Regression

INSTRUCTOR: HONGJIE CHEN

JUNE 6TH 2022

Extend Y from Discrete to Continuous

- Classification:

  $P(Y \mid X)$ where Y is discrete

- Regression:

  $P(Y \mid X)$ where Y is continuous

- For example…

Problem setting

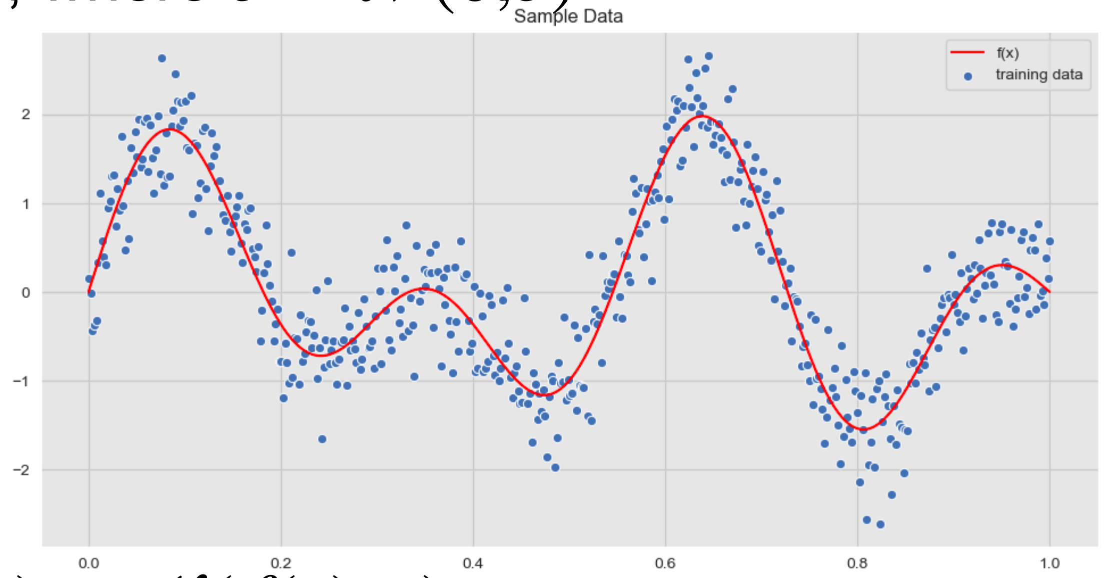
- Learn a function $f : X \rightarrow Y, Y \in \mathscr{R}$

- Approach:
  - Choose some parameterized form for $P(Y|X, \theta)$ where θ is called a parameter vector.
  - Estimate θ via MLE or MAP

VIRGINIA TECH

Parameterized Form for $P(Y|X,\theta)$

- Assume Y is some deterministic $f(X)$, plus random noise
  - $y = f(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0,\sigma)$



- Therefore $p(y|x) = \mathcal{N}(f(x), \epsilon)$
- Expectation: $\mathbb{E}[Y] = f(X)$

Figure credit: [link]

VIRGINIA TECH

Where we were last time

- Review the logistic regression model
- Distinguish generative models and discriminative models
  - Generative models describe how data are generated
  - Discriminative models distinguish data by boundaries
- Regression model

- HW1, HW2, HW3

Linear Regression

- $p(y \mid x) = \mathcal{N}(f(x), \epsilon)$

- Assume $f(x)$ is a linear function

  $p(y \mid x) = \mathcal{N}(w_1 x + w_0, \sigma)$

  $\mathbb{E}(y \mid x) = w_1 x + w_0$

- Make parameters explicit
  - $W = [w_1, w_0]$
  - $p(y \mid x, W) = \mathcal{N}(w_1 x + w_0, \sigma)$

VIRGINIA TECH

Training Linear Regression Model

- MCLE

$$W_{MCLE} = \text{argmax}_W \prod_l P(Y^l | X^l, W)$$

$$W_{MCLE} = \text{argmax}_W \sum_l \ln P(y^l | x^l, W)$$ **Log likelihood**

$$\text{Where } P(y | x, W) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \mu = f(x, W)$$

Simplify

$$P(y|x,W) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(x,W))^2}{2\sigma^2}}$$

$$W_{MCLE} = \text{argmax}_W \sum_l \ln P(y^l|x^l,W)$$

$$= \text{argmax}_W \sum_l \boxed{\ln \frac{1}{\sqrt{2\pi\sigma^2}}} + \left(-\frac{(y^l - f(x^l,W))^2}{2\sigma^2}\right)$$

Constant

$$= \text{argmax}_W - \frac{1}{2\sigma^2} \sum_l (y^l - f(x^l,W))^2$$

$$= \text{argmin}_W \frac{1}{2\sigma^2} \sum_l (y^l - f(x^l,W))^2$$

$$= \text{argmin}_W \sum_l (y^l - (w_1 x^l + w_0))^2$$

Gradient Descent

- $W_{MCLE} = \text{argmin}_W \boxed{\sum_l (y^l - f(x^l, W))^2}$ E
- Gradient $\nabla E(\vec{w}) = [\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n}]$, is a vector
  - Training rule: $\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta \nabla E(\vec{w})$
  - View from one feature dimension $\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$

VIRGINIA TECH

Calculate Derivative

- $E = \sum_l (y^l - (w_1 x^l + w_0))^2$
- $\frac{\partial E}{\partial w_0} =$
- $\frac{\partial E}{\partial w_1} =$

Calculate Derivative

- $E = \sum_l (y^l - (w_1 x^l + w_0))^2$

- $\frac{\partial E}{\partial w_0} = -2 \sum_l (y^l - (w_1 x^l + w_0))$

- $\frac{\partial E}{\partial w_1} = -2 \sum_l (y^l - (w_1 x^l + w_0))x^l$

Vectorize $X = [x_1, x_2, \ldots, x_n]$

$$f(x) = w_0 + \sum_{j=1}^{n} w_j x_j$$

$$\overrightarrow{w} = [w_0, w_1, \ldots, w_n]$$

$$\frac{\partial E}{\partial w_i} = -2 \sum_{l} (y^l - (w_0 + \sum_{j=1}^{n} w_j x_j)) x_i^l$$

$$\overrightarrow{w}^{(t+1)} \leftarrow \overrightarrow{w}^{(t)} - \eta \nabla E(\overrightarrow{w})$$

$$w_i \leftarrow w_i + 2\eta \sum_{l} (y^l - (w_0 + \sum_{j=1}^{n} w_j x_j)) x_i^l$$

Gradient Descent Algorithm

- Repeat until $\Delta w < \epsilon$
  - For all dimension i,

$$w_i \leftarrow w_i + 2\eta \sum_l (y^l - (w_0 + \sum_{j=1}^{n} w_j x_j)) x_i^l$$

  - Assume $x_0 = 1$ to incoporate w_0

MAP

- $$W_{MAP} = \text{argmin}_W \underbrace{\boxed{(-c\sum_i w_i^2)}}_{\text{Regularization}} + \sum_l (y^l - f(x^l, W))^2$$

- Remember advantages of regularization?
- Demo: https://lukaszkujawa.github.io/gradient-descent.html

- Question: must f a linear function to x?

VIRGINIA TECH