Supervise Learning:Given data X in feature space and labels Y,
learn to predict Y from given X.

Label: 可以是 discrete 的 or continuous 的
对于 discrete 的 label, 这类问题称为 classification
对于 continuous 的 label, 这类问题称为 regression.,

Linear regression 的 topics:
1. Objective function
2. Vectorization
3. 计算 gradient
4. Batch v.s. Stochastic gradient
5. Closed form solution

Lec2 linear regression I

Notation $x \in \mathbb{R}^d$: data

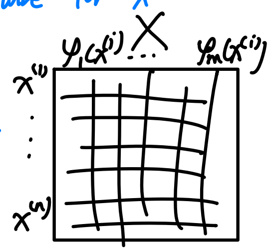
$\varphi(x) \in \mathbb{R}^m$: features for x

$\varphi_j(x) \in \mathbb{R}$: the j^{th} feature for x

$y \in \mathbb{R}$: ctn label

$x^{(n)}$: the n^{th} training example

$y^{(n)}$: the n^{th} training label



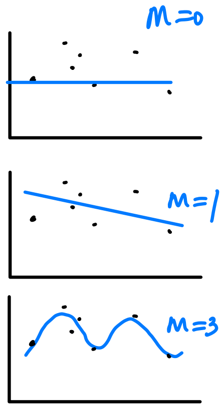
I. regression expression

1-d case $D=1$

$\{x^{(1)}, \dots, x^{(N)}\}$, $\{y^{(1)}, \dots, y^{(N)}\}$

We want: $h(x, w) \approx y$

input parameter



general case

Bias term

$$h(x,w) = w_0 + \sum_{j=1}^{M-1} w_j \varphi_j(x)$$

$$= w^T \varphi(x) \quad \varphi_0(x) = 1$$

$(= w \cdot \varphi(x))$

$w = (w_0, \dots, w_{M-1})^T$, $\varphi(x) = (1, \varphi_1(x), \dots, \varphi_{M-1}(x))^T$

$$w = \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix}, \quad \varphi(x) = \begin{bmatrix} 1 \\ \varphi_1(x) \\ \vdots \\ \varphi_{M-1}(x) \end{bmatrix}$$

Rmk basis function φ_j need not be linear ex $y = w_0 + w_1^2 x_1$, $y = e^{w_0 + w_1 x}$ 不是 linear regression

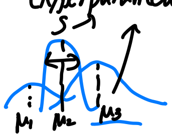
linear regression 指 y 和 weight vector 之间的关系

i.e. 固定x, $h(x, av+bw) = a h(x,v) + b h(x,w)$

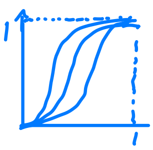
ex $\varphi_j(x) = x^j$ (polynomial)

$$\varphi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right) \quad \text{(Gaussian)}$$

(hyperparameter)



$$\varphi_j(x) = \frac{1}{1+\exp\left(-\frac{x-\mu_j}{s}\right)} \quad \text{(sigmoid)}$$

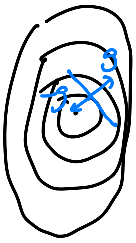


II. objective function (loss ~)

We use: sum of squares error

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} \| h(x^{(n)}, w) - y^{(n)} \|_{L2}^2$$

$$= \frac{1}{2}\sum_{n=1}^{N} \left(\sum_{i=0}^{M-1} w_i \varphi_i(x^{(n)}) - y^{(n)}\right)^2$$



$$\nabla E(w) = \begin{bmatrix} \frac{\partial E}{\partial w_0}(w) \\ \vdots \\ \frac{\partial E}{\partial w_m}(w) \end{bmatrix}$$

$$\text{where } \frac{\partial E}{\partial w_k}(w) = \frac{\partial}{\partial w_k}\left[\frac{1}{2}\sum_{n=1}^{N}\left(\sum_{i=0}^{M-1} w_i \varphi_i(x^{(n)}) - y^{(n)}\right)^2\right]$$

II. computing gradient

$$\frac{\partial E}{\partial w_k}(w) = \frac{\partial}{\partial w_k}\left[\frac{1}{2}\sum_{n=1}^{N}\left(\sum_{i=0}^{M-1} w_i \varphi_i(x^{(n)}) - y^{(n)}\right)^2\right]$$

$$= \frac{1}{2}\sum_{n=1}^{N} \frac{\partial}{\partial w_k}\left(w_0\varphi_0(x^{(n)}) + \dots + w_k\varphi_k(x^{(n)}) + \dots + w_{M-1}\varphi_{M-1}(x^{(n)}) - y^{(n)}\right)^2$$

const

即 $\frac{d}{dw_k}(a w_k + b)^2 = 2a(a w_k + b)$

where $a = \varphi_k(x^{(n)})$

$$= \sum_{n=1}^{N}\left[\left(\sum_{i=0}^{M-1} w_i \varphi_i(x^{(n)})\right) - y^{(n)}\right]\varphi_k(x^{(n)})$$

$=$

So

$$\nabla E(w) = \sum_{n=1}^{N}\left\{\left(\left(\sum_{i=0}^{M-1} w_i \varphi_i(x^{(n)})\right) - y^{(n)}\right)\begin{bmatrix}\varphi_0(x^{(n)}) \\ \vdots \\ \varphi_{M-1}(x^{(n)})\end{bmatrix}\right\}$$

(pred − actual) for $x^{(n)}$

$$= \sum_{n=1}^{N}\left(w^T\varphi(x^{(n)}) - y^{(n)}\right)\varphi(x^{(n)})$$

$$= \sum_{n=1}^{N}\left(\text{pred}^{(n)} - \text{actual}^{(n)}\right) \text{feature.vector}^{(n)}$$

III. Batch Gradient Desent v.s Stochastic Gradient Descent

Batch Gradient Desent:

repeat till convergence:

$$w := w - \eta \nabla E(w)$$

通常设置 $\mu = 0.01$ 等

Stochastic Gradient Descent:

repeat until conv:

Randomly suffle N samples

for $n = 1, \dots, N$

$$w := w - \eta \nabla E(w \mid x^{(n)})$$

note: $\nabla E(w) = \sum_{n=1}^{N} (\text{pred}^{(n)} - \text{actual}^{(n)}) \text{feature}^{(n)}$

而 $\nabla E(w \mid x^{(n)}) = (\text{pred}^{(n)} - \text{actual}^{(n)}) \text{feature}^{(n)}$

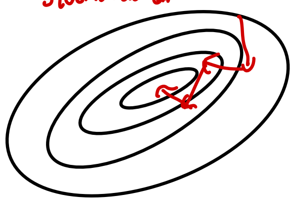
即其中一项

Rmk usually set

$$\eta_t \propto \frac{1}{t} \text{ 或 } \eta_t = \eta_1 \frac{1}{1 + \frac{t-1}{r}}$$

batch GD

Stochastic GD



Closed form sol

let $\Phi = \begin{pmatrix} \varphi_0(x^{(1)}) & \cdots & \varphi_{M-1}(x^{(1)}) \\ \vdots & & \vdots \\ \varphi_0(x^{(N)}) & \cdots & \varphi_{M-1}(x^{(N)}) \end{pmatrix}$

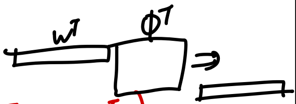
$$E(w) = \frac{1}{2} \sum_{n=1}^{N} (w^T \varphi(x^{(n)}) - y^{(n)})^2 \quad \left(= \frac{1}{2} \| \Phi w - y \|^2\right)$$

$E: \mathbb{R}^M \to \mathbb{R}$

$$= \frac{1}{2} \| \Phi w - y \|^2$$

$$= \frac{1}{2} (\Phi w - y)^T (\Phi w - y)$$

$$= \frac{1}{2} (w^T \Phi^T - y^T)(\Phi w - y)$$



$$= \frac{1}{2} (w^T \Phi^T \Phi w - \underbrace{w^T \Phi^T y}_{} - \underbrace{y^T \Phi w}_{} + y^T y)$$

the same, scalar

$$= \frac{1}{2} w^T \Phi^T \Phi w - w^T \Phi^T y + \frac{1}{2} y^T y$$

(recall 395: chain rule to differentiate $f: \mathbb{R}^m \to \mathbb{R}^n$)