linear Regression

Dataset: $D = \{(x_1, y_1), \dots (x_m, y_m)\}$

prediction: $\hat{y} = wx_i + b$

cost function: $E(w, b) = \sum_{i=1}^{m}(y_i - wx_i - b)^2$

minimize cost function:

$\frac{\partial E}{\partial w} = 2(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i) = 0$

$\frac{\partial E}{\partial b} = 2(mb - \sum_{i=1}^{m}(y_i - wx_i)) = 0$

so,

$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i)$

$mw\sum_{i=1}^{m} x_i^2 = m\sum_{i=1}^{m} y_i(x_i - \overline{x}) + (\sum_{i=1}^{m} x_i)^2$

then, $w = \frac{\sum_{i=1}^{m} y_i(x_i - \overline{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum_{i=1}^{m} x_i)^2}$

least square

x is a multivariable with a dimension of d, with the expression of $x_i = (x_{i1}, \dots x_{id})$, then the dataset can be expressed as a matrix X as:

$$\left\{\begin{matrix} x_{11} & \dots & x_{1d} & 1 \\ x_{21} & \dots & x_{2d} & 1 \\ \vdots & \ddots & \vdots & \\ x_{m1} & \dots & x_{md} & 1 \end{matrix}\right\} = \left\{\begin{matrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{matrix}\right\}$$

and the label can also be written as vector $y = (y_1, \dots, y_m)^T$, then the predicted w is also a vector with the expression $\hat{w} = (w, b)$, and the predicted label $\hat{y} = X\hat{w}$

cost function:

$E_{\hat{w}} = (y - X\hat{w})^T(y - X\hat{w})$

$\frac{\partial E_{\hat{w}}}{\partial \hat{w}} = 2X^T(X\hat{w} - y) = 0$

if X^TX is full-rank matrix or positive definite determined matrix, we can obtain:

$\hat{w}^* = (X^Tx)^{-1}x^Ty$

regularization terms

Since X^Tx can be non-full-rank, we can introducte regularization terms. A general regularizer is:

$$\frac{1}{2}(y - X\hat{w})^T(y - X\hat{w}) + \frac{\lambda}{2}\sum_{i=1}^{m}|w_i|^q \quad (1)$$

- L2 regularization ($q = 2$)

By adding a term of $\frac{\lambda}{2}\hat{w}^T\hat{w}$ in the cost function, we obtain: $\hat{w}^* = (X^Tx + \lambda I)^{-1}x^Ty$

- lasso ($q = 1$)

If λ is sufficiently large, then some coefficients are zero, then w is a sparse matrix.

Logistic regression

sigmoid function: $y = \frac{1}{1+exp(-(w^Tx+b))}$ which is equivalent to : $ln\frac{y}{1-y} = w^Tx + b$

let's see y as the probability of label "1" and $1 - y$ as label "0", then the ratio between them represents the relative possibility of "1".

$p(y = 1|x) = \frac{exp(w^Tx+b)}{1+exp(w^Tx+b)} = p_0(x)$ $p(y = 0|x) = \frac{1}{1+exp(w^Tx+b)} = p_1(x)$

Maximum Likelyhood Method

$$l(w, b) = \sum_{i=1}^m lnp(y_i|x_i; w, b) \quad (2)$$

let $\beta = (w, b), \hat{x} = (x; 1)$, then $w^Tx + b = \beta^T\hat{x}$

then $p(y_i|x_i; w, b) = y_ip_1(\hat{x}_i; \beta) + (1 - y_i)p_0(\hat{x}_i; \beta)$

so the maximum of (2) is equivalent to minimize: $l(\beta) = \sum_{i=1}^m(-y_i\beta^T\hat{x}_i + ln(1 + e^{\beta^T\hat{x}_i})*)$

partial differential:

$l^{(1)} = \frac{\partial l(\beta)}{\partial\beta} = -\sum_{i=1}^m \hat{x}_i(y_i - p_1)$ $l^{(2)} = \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^m \hat{x}_i\hat{x}_i^Tp_1(1 - p_1)$

- Newton's iteration of parameter update

$\beta = \beta^t - (l^{(2)})^{(-1)}l^{(1)}$

Linear Discriminat Analysis

By projecting the dataset to a line w, then the projection centers are $w^T\mu_0$ and $w^T\mu_1$, respectively; the covariances are $w^T\Sigma_0w$ and $w^T\Sigma_1w$, respectively.

- To minimize the projected distance within the same class, we can minimize the covariance within the class (minimize $w^T\Sigma_0w + w^T\Sigma_1w$);
- To maximize the projected distance between different classes, we can maximize the center distance (maximize $||w^T\mu_0 - w^T\mu_1||^2$)

Considering both goals, we can set :

$$J = \frac{||w^T\mu_0-w^T\mu_1||^2}{w^T\Sigma_0w+w^T\Sigma_1w} = \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^Tw}{w^T(\Sigma_0+\Sigma_1)w} \quad (3)$$

Hence, we can define the within-class scatter matrix: $S_w = \Sigma_0 + \Sigma_1 = \sum_{x\in X_0}(x - \mu_0)(x - \mu_0)^T + \sum_{x\in X_1}(x - \mu_1)(x - \mu_1)^T$ and between-class scatter matrix $S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$

then (3) can be re-written as: $J = \frac{w^TS_bw}{w^TS_ww}$, we can maximize this term with the subjection of $w^TS_ww = 1$ since only the direction of w matters.

Using Lagurange method, the solutions is: $$S_b w = \lambda S_w w \quad (4)$$ Since the direction of $S_b w = (\mu_0 - \mu_1)^T w(\mu_0 - \mu_1)$ is along $(\mu_0 - \mu_1)$ we assume $S_b w = \lambda(\mu_0 - \mu_1)$.

With (4), with can obtain: $w = S_w^{-1}(\mu_0 - \mu_1)$