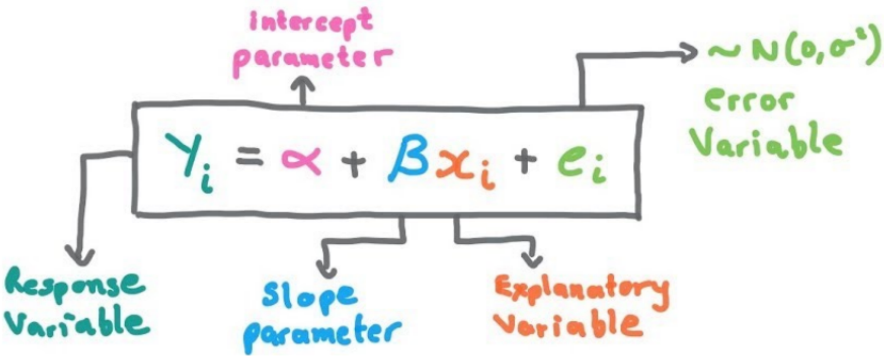
ch2: Linear Regression

Regression Model Definition(P4)

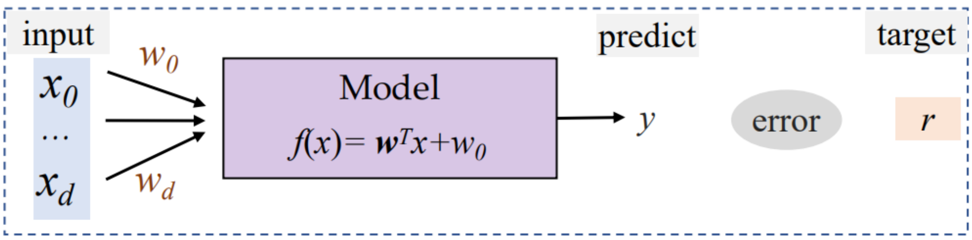
- A regression model provides a function that describes the relationship between one or more **independent variables** and a response, **dependent**, or target variable.

回归模型提供了一个函数，用来描述一个或多个自变量与响应、因变量或目标变量之间的关系



Model Architecture(P6)

A simple linear function.



- Train:
  - estimate the parameters $\boldsymbol{w}$ and w_0 from data
- Test:
  - calculate $f(x) = \boldsymbol{w}^T x + w_0$.

Loss Function(P7)

- For a given input x, the model outputs a real value y. Let $r \in \mathbb{R}$ be target value, the square error is :

$$l\,(\boldsymbol{w}, w_0 \mid x, r) = (r - y)^2$$

- Given: $D = \{(x^{(1)}, r^{(1)}), \ldots, (x^{(N)}, r^{(N)})\}$, the loss over the dataset is defined as the mean square error (MSE):

$$L(\boldsymbol{w}, w_0 \mid D) = \frac{1}{2N} \sum_{\ell=1}^{N} (r^{(\ell)} - y^{(\ell)})^2$$

Optimization(P8-9)

Given: $D = \{(x^{(1)}, r^{(1)}), \ldots, (x^{(N)}, r^{(N)})\}$

minimize the loss function using gradient descend:

- Goal:

$$\min_w L(\boldsymbol{w})$$

- Iteration:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \frac{\partial L}{\partial w}$$

$$L(\boldsymbol{w}, w_0 \mid D) = -\ 1/2\mathrm{N} \textstyle\sum_{\ell=1}^{N} (r^{(\ell)} - y^{(\ell)})^2$$

What is $\frac{\partial L}{\partial w}$?

For each $\boldsymbol{w}_j$ ($j = 1,\ldots,d$):

$$\frac{\partial L}{\partial w_j} = -\frac{1}{N}\underbrace{\sum_{\ell}\left(r^{(\ell)} - y^{(\ell)}\right)\frac{\partial y^{(\ell)}}{\partial w_j}}_{\text{Chain rule}} = -\frac{1}{N}\sum_{\ell}\left(r^{(\ell)} - y^{(\ell)}\right)x^{(\ell)}$$

$$\boldsymbol{w}_{\mathrm{new}} = \boldsymbol{w}_{\mathrm{old}} + \tfrac{1}{N}\textstyle\sum_{\ell=1}^{N} (r^{(\ell)} - y^{(\ell)})\, x^{(\ell)}$$

- Gradient Descend for Liner Regression(P9)(没细看，应该不会考吧)

Gradient Descend for Liner Regression
Input: $D = \{(\mathrm{x}^{(l)}, \mathrm{r}^{(l)})\}$ $(l = 1{:}N)$
for $j = 0,\ldots,\mathrm{d}$
$\quad \boldsymbol{w}_j \leftarrow rand\,(-0.01, 0.01)$
repeat
$\quad$ **for** $j = 0,\ldots,d$
$\quad\quad \Delta\boldsymbol{w}_j \leftarrow 0$
$\quad$ **for** $l = 1,\ldots,N$
$\quad\quad y \leftarrow 0$
$\quad\quad$ **for** $j = 0, \ldots, d$
$\quad\quad\quad y \leftarrow y + \boldsymbol{w}_j x_j^{(l)}$
$\quad\quad \Delta\boldsymbol{w}_j \leftarrow \Delta\boldsymbol{w}_j + (r^{(l)} - y) x_j^{(l)}$
$\quad \Delta\boldsymbol{w}_j = \Delta\boldsymbol{w}_j / \mathrm{N}$
$\quad$ **for** $j = 0,\ldots,d$
$\quad\quad w_j \leftarrow w_j + \eta\Delta w_j$
until convergence

The Matrix Form（P11）

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \vdots \\ \boldsymbol{x}^{(N)} \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^{(N)} & x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{bmatrix} \qquad \boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \qquad \boldsymbol{r} = \begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(N)} \end{bmatrix}$$

- Prediction: $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{w}=\begin{bmatrix} \boldsymbol{x}^{(1)}\boldsymbol{w} \\ \boldsymbol{x}^{(2)}\boldsymbol{w} \\ \vdots \\ \boldsymbol{x}^{(N)}\boldsymbol{w} \end{bmatrix}$
- Objective: $L(\boldsymbol{w})=\frac{1}{2}(\boldsymbol{r}-\boldsymbol{y})^T(\boldsymbol{r}-\boldsymbol{y})=\frac{1}{2}(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})$
- Gradient

$$\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}=-\boldsymbol{X}^T(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})$$

- Solution

loss最小的时候

$$\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}=0 \Rightarrow \boldsymbol{X}^T(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})=0$$

$$\Rightarrow \boldsymbol{X}^T\boldsymbol{r}=\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

$$\Rightarrow \boldsymbol{w}*=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{r}$$

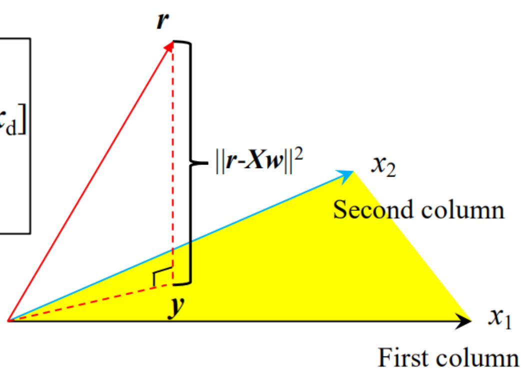
- Then the predicted values are

$$\boldsymbol{y}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{r}$$

$$=\boldsymbol{H}\boldsymbol{r}$$

Geometrical Explanation

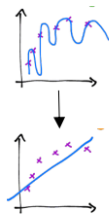
- The column vectors $[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_d]$ form a subspace of $\mathbb{R}^n$.
- $\boldsymbol{H}$ is a least square projection



出现奇异值无法直接计算的时候：

When some column vectors are not independent (e.g., $\boldsymbol{x}_2=3\boldsymbol{x}_1$), then $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is singular, thus $\boldsymbol{w}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$ cannot be directly calculated.

Solution: Regularization



$$\boldsymbol{L}(\boldsymbol{w})=\frac{1}{2}(\boldsymbol{r}-\boldsymbol{y})^T(\boldsymbol{r}-\boldsymbol{y})=\frac{1}{2}(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

New gradient: $\dfrac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}=-\boldsymbol{X}^T(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})+\lambda\boldsymbol{w}$

New optimal solution: $\dfrac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}}=0 \Rightarrow -\boldsymbol{X}^T(\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w})+\lambda\boldsymbol{w}=0$

$$\Rightarrow \boldsymbol{X}^T\boldsymbol{r}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\mathbf{I})\boldsymbol{w}$$

$$\Rightarrow \boldsymbol{w}^*=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\mathbf{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$$