

## Regularization

### 1. Overfitting

- Underfitting, or high bias, is when the form of our hypothesis function  $h$  maps poorly to the trend of the data.
- Overfitting, or high variance, is caused by a hypothesis function that fits the available data but does not generalize well to predict new data.

#### Addressing

- Reduce the number of features (manually select, model selection).
- Regularization (keep all the features, but reduce the magnitude of parameters).

### 2. Regularized Linear Regression

$$\min_{\theta} \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right], \text{ where } h_{\theta}(x) = \theta^T x$$

- Normal equation

$$\theta = (X^T X + \lambda \cdot L)^{-1} X^T y \text{ where } L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

Proof.

$$\begin{cases} \frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{k=1}^m (\theta^T x^{(k)} - y^{(k)}) x_j^{(k)} & (j = 0) \\ \frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{k=1}^m (\theta^T x^{(k)} - y^{(k)}) x_j^{(k)} + \frac{\lambda}{m} \theta_j & (j \in N^+) \end{cases} \Rightarrow \nabla_{\theta} J(\theta) = \frac{1}{m} (X^T X \theta - X^T y) + \frac{\lambda}{m} L \theta$$

### 3. Regularized Logistic Regression

$$\min_{\theta} \left[ -\frac{1}{m} \sum_{i=1}^m (y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))) + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2 \right]$$

### 4. MLE & MAP

#### Preliminaries

- Assume data are generated via  $d \sim p(d; \theta)$
- $D = \{d^{(i)}\}_{i=1,2,\dots,m}$ , where  $d^{(i)}$  is i.i.d. (independent of others and same distribution).
- Goal: Estimate parameter  $\theta$  that best models the data.

### Maximum Likelihood Estimation (MLE)

- Likelihood:  $L(\theta) = p(D; \theta) = \prod_{i=1}^m p(d^{(i)}; \theta)$
- MLE typically maximizes the log-likelihood  $l(\theta)$ .
- $\theta_{MLE} = \arg \max_{\theta} \sum_{i=1}^m \log p(d^{(i)}; \theta)$

### Maximum-a-Posteriori Estimation (MAP)

- Posterior probability of  $\theta$  is  $p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$
- $p(\theta)$  is prior probability of  $\theta$ , where  $p(D)$  is probability of the data.
- MAP usually maximizes the log of the posteriori probability
- $\theta_{MAP} = \arg \max_{\theta} (\log p(\theta) + \sum_{i=1}^m \log p(d^{(i)}|\theta))$

### Linear Regression

#### 1. MLE

- Suppose  $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$ , where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 
  - Normal Distribution  $p(x; \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$
- $p(d^{(i)}; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(y^{(i)} - \theta^T x^{(i)})^2) \Rightarrow \log p(d^{(i)}; \theta) = \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}(y^{(i)} - \theta^T x^{(i)})^2$
- $\theta_{MLE} = \arg \min_{\theta} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2$

#### 2. MAP

- Suppose  $\epsilon \sim \mathcal{N}(0, \sigma^2), \theta \sim \mathcal{N}(0, \lambda^2 I)$ 
  - Multivariate normal distribution  $p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu))$
  - where  $\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}$  is symmetric and positive semidefinite
- $p(\theta) = \frac{1}{(\sqrt{2\pi}\lambda)^n} \exp(-\frac{1}{2\lambda^2} \theta^T \theta) \Rightarrow \log p(\theta) = n \log \frac{1}{\sqrt{2\pi}\lambda} - \frac{\theta^T \theta}{2\lambda^2}$
- $\theta_{MAP} = \arg \min_{\theta} \left\{ \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 + \frac{\theta^T \theta}{2\lambda^2} \right\}$

#### 3. MLE vs MAP

- MLE (unregularized solution) vs MAP (regularized solution)
- The prior distribution acts as a regularizer in MAP estimation

### Logistic Regression

Similar conclusion as above.