Logistic Regression

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

$$P(y=1|x) = \frac{e^{X\theta}}{1+e^{X\theta}}$$

Finding $\boldsymbol{\theta}$ with Gradient Ascent

- $\hat{\theta} = argmax_\theta \sum_{1 \le i \le N} log(P(Y_i|X_i;\theta))$
  - $f(\theta) = \sum_{1 \le i \le N} log(P(Y_i|X_i;\theta))$
  - $\frac{\partial f(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\{\sum_{1 \le i \le N} log(P(Y_i|X_i;\theta))\} = \sum_{1 \le i \le N} X_{i,j}(Y_i - P(y=1|x;\theta))$
- To utilize the gradient method
  - We need to know *f'(x)* which are above
    - Case of ascent: $x_{t+1} \leftarrow x_t + h\mathbf{u}^* = x_t + h\frac{f'(x_t)}{|f'(x_t)|}$
  - Then, how to iteratively update the parameter, $\boldsymbol{\theta}$
  - $\theta_j^{t+1} \leftarrow \theta_j^t + h\frac{\partial f(\theta^t)}{\partial \theta_j^t} = \theta_j^t + h\{\sum_{1 \le i \le N} X_{i,j}\left(Y_i - P(Y=1|X_i;\theta^t)\right)\}$

$$= \theta_j^t + \frac{h}{C}\{\sum_{1 \le i \le N} X_{i,j}\left(Y_i - \frac{e^{X_i\theta^t}}{1+e^{X_i\theta^t}}\right)\}$$

C=Normalization to the unit vector

  - θ_j^0 can be arbitrarily chosen.

Linear Regression Revisited

- Previously,
  - $\hat{\theta} = argmin_{\theta}(f - \hat{f})^2 = argmin_{\theta}(Y - X\theta)^2$
    $= argmin_{\theta}(Y - X\theta)^T\ (Y - X\theta) = argmin_{\theta}(Y - X\theta)^T\ (Y - X\theta)$
    $= argmin_{\theta}(\theta^T X^T X\theta - 2\theta^T X^T Y + Y^T Y) = argmin_{\theta}(\theta^T X^T X\theta - 2\theta^T X^T Y)$
  - $\nabla_{\theta}(\theta^T X^T X\theta - 2\theta^T X^T Y)=0$
    - $2X^T X\theta - 2X^T Y = 0$
  - $\theta = (X^T X)^{-1} X^T Y$
- Any problem???
- Gradient descent can be a solution
  - $\hat{\theta} = argmin_{\theta}(f - \hat{f})^2 = argmin_{\theta}(Y - X\theta)^2 =$
    $argmin_{\theta} \sum_{1\le i\le N}(Y^i - \sum_{1\le j\le d} X_j^i \theta_j)^2$
  - $\frac{\partial}{\partial\theta_k}\sum_{1\le i\le N}(Y^i - \sum_{1\le j\le d} X_j^i\theta_j)^2 = -\sum_{1\le i\le N} 2(Y^i - \sum_{1\le j\le d} X_j^i\theta_j)X_k^i$
  - $\theta_k^{t+1} \leftarrow \theta_k^t - h\frac{\partial f(\theta^t)}{\partial\theta_k^t} = \theta_k^t + h\sum_{1\le i\le N} 2(Y^i - \sum_{1\le j\le d} X_j^i\theta_j)X_k^i$