PS: Pre-Learning & 先导课 & Lecture.

Linear Regression and Beyond.

(I) General Formation

Suppose x describe the characteristic of sample and $x = \{x_1; x_2; \cdots; x_d\}$
And the Linear Model attempt to learn a predict function by the combination of characteristic.

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$
$$f(x) = w^T x + b$$

General Formation

(II) Linear Regression

Input: Supposet Data Set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\} = \{(x_i, y_i)\}_{i=1}^{m}$

Output: The predic function:

$$f(x_i) = w^T x_i + b$$

Q: How to determine the value w^T and b

A: Peformance Measurement:

$$\langle w^*, b^* \rangle = \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

$$= \arg\min_{(w,b)} \sum_{i=1}^{m} (y_i - w x_i - b)^2$$

Above equation is (均方误差) and it is similar to 《欧式距离》
According to Linear Regression. "最小二乘法" (least square method) attempt to find the line, which the Euclidien Distances of the whole samples are Least.

Solve:

$$E(w,b) = \sum_{i=1}^{m} (y_i - w x_i - b)^2$$ the process of minimize.

对 w, b 分别求导:

假设 $g(x_i) = f(x_i) - w_i x_i - b$

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

$$= 2\left(w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b) x_i\right) \quad ①$$

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

$$\frac{\partial E(w,b)}{\partial b} = \sum_{i=1}^{m} 2\cdot\left(f(x_i) - w x_i - b\right)\cdot(-1)$$

$$= 2\left(mb - \sum_{i=1}^{m}(y_i - w x_i)\right) \quad ①$$

Let the eqn ① & ② to be zero. Thus, we can arrive the closed form solution.

①: $w = \dfrac{\sum_{i=1}^{m}(y_i - b)x_i}{\sum_{i=1}^{m} x_i^2}$ ②: $b = \dfrac{\sum_{i=1}^{m}(y_i - w x_i)}{m}$

③→①: ③:

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

$$\Rightarrow \sum_{i=1}^{m}(y_i - w x_i)\cdot\sum_{i=1}^{m} x_i = \left(\sum_{i=1}^{m} y_i x_i - w\sum_{i=1}^{m} x_i^2\right)\cdot(m)$$

$$w\cdot\sum_{i=1}^{m} x_i^2\cdot m = \sum_{i=1}^{m} y_i x_i m - \sum_{i=1}^{m}(y_i - w x_i)\sum_{i=1}^{m} x_i$$

$$\Rightarrow w = \frac{\sum_{i=1}^{m} y_i\cdot x_i - m\sum_{i=1}^{m} y_i\cdot x_i}{\left(\sum_{i=1}^{m} x_i\right)^2 - m\sum_{i=1}^{m} x_i^2} = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\sum_{i=1}^{m} x_i^2}$$

- Empirical Variance: $SS_{tot} = \sum_{i=1}^{m}(y_i - \bar{y})^2$
- Squared Error: $SS_{res} = \sum_{i=1}^{m}(y_i - \tilde{y}_i)^2$
- R^2 value: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$ → 评价指标

(Ⅲ) 对数几率回归

$f(z) = \dfrac{1}{1+e^{-z}}$

若 $z = f(x) = wx + b$.

$$\Rightarrow y = \frac{1}{1+e^{-(wx+b)}}$$

$$\Rightarrow (wx+b) = \ln\frac{y}{1-y} \quad ①$$

若是二分类：y 视为类后验概率估计.

$$\ln\frac{P\{y=1|x\}}{P\{y=0|x\}} = w^T x + b.$$

$$\begin{cases} P\{y=1|x\} = \dfrac{e^{w^Tx+b}}{1+e^{w^Tx+b}} \\ P\{y=0|x\} = \dfrac{1}{1+e^{w^Tx+b}} \end{cases}$$

"Maximum Likelihood Method" 极大似然估计.

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

in which, $p(y_i|x_i;w,b)=y_i\cdot p_1(\hat{x}_i;\beta)+(1-y_i)\cdot p_0(\hat{x}_i;\beta)$.

$$l(w,b)=\sum_{i=1}^{m}\ln(y_i\cdot p_1(\hat{x}_i;\beta)+(1-y_i)\cdot p_0(\hat{x}_i;\beta))$$

$$p_1(\hat{x}_i;\beta)=\frac{e^{\beta x}}{1+e^{\beta x}}\ ;\ p_0(\hat{x}_i;\beta)=\frac{1}{1+e^{\beta x}}$$

$$l(w,b)=\sum_{i=1}^{m}\ln\left(y_i\cdot\frac{e^{\beta x}}{1+e^{\beta x}}+(1-y_i)\cdot\frac{1}{1+e^{\beta x}}\right)$$

最大似然

$$=\sum_{i=1}^{m}\ln\left(\frac{1}{1+e^{\beta x}}+y_i\left(\frac{e^{\beta x}-1}{1+e^{\beta x}}\right)\right)$$

↳ 等同于 $\min=\sum_{i=1}^{m}-y_i\beta^T x+\ln(1+e^{\beta x})$.

- 梯度下降法 / 牛顿法.

→ 牛顿法:

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

$$\beta^*=\arg\min_\beta l(\beta)$$

二阶项　一阶项

$$\beta^{t+1}=\beta^t-\left(\frac{\partial^2 l(\beta)}{\partial\beta\cdot\partial\beta^T}\right)^{-1}\frac{\partial l(\beta)}{\partial\beta}$$

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

(IV) Solving Linear Regression: Gradient Descent

Regression: input: n data points (d dimension)

output: Find a map $f: R^{d+1}\to R$

$$f(x)=w^Tx+b=w_1x_1+w_2x_2+\cdots+w_dx_d+w_0.$$
$$=(w,w_0)^T\cdot(x,1)$$

Loss-Function: Mean Squared Error

$$L(f)=\sum_{i=1}^{n}(f(x_i)-y_i)^2$$

Minimize the Loss-Function how to find a group of w.

↳ To solve

$$\min_w\sum_{i=1}^{n}\left(w_0+\sum_{j=1}^{d}w_jx_j^{(i)}-y^{(i)}\right)^2$$

Above is the Convex Function (凸函数). About it, there are some typical theory to convenient (优化):

Gradient Descent Method (梯度下降法)

Newton Method (牛顿法) which has mentioned above.

- **Gradient Descent Method (梯度下降法):**

Def.: f is any function in $w=(w_0, w_1, \cdots, w_d)$ (Loss Function)

◦ Gradient Descent Method (梯度下降法):

- L: it is any function in $w = (w_0, w_1, \cdots, w_d)$ (Loss Function)
- The Gradient is:

$$\nabla L(w) = \left(\frac{dL(w)}{dw_0}, \frac{dL(w)}{dw_1}, \cdots, \frac{dL(w)}{dw_d}\right)$$

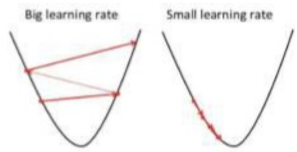
And For Descent, it's need for choose a Learning Rate $\delta > 0$.
After Initialize the first parm: $w^{(1)} = (w_0^{(1)}, w_1^{(1)}, \cdots, w_d^{(1)})$

Loop: for $i = 1$ to T:

$$w^{(i+1)} = w^{(i)} - \delta \nabla L(w^{(t)})$$

return $w^{(T+1)}$

Gradient Descent for minimizing $L(\mathbf{w})$
Parameters: Learning Rate $\delta > 0$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L(\mathbf{w}^{(t)})$
Return $w^{(T)}$



Q: In every iteration, we need to recompute $L(w^{(t)})$
A:

Stochastic Gradient Descent for Loss Functions

Parameters: Learning rate $\delta > 0$, Loss function $L(\mathbf{w}) = \sum_{i=1}^{n} L_i(\mathbf{w})$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
Pick one training example $i = 1, 2, \ldots, n$ uniformly at random
Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L_i(\mathbf{w}^{(t)}, b)$
Return $w^{(T)}$

Compare Two Method:

① Gradient Descent: Coverages Quickly and Smoothly
② Stochastic Gradient Descent: May converges more slowly and unsteadily.
(※) ↳ But can compute more easily and fastly.

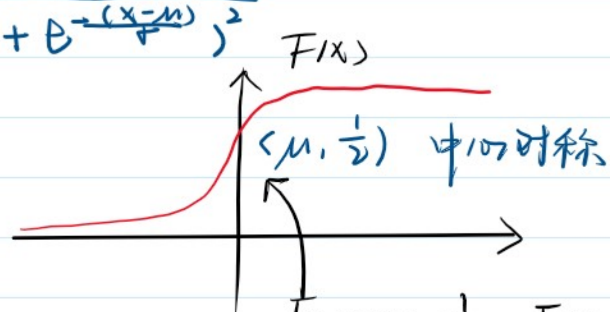
(V) Logistic Regression (逻辑回归)

Logistic Regression is a typical classification Method.

◦ Logistic Distribution:

$$F(x) = P(X \leq x) = \frac{1}{1 + e^{-\frac{(x-\mu)}{\gamma}}}$$

$$f(x) = F'(x) = \frac{e^{-\frac{(x-\mu)}{\gamma}}}{\gamma\left(1 + e^{-\frac{(x-\mu)}{\gamma}}\right)^2}$$



$$F(-x+\mu) - \frac{1}{2} = -F(x+\mu) + \frac{1}{2}$$

Similar to Expolominal Regression.

$$\left\{ P\{Y=1 \mid x\} = \frac{\exp(w \cdot x + b)}{\quad} \right.$$

Similar to Polynomial Regression:

$$\begin{cases} P\{Y=1 \mid x\} = \dfrac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)} \\ P\{Y=0 \mid x\} = \dfrac{1}{1 + \exp(w x + b)} \end{cases} \rightarrow \log \frac{P\{Y=1 \mid x\}}{1 - P\{Y=1 \mid x\}} = wx + b.$$