Chapter04 Gradient Descent

1. Learning Rate

   Vanilla Gradient Descent: $\theta^i = \theta^{i-1} - \eta \nabla L(\theta^{i-1})$

2. Adagrad

   a) Divide the learning rate of each parameter by the root mean square of its previous derivatives

   b) $w^{t+1} = w^t - \frac{\eta^t}{\sigma^t} g^t \quad \eta^t = \frac{\eta}{\sqrt{t+1}} \quad \sigma^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^{t} (g^i)^2}$

   c) $w^{t+1} = w^t - \frac{\eta}{\sqrt{\sum_{i=0}^{t} (g^i)^2}} g^t$

   d) The best step is: $\frac{|First\ Derivative|}{|Second\ Derivative|}$

3. Stochastic Gradient Descent

   a) Make the training faster

   b) Loss for only one example

4. Feature Scaling

   a) $y = b + w_1 x_1 + w_2 x_2$ make different features have the same scaling

   b) $x_r^i = \frac{x_r^i - m_i}{\sigma_i}$

Chapter05 Classification

1. Classification as Regression: Penalize to the examples that are "too correct"

2. $P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$

3. Using maximum likelihood to get $P(x|C_1)$ and $P(x|C_2)$: The Gaussian with and mean μ and covariance matrix Σ can generate all the points, likelihood function is written as: $L(\mu, \Sigma) = \prod f_{\mu,\Sigma}(x_i)$ $\mu^*, \Sigma^* = argmaxL(\mu, \Sigma)$

   $\mu^* = average(x_i)$ $\Sigma^* = average(x_i - \mu^*)(x_i - \mu^*)^T$

4. To get better performance, we usually use the same covariance matrix

   a) $\Sigma = weighted\_average(\Sigma_1, \Sigma_2)$

   b) The modified model's boundary line is linear

5. Posterior Probability

   a) $P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = \frac{1}{1+exp(-z)} = \sigma(z) = \sigma(w \cdot x + b)$

   b) $z = ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} = w^T x + b$

   $w^T = (\mu^1 - \mu^2)^T \Sigma^{-1} \quad b = ln \frac{N_1}{N_2} + \frac{1}{2}((\mu^2)^T \Sigma^{-1} \mu^2 - (\mu^1)^T \Sigma^{-1} \mu^1)$