x}, y)$ joint probability of $\mathbf{x}$ and y
- How does the label depend on the label?

Probability

- Where does the samples and label come from?
- *Joint probability*: $p_{\text{data}}(\mathbf{x}, y)$ joint probability of $\mathbf{x}$ and y
- How does the label depend on the label?
- *Conditional probability*: $p_{\text{data}}(y \mid \mathbf{x})$ probability of y given $\mathbf{x}$

Probability

- Where does the samples and label come from?
- *Joint probability*: $p_{\text{data}}(\mathbf{x}, y)$ joint probability of $\mathbf{x}$ and y
- How does the label depend on the label?
- *Conditional probability*: $p_{\text{data}}(y \mid \mathbf{x})$ probability of y given $\mathbf{x}$
- Where does probabilities come from?

Generating synthetic datasets demo

Notations (a bit different from Lec 1)

- Model params $\boldsymbol{\theta} \in \Theta$

Notations (a bit different from Lec 1)

- Model params $\boldsymbol{\theta} \in \Theta$
- The model $f(\cdot\,; \boldsymbol{\theta})$

Notations (a bit different from Lec 1)

- Model params $\boldsymbol{\theta} \in \Theta$
- The model $f(\cdot\,; \boldsymbol{\theta})$
- The model's prediction $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$ at a point $\mathbf{x}$

Notations (a bit different from Lec 1)

- Model params $\boldsymbol{\theta} \in \Theta$
- The model $f(\cdot\,; \boldsymbol{\theta})$
- The model's prediction $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$ at a point $\mathbf{x}$
- *Model-specified probability* $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$

Notations (a bit different from Lec 1)

- Model params $\boldsymbol{\theta} \in \Theta$
- The model $f(\cdot\,; \boldsymbol{\theta})$
- The model's prediction $\hat{y} = f(\mathbf{x}; \boldsymbol{\theta})$ at a point $\mathbf{x}$
- *Model-specified probability* $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$
- What exactly is $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$?

Gaussian/normal distribution

- Gaussian distribution with mean μ and variance σ^2

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2) \qquad (\epsilon \text{ for "error"})$$

- The probability density function (PDF)

$$\mathcal{N}(\epsilon; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\epsilon - \mu)^2}{\sigma^2}\right)$$

What exactly is $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$?

- Model params in linear regression: $\boldsymbol{\theta} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$

What exactly is $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$?

- Model params in linear regression: $\boldsymbol{\theta} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$
- The model: $f(\mathbf{x}; \boldsymbol{\theta}) := \mathbf{w}^\top \mathbf{x} + b$

What exactly is $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$?

- Model params in linear regression: $\boldsymbol{\theta} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$
- The model: $f(\mathbf{x}; \boldsymbol{\theta}) := \mathbf{w}^\top \mathbf{x} + b$
- Gaussian/normally distributed noise: there exists $\sigma^2 > 0$ such that

$$y \mid \mathbf{x} \sim \mathcal{N}(f(\mathbf{x}; \boldsymbol{\theta}), \sigma^2)$$

What exactly is $p_{\text{model}}(y \mid \mathbf{x}; \boldsymbol{\theta})$?

- Model params in linear regression: $\boldsymbol{\theta} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$
- The model: $f(\mathbf{x}; \boldsymbol{\theta}) := \mathbf{w}^\top \mathbf{x} + b$
- Gaussian/normally distributed noise: there exists $\sigma^2 > 0$ such that

$$y \mid \mathbf{x} \sim \mathcal{N}(f(\mathbf{x}; \boldsymbol{\theta}), \sigma^2)$$

- *The "well-specified" assumption:* there exists $\boldsymbol{\theta}^* \in \Theta$ such that $p_{\text{data}}(\cdot \mid \cdot) = p_{\text{model}}(\cdot \mid \cdot; \boldsymbol{\theta}^*)$

Maximum likelihood

Likelihood:

$$\prod_{i=1}^{N} p(y^{(i)} \mid \mathbf{x}^{(i)}; \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y^{(i)} - f(\mathbf{x}^{(i)}; \boldsymbol{\theta}))^2}{\sigma^2}\right)$$

Max. likelihood & empirical risk minimization

- Squared error loss $L(\hat{y}, y) = (\hat{y} - y)^2$

Max. likelihood & empirical risk minimization

- Squared error loss $L(\hat{y}, y) = (\hat{y} - y)^2$
- Empirical risk minimization (ERM) (more commonly training error or the training mean squared error (MSE))

$$\min_{\boldsymbol{\theta} \in \Theta} \quad J(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), y^{(i)})$$

Linear regression in 1-D

$$\boldsymbol{\theta} = \begin{bmatrix} w \\ b \end{bmatrix}$$

$$J\left(\begin{bmatrix} w \\ b \end{bmatrix}\right) := \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (wx^{(i)} + b))^2$$

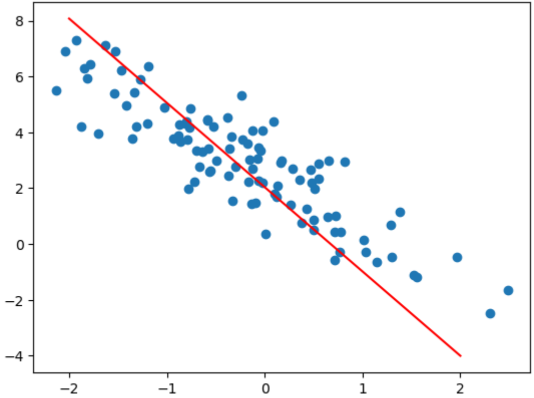
$$\frac{\partial J}{\partial w} =$$

$$\frac{\partial J}{\partial b} =$$

Gradient descent:

Exercise 1

Verifying your solution to Exercise 1.c



Linear regression in 2-D

$$\boldsymbol{\theta} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} \qquad \tilde{\mathbf{X}} = \begin{bmatrix} \left(\mathbf{x}^{(1)}\right)^\top & 1 \\ \vdots & \vdots \\ \left(\mathbf{x}^{(N)}\right)^\top & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

$$J\left(\boldsymbol{\theta}\right) := \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (y^{(i)} - (\mathbf{w}^\top \mathbf{x}^{(i)} + b))^2 = \tfrac{1}{N} \|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\theta}\|^2$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \tfrac{2}{N}(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}^\top \mathbf{y})$$

$$\arg\min_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \boldsymbol{\theta}_\star = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}\mathbf{y}$$

Gradient descent:

Train/population-level/test

Training error/risk

$$J(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), y^{(i)})$$

(Population) Risk

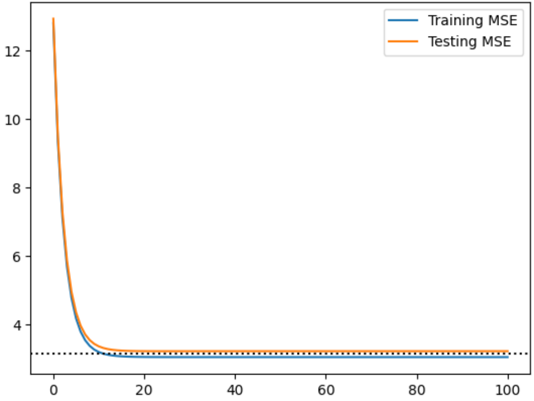
$$J^*(\boldsymbol{\theta}) := \mathbb{E}_{(\mathbf{x},y)\sim p_{\text{data}}} L(f(\mathbf{x}; \boldsymbol{\theta}), y)$$

Test error/risk

$$J^{(\text{test})}(\boldsymbol{\theta}) := \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} L(f(\mathbf{x}^{(\text{test},i)}; \boldsymbol{\theta}), y^{(\text{test},i)})$$

Exercise 2

Verifying your solution to Exercise 2.d

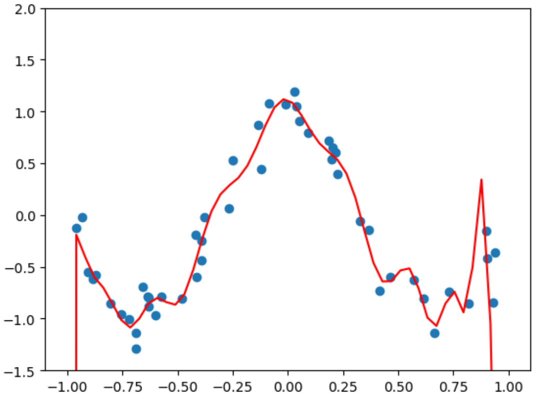


What I want to fit nonlinear functions?

1. Feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$
2. Transform the data $\tilde{\mathbf{x}} = \phi(\mathbf{x})$

Exercise 3

Exercise 3



Activation function

Rectified linear unit or ReLU

$$\mathrm{relu}(z) := \max\{0, z\}$$

A simple "bias-only" neural network

- Model params $\boldsymbol{\theta} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \in \mathbb{R}^2$
- The model $f(\cdot\,; \boldsymbol{\theta}) : \mathbb{R} \to \mathbb{R}$

$$f(x; \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}) = \mathrm{relu}(x + b_1) - \mathrm{relu}(x + b_2)$$

References I