Entropy of a dataset with C classes

$$H(S) = -\sum_{i}^{C} p_i \log_2 p_i$$

Entropy of the dataset after splitting from feature x

$$H(S|x) = \sum_{c \in x} P(c) \cdot H(c)$$

Information gain of splitting from feature x

$$IG(S,x) = H(S) - H(S|x)$$

Gini impurity

$$G = 1 - \sum_{i=1}^{C} P(i)^2 \quad \text{or} \quad G = \sum_{i=1}^{C} P(i)(1 - P(i))$$

Linear regression

Hypothesis (Final hypothesis)

$$g(x) = \theta^0 + \theta^1 x$$

Error

$$\text{Error} = [h(x) - y]^2$$

Cost function

$$J(\theta^0, \theta^1) = \frac{1}{2N} \sum_{i=1}^{N} (h(x_i) - y_i)^2$$

Gradient Descent

1. Repeat
2. $\theta^j = \theta^j - \alpha \frac{dJ(\theta^0, \theta^1)}{d\theta^j}$
3. Until converge

$\Downarrow$

Batch gradient descent

1. Repeat
2. $\theta_0 = \theta_0 - \frac{\alpha}{N} \sum_{i=1}^{N} (h(x_i) - y_i)$
3. $\theta_1 = \theta_1 - \frac{\alpha}{N} \sum_{i=1}^{N} (h(x_i) - y_i) x$
4. Until converge

$h(x) = \theta^0 + \theta^1 x$

Multivariate linear regression

Hypothesis function

$$g(x) = \theta^0 x^0 + \theta^1 x^1 + \theta^3 x^3 + \dots + \theta^d x^d$$ where $x^0 = 1$

$$g(x) = \theta^T x \Rightarrow \theta = \begin{bmatrix} \theta^0 \\ \theta^1 \\ \vdots \\ \theta^d \end{bmatrix} \quad x = \begin{bmatrix} x^0 \\ x^1 \\ \vdots \\ x^d \end{bmatrix}$$

Cost function

$$J(\theta^0, \theta^1, \dots, \theta^d) = \frac{1}{2N} \sum_{i=1}^{N} (\theta^T x_i - y_i)^2$$

Gradient descent

1. Repeat
2. $\theta^0 = \theta^0 - \frac{\alpha}{N} \sum_{i=1}^{N} (h(x_i) - y_i) x_i^0$
3. $\theta^1 = \theta^1 - \frac{\alpha}{N} \sum_{i=1}^{N} (h(x_i) - y_i) x_i^1$
4. $\theta^d = \theta^d - \frac{\alpha}{N} \sum_{i=1}^{N} (h(x_i) - y_i) x_i^d$
5. Until converge.

~~Cos~~

Normal equations

Cost function

$$J(\theta^0, \theta^1, \dots, \theta^d) = \frac{1}{2N} \| x\theta - y \|^2$$

Compute θ

$$\theta = (x^T x)^{-1} x^T y$$

$$X = \begin{bmatrix} \dots x_1^T \dots \\ \dots x_2^T \dots \\ \vdots \\ \dots x_N^T \dots \end{bmatrix} \quad \theta = \begin{bmatrix} \theta^1 \\ \theta^2 \\ \theta^3 \\ \vdots \\ \theta^d \end{bmatrix} \quad y = [y_1, y_2, \dots, y_N]$$

Logistic regression

$$k(x) = h(g(x))$$

Classification hypothesis

Linear regression hypothesis.

Restrictive function/logistic function

$$g_\theta(x) = \theta^T x$$

Logistic function/Sigmoid function

$$S_\theta(x) = \frac{1}{1 - e^{-\theta^T x}}$$

Cost function

$$J(\theta) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \log(h_\theta(x_i)) + (1-y_i)\log(1-h_\theta(x_i))\right]$$

Gradient descent

1. Repeat
2. $\theta^j = \theta^j - \alpha \frac{dJ(\theta)}{d\theta^j}$
3. Until converge

For any j

$$\frac{dJ(\theta)}{d\theta^j} = \frac{1}{N}\sum_{i=1}^{N}(h_\theta(x_i) - y_i)\, x_i^j$$

where $h_\theta(x_i) = \frac{1}{1-e^{-\theta^T x}}$

Regularization

Cost function

$$J(\theta) = \frac{1}{2N}\left[\sum_{i=1}^{N}(h(x_i) - y_i)^2 + \lambda \sum_{i=1}^{N}\theta_j^2\right]$$

regularization parameter

Gradiend descent with regularization

1. Repeat
2. $\theta^0 = \theta^0 - \alpha \frac{1}{N}\sum_{i=1}^{N}(h(x_i) - y_i)\, x_i^0$
3. $\theta^j = \theta^j - \alpha\left(\frac{1}{N}\sum_{i=1}^{N}(h(x_i) - y_i) . x_i^j + \frac{\lambda}{N}\theta^j\right)$
4. Until converge

$$\theta^j = \theta^j\left(1 - \frac{\alpha\lambda}{N}\right) - \frac{\alpha}{N}\sum_{i=1}^{N}(h(x_i) - y_i)\, x_i^j$$

Usually $0 < \left(1 - \alpha\frac{\lambda}{N}\right) < 1$

Logistic regression with regularization

Cost function

$$J(\theta) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \log(h_\theta(x_i)) + (1-y_i)\log(1-h_\theta(x_i))\right] + \frac{\lambda}{2N}\sum_{j=1}^{d}\theta_j^2$$

Gradient descent

1. Repeat
2. $\theta^0 = \theta^0 - \alpha \frac{1}{N}\sum_{i=1}^{N}(h(x_i) - y_i)x_i^0$
3. $\theta^j = \theta^j - \alpha\left(\frac{1}{N}\sum_{i=1}^{N}(h(x_i) - y_i).x_i^j + \frac{\lambda}{N}\theta^j\right)$
4. Until converge

$$\theta^j = \theta^j\left(1 - \alpha\frac{\lambda}{N}\right) - \alpha\frac{1}{N}\sum_{i=1}^{N}(h(x_i) - y_i)x_i^j$$

Usually $0 < \left(1 - \alpha\frac{\lambda}{N}\right) < 1$