Deep Learning Reviews

Ao Zhang

July 2, 2021

1 Some Definitions

- **Supervised Learning**: Given $\mathcal{D} = \{(x_i, y_i) : i = 1, 2, ..., n\}$, develop a program that predicts Y from X, or finds how Y depends on X.
- **Unsupervised Learning**: Given $\mathcal{D} = \{x_i : i = 1, 2, ..., n\}$, develop a program that finds the structure in X, or generates an (new) X that conforms to the structure.
- **Reinforcement Learning**: Suppose there is an "environment" which can interact with an agent by changing the "state" and generating a reward for the agent. Develop an agent program that maximize some accumulated reward it receives.
- **Model**: A restricted family $\mathcal{H}$ of hypotheses.

$$Y = f(X; \hat{\theta}) \tag{1}$$

- **Parametric Models**: Models have a fixed number of parameters, independent of sample size.
- **Non-Parametric Models**: The number of parameters increases with sample size. (usually not considered.)
- **Loss Functions**: A model is usually characterized by Loss Function $\mathcal{L}(\theta)$ over the space Θ of model parameters θ. An example using Mean Square Error (MSE) is shown as below.

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \mathcal{L}(\theta) = ||Y - X\theta||^2 \tag{2}$$

- **Gradient Descent (GD)**: Based on Equation 2,

$$\theta^{new} = \theta^{old} - \lambda \frac{d\mathcal{L}}{d\theta}(\theta^{old}) \tag{3}$$

$$= \theta^{old} + \lambda \frac{1}{N} \sum_{i=1}^{N} 2(y_i - \theta^{old} x_i) x_i \tag{4}$$

- **Stochastic Gradient Descent (SGD)**: Based on Equation 2,

$$\theta^{new} = \theta^{old} + \lambda 2(y_i - \theta^{old} x_i) x_i \tag{5}$$

- **Mini-Batched SGD**: Based on Equation 2,

$$\theta^{new} = \theta^{old} + \lambda \frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} 2(y - \theta^{old} x) x \tag{6}$$

2 Activation Functions

- **Sigmoid Function**:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{7}$$

$$\sigma'(x) = \sigma(x)(1-\sigma(x)) \tag{8}$$

- **Softmax Function**:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}, \quad x_i \in \mathbb{R}^K \tag{9}$$

3 Normalization

4 Regularization

First, we need to review what is **Overfitting** and what is **Underfitting**. Some definitions need to be introduced at the beginning.

- **In-Sample Error**: Also known as **training error**, notation E_{in}.
- **Out-Sample Error**: Also known as **testing error**, notation E_{out}.
- **Generalization Gap**: Notation E_{gen}

$$E_{gen} = E_{out} - E_{in} \tag{10}$$

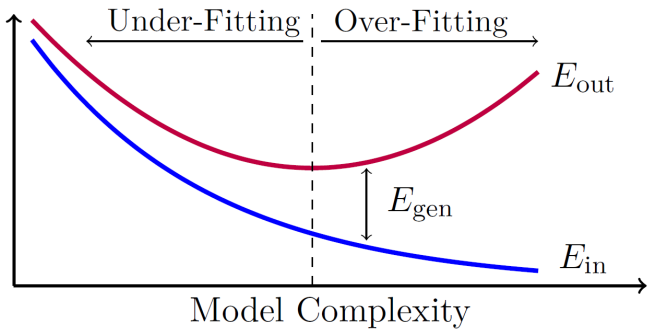

Figure 1: Overfitting and Underfitting

Regularization: It refers to techniques that reduce over-fitting when learning with complex models.

$$\mathcal{L}_{Reg}(\theta) = \mathcal{L} + \Omega(\theta) \tag{11}$$

Popular regularziation methods can be listed as:

- **Early Stop**
- **Dropout**
- **L1 Regularizer**: $\Omega(\theta) := \lambda_{Reg}|\theta|_2$
- **L2 Regularizer**: $\Omega(\theta) := \lambda_{Reg}||\theta||_2^2$

5 Loss Functions

- **Mean Square Error**:

$$\hat{\theta} = \arg\min_{\theta}(Y - X\theta)^2 \tag{12}$$

- **Cross Entropy**: Minimize Cross Entropy = Maximize Likelihood

$$CE(\tilde{p}; p) = -\sum_{y \in Y} \tilde{p}(y) \log p(y) \tag{13}$$

$$= -\mathbb{1}_{y_i=1} \log p_{Y|X}(1|x_i) - \mathbb{1}_{y_i=0} \log p_{Y|X}(0|x_i) \tag{14}$$

- **Focal Loss function**
- **IoU Loss function**
- **Dice Loss function**

6 Optimization Methods

7 Metrics

References