Linear Regression

CSE512 – Machine Learning, Spring 2018, Stony Brook University

Instructor: Minh Hoai Nguyen (minhhoai@cs.stonybrook.edu)

Date: 29-Jan-2018

Last Lecture: Parameter Estimation

We observe some data $X_1, \cdots, X_n$

Make an assumption $X \sim P(X|\theta)$

Choose a loss function:

MLE: $$\hat{\theta} = \operatorname*{argmax}_{\theta} P(\mathcal{D}|\theta)$$

MAP: $$\hat{\theta} = \operatorname*{argmax}_{\theta} P(\theta|\mathcal{D})$$

Optimization:

E.g., Take derivative and set to 0

Outline

- Supervised Learning
- Linear Regression

Supervised Learning

Input (features) $\longrightarrow$ Output (targets, labels)

Outside Temp, #people in building	Energy consumption
GRE scores, LORs, GPA	Job/No-job

Energy Consumption Prediction

Outside Temp	#people	Energy consumption
72	4	10
32	3	50
50	10	75
60	7	56

Problem formulation

Given the training set: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$

Task: find a predictor function: $y = f(\mathbf{x})$

General Machine Learning Approach

- Using domain/prior knowledge, assume a model for the predictor
  - E.g., linear model, quadratic model

- The functional form of the model is fixed, but it has unknown parameters $y = f(\mathbf{x}; \Theta)$

- Use training data to <u>learn</u> the model's parameters

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n) \longrightarrow \Theta$$

- Use the <u>learned</u> parameters for prediction: $\mathbf{x}, \Theta \longrightarrow y$

Outline

- Supervised Learning
- Linear Regression

Linear Regression

Assume the output is a linear function of input features

$$\hat{y} = f(\mathbf{x}; \Theta) = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d + \theta_{d+1}$$

Learn the parameters so that

$$y_i \approx \hat{y}_i = \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_d x_{id} + \theta_{d+1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1d} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{d+1} \end{bmatrix}$$

Minimize the prediction loss

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$L(\boldsymbol{\theta}) = ||\mathbf{y} - \bar{\mathbf{X}}^T \boldsymbol{\theta}||^2$$

$$\bar{\mathbf{x}}_i = [\mathbf{x}_i; 1]$$

$$\hat{y}_i = \bar{\mathbf{x}}_i^T \boldsymbol{\theta}$$

Optimization

Minimize $$L(\boldsymbol{\theta}) = ||\mathbf{y} - \bar{\mathbf{X}}^T\boldsymbol{\theta}||^2$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}} =$$

Closed-form solution

But Why Sum of Squared Errors?

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Why not sum absolute errors or squared squared errors?

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^4$$

Because: Gaussian noise assumption!

Outline

- Supervised Learning
- Linear Regression
- Linear Regression and Gaussian Connection

Linear Regression: Gaussian Noise

Linear model with additive Gaussian noise

$$y_i = \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i^T \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$$

The Likelihood of Training Data

Linear model with additive Gaussian noise

$$y_i = \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i^T \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$$

Conditional likelihood

$$P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \sigma)$$

$$= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}^T\boldsymbol{\theta}||^2}$$

Maximizing the Data Likelihood

Find the model parameters to maximize the conditional likelihood

$$\hat{\boldsymbol{\theta}}, \hat{\sigma} = \underset{\boldsymbol{\theta}, \sigma}{\operatorname{argmax}} \, P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma)$$

Equivalent to maximizing the conditional log-likelihood

$$\hat{\boldsymbol{\theta}}, \hat{\sigma} = \underset{\boldsymbol{\theta}, \sigma}{\operatorname{argmax}} \log(P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma))$$

The log-likelihood:

$$\log(P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma)) = -\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}^T\boldsymbol{\theta}||^2 - n\log(\sigma) - \frac{n}{2}\log(2\pi)$$

Maximum Likelihood Estimate (MLE)

MLE estimate for $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} -||\mathbf{y} - \mathbf{X}^T\boldsymbol{\theta}||^2 = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$$

MLE estimate for σ

$$\frac{\partial L}{\partial \sigma} =$$

Making Prediction

$$\hat{y} = \mathcal{N}(\mathbf{x}_{new}^T \boldsymbol{\theta}, \sigma^2)$$

MAP for Linear Regression?

$P(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \lambda^2\mathbf{I})$ Equivalently $P(\theta_i) \sim \mathcal{N}(0, \lambda^2)$

The optimization problem correspond to

Minimize $L(\boldsymbol{\theta}) = ||\mathbf{y} - \mathbf{X}^T\boldsymbol{\theta}||^2 + \gamma||\boldsymbol{\theta}||^2$

This is called Ridge Regression

Things You Need to Know

- ML is to learn parameters of a function
- Least-squares solution
- Connection between Linear Regression and Gaussian Noise