PART 1

Linear regression

x_0	x_1	y
1	1	2
1	4	3
1	6	4

Dummy variable

$y = w_0 + w_1 x_1$

$= \bar{w}^T \bar{x}$

y observed $\quad \hat{y}$ predicted

Mean squared error

$$J(\bar{w}) = \frac{1}{3}\left[(2 - (w_0 + w_1))^2 + (3 - (w_0 + 4w_1))^2 + (4 - (w_0 + 6w_1)) \right]$$

$$\bar{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \quad \bar{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\nabla J = \left[\frac{\partial J}{w_0}, \frac{\partial J}{\partial w_1} \right]^T = \bar{0} \to \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \quad w_0 = 1, \; w_1 = 0.5$$

Matrix Notation

x_0	x_1	y
1	1	2
1	4	3
1	6	4

$\bar{X}$ 3x2 $\bar{y}$ 3x1

$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$ 2x1

Design Matrix

$$J(\bar{w}) = \frac{1}{3}(\bar{y} - \bar{X}\bar{w})^T(\bar{y} - \bar{X}\bar{w})$$

3x1 3x2 2x2

3x1

$$\nabla J = 0$$

$$\Rightarrow \bar{w} = (\bar{X}^T\bar{X})^{-1}\bar{X}^T\bar{y}$$

2x2 2x3 3x1

2x1

Ordinary least square (OLS)

Solution

Probabilistic Linear Regression

So far: $\bar{y} = \bar{w}^T \bar{x}$ $\bar{y}$ completed determined by $\bar{x}$

More reasonable model

$$\bar{y} = \bar{w}^T \bar{x} + \text{impact of other minor factors}$$

$$= \bar{w}^T \bar{x} + \underline{\varepsilon} \qquad \varepsilon \sim N(0, \sigma^2)$$

Noise ↖ Normal distribution

$$p(\bar{y} \mid \bar{x}, \theta) = N(\underbrace{\mu(\bar{x})}_{\bar{w}^T \bar{x}}, \sigma^2)$$

$\theta = (\bar{w}, \sigma)$

$$D = \{\bar{x}_i, y_i\}_{i=1}^{N}$$

Minimize $$-\frac{1}{N}\sum_{i=1}^{N} \log p(y_i | \bar{x}_i, \theta)$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \bar{w}^T \bar{x}_i)^2}{2\sigma^2}} \right]$$

$$= -\frac{1}{N}\sum_{i=0}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2}\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{w}^T \bar{x}_i)^2$$

$\sigma = 1$

MSE

Polynomial Regression

x_0	x_1	y
1	1	2
1	4	3
1	6	4

$$y = w_0 + w_1 x_1 = \bar{w}^T \bar{x}$$

More Complex model

$$y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 x_1^3$$

		x_2	x_3	
x_0	x_1	x_1^2	x_1^3	y
1	1	1	1	2
1	4	16	64	3
1	6	36	216	4

$$\bar{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \Rightarrow \phi(\bar{x}) = \begin{bmatrix} x_0 \\ x_1 \\ x_1^2 \\ x_1^3 \end{bmatrix}$$

feature transformation mapping

Poly Reg. = Poly feature transformation + Linear Reg

Q: what polynomial should be used?

order d

1

3

10

⋮

Hypothesis Space

$Y = w_0 + w_1 x_1$

$\bar{w}_1 = [w_0, w_1]$

$S_1 = \{ Y = w_0 + w_1 x_1 \mid w_0, w_1 \in R^1 \}$

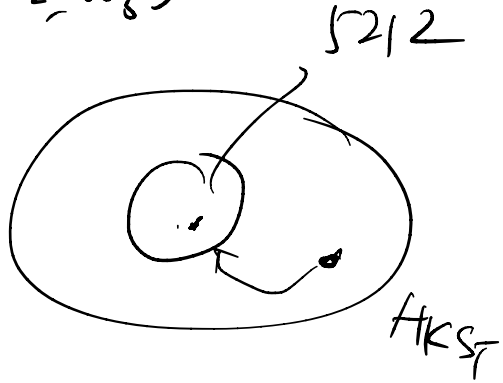
$Y = 1 + x_1$

$Y = -0.2 + 5x_1$

$S_2 = \{ Y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 x_1^3 \mid w_0, w_1, w_2, w_3 \in R^1 \}$

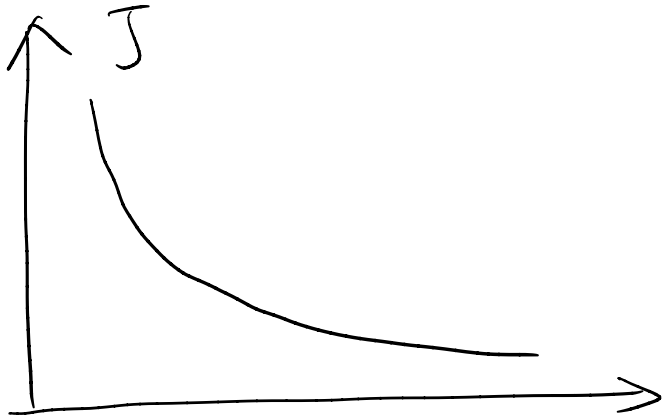
$S_1 \subset S_2$

$\bar{w}_2 = (w_0, w_1, w_2, w_3)^T$



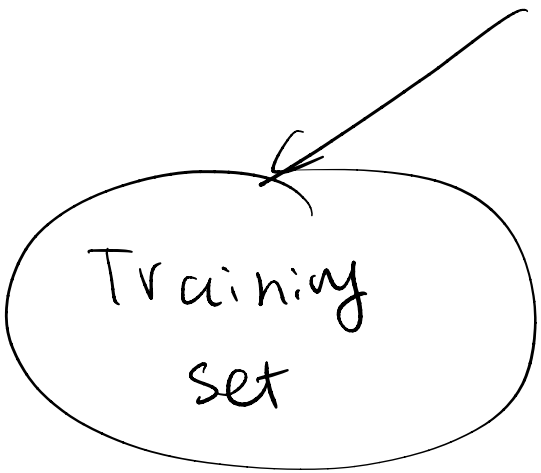
$\min_{\bar{w}_1} J(\bar{w}_1) \geq \min_{w_2} J(\bar{w}_2)$

height of shortest person in Comp5212 $\geq$ (B) height of shortest person in HKUST

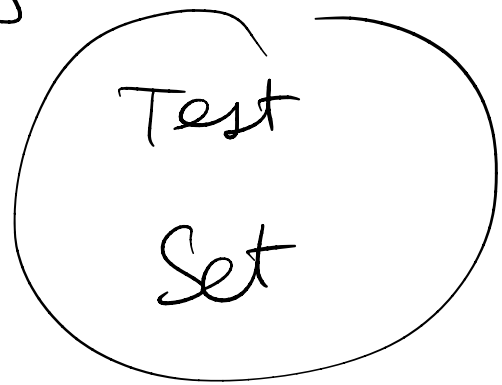


$P(\bar{x}, y)$

iid



You

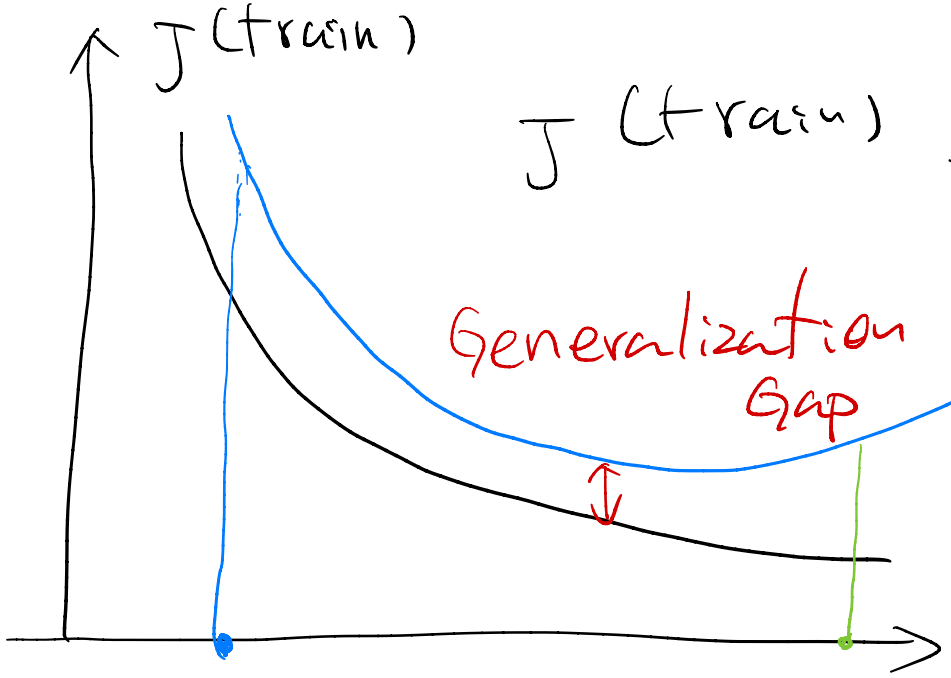


User

Domain Shift

Domain Adaptation

Domain Generalized



$J(\text{train}) \leq J(\text{test})$

Overfitting

Training set: Pattern + noise1

Test set: Pattern + noise2

Model too simple

* cannot fit pattern

* $J(\text{train})$, $J(\text{test})$ large

Underfitting

Model too complex

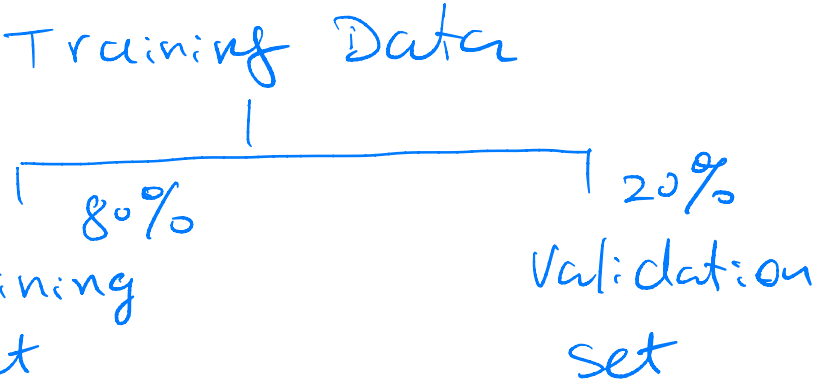
* Fit pattern + noise1

* $J(\text{train})$ small

* $J(\text{test})$ large

dist(noise1, noise2)

Validation for choosing hyperparamters d

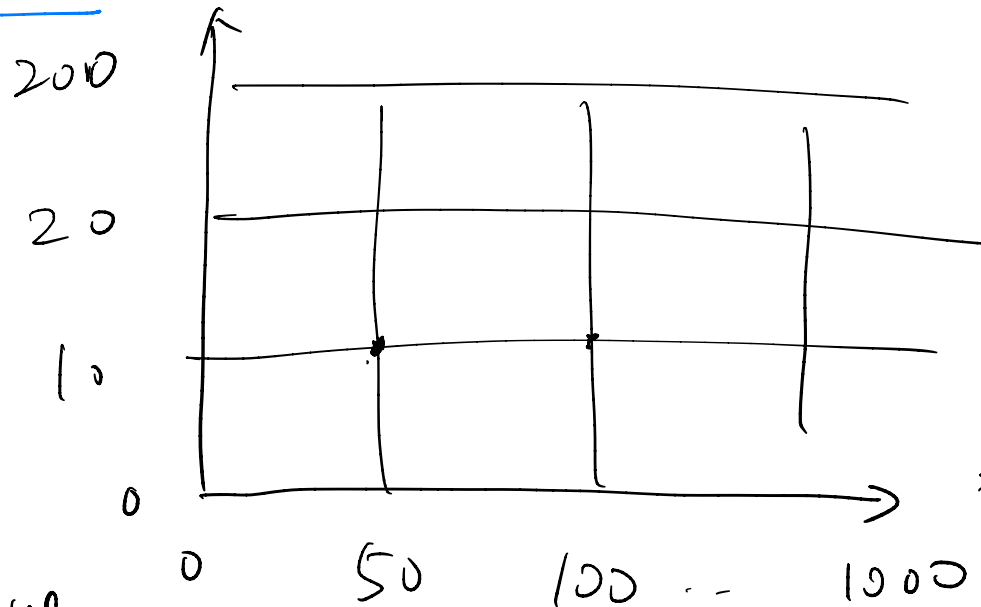


$d = 1$	1.5	1.6
$d = 5$	0.3	0.35 ✓
$d = 10$	0	0.5

what values to consider for d?

1, 2, 3, - - - -

Grid Search

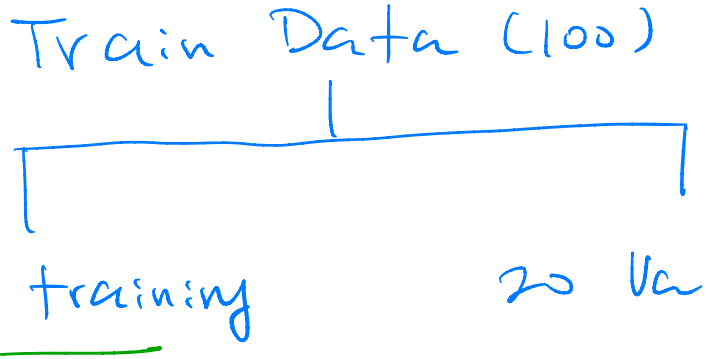


Auto ML

Exhaustive

200,000 models

When data is scarce

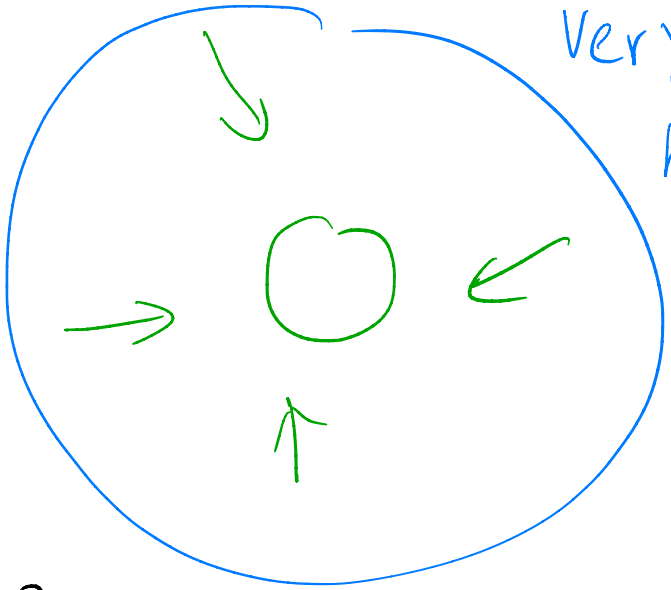


* Not use all data in model construction
* Validation error has high Variance

Income of Hker :

		Team1	Team2
		12K	21K
Method 1:	Survey 5 people Average	$\delta_1 = \vert$ 0 —	0 $\vert$
		V 18K	18.1K
Method 2:	Survey 1000 people Average	$\delta_2 = \vert$ 0 —	0 $\vert$

Regularization



Very large hypothesis space

Model space

$$\min_{J(\bar{w})} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \bar{w}^T \phi(\bar{x}_i))^2$$

Bias $w = \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix}$

Prior: Simpler model

$$+ \lambda \|\bar{w}\|_2^2 \leftarrow w_1^2 + w_2^2 + \cdots + w_D^2$$

Two competing objective

- min loss
- min weights

pushing down the weights for the feature

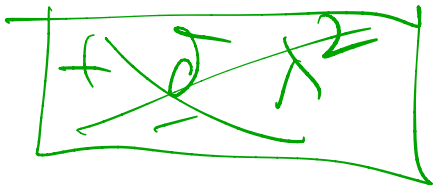
Un important

$$y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 x_1^3 + w_4 x_1^4$$

0.8 1.5 0.2 0.001 0.0001

$$J(\bar{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + \bar{w}^T \phi(\bar{x}_i))^2 + \lambda \|\bar{w}\|_2^2$$

λ: balance parameter



$\lambda = 0$

$\lambda = 10$

$\lambda = 1000$

How to choose λ?

$$J(\bar{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \bar{w}^T \phi(\bar{x}_i))^2 + \lambda \|\bar{w}\|_2^2$$

$w_1^2 + w_2^2 + \cdots$

L_2 regularization, weight decay, ridge regression

$$J(\bar{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \bar{w}^T \phi(\bar{x}_i))^2 + \lambda \|\bar{w}\|_1,$$

$|w_1| + |w_2| + \cdots$

L_1 regularization

LASSO (least absolute shrinkage and selection operation)

sparse Model

$J(\bar{w}) = l(\bar{w}) + r(\bar{w})$ minimized at w^*

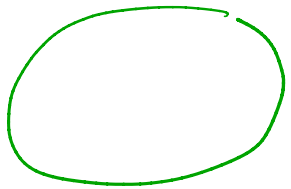
$a = l(\bar{w}^*)$, $b = r(\bar{w}^*)$

$l(\bar{w}) = a$ Contour ①

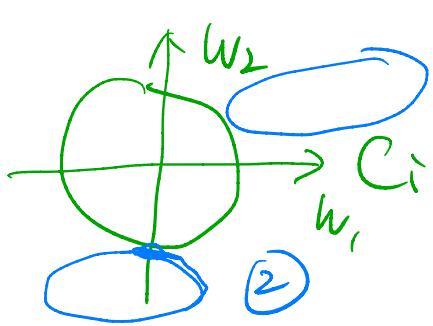
$r(\bar{w}) = b$ Contour ②

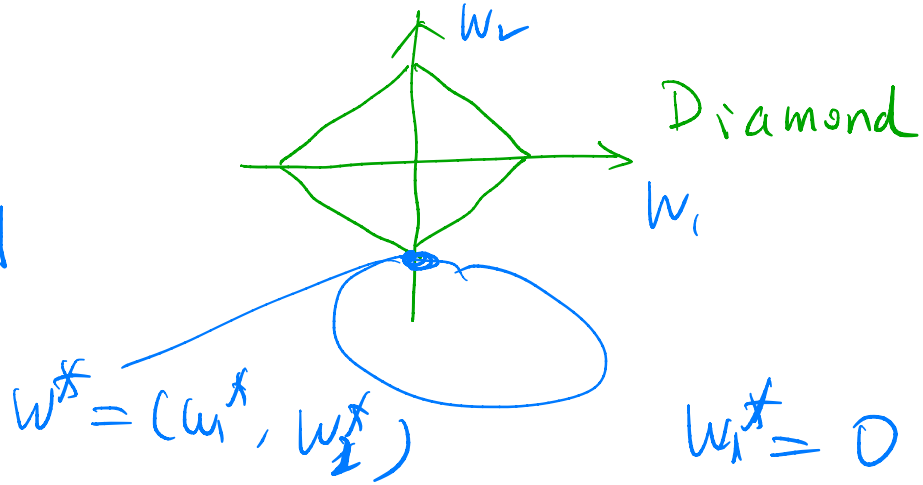
$$\begin{cases} L_2: & w_1^2 + w_2^2 = b \\ L_1: & |w_1| + |w_2| = b \end{cases}$$

w^* is where ① & ② meet



ellipse





Regularization

$$\min \ell(D, Q) + \lambda f(Q)$$

Encourage Q to have certain properties

$\ell(D, Q) + \lambda \|W\|_2^2$ Simple

$\ell(D, Q) + \lambda I(z_1, z_2)$ Disentanglement

latent variables in model

$\ell(D, Q) - \lambda H[Q(Y|x)]$ reduce confidence of result

max entropy

0.99

classes ! . . . .

Linear Regression: $\{\bar{x}_i, y_i\}_{i=1}^N$ $\quad y_i \in \mathbb{R}^1$

$\bar{x} \longrightarrow \bar{w}^T \bar{x} \qquad p(y|\bar{x}) = N(\bar{w}^T\bar{x}, \sigma)$

Logistic Regression: $\{\bar{x}_i, y_i\}_{i=1}^N \qquad y_i \in \{0, 1\}$

Want : $p(y|\bar{x}, \bar{w}) \in [0, 1]$

$\bar{x} \longrightarrow \bar{w}^T\bar{x} \in \mathbb{R}^1$ ← Transformation

$$p(y=1|\bar{x}, \bar{w}) = \sigma(\bar{w}^T\bar{x}) = \frac{1}{1+e^{-\bar{w}^T\bar{x}}}$$

$$p(y=0|\bar{x}, \bar{w}) = 1 - \frac{1}{1+e^{-\bar{w}^T\bar{x}}} = \frac{e^{-\bar{w}^T\bar{x}}}{1+e^{-\bar{w}^T\bar{x}}}$$

$$P(Y=1 \mid \bar{x}, \bar{w}) = \sigma(\bar{w}^T \bar{x}) = \frac{1}{1+e^{-\bar{w}^T \bar{x}}}$$

$$P(Y=0 \mid \bar{x}, \bar{w}) = 1 - \frac{1}{1+e^{-\bar{w}^T \bar{x}}} = \frac{e^{-\bar{w}^T \bar{x}}}{1+e^{-\bar{w}^T \bar{x}}}$$

$p \in [0, 1]$

$$\text{logit}(p) = \log \frac{p}{1-p}$$

$\hookrightarrow \in (-\infty, \infty)$

$$\text{Logit}\left[P(Y=1 \mid \bar{w}, \bar{x})\right] = \log \frac{P(Y=1 \mid \bar{x}, \bar{w})}{P(Y=0 \mid \bar{x}, \bar{w})}$$

$$= \log e^{\bar{w}^T \bar{x}}$$

$$= \bar{w}^T \bar{x}$$

logistic regression : Linear regression for logit

2022-02-0

$$\hat{Y} = 1 \iff P(Y=1 \mid \bar{x}, \bar{w}) > P(Y=0 \mid \bar{x}, \bar{w})$$

$$\iff \frac{P(Y=1 \mid \bar{x}, \bar{w})}{P(Y=0 \mid \bar{x}, \bar{w})} > 1$$

$$\iff e^{\bar{w}^T \bar{x}} > 1$$

$$\bar{x} = [x_1, x_2 \cdots x_D]$$

$$\iff \bar{w}^T \bar{x} > 0$$

$$w_0 + w_1 x_1 + \cdots + w_D x_D > 0$$

Logistic Regression is a

Linear classifier

To learn $\bar{w}$, minimize cross entropy $\ell(\bar{w})$

$$J(\bar{w}) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \bar{x}_i, \bar{w})$$

$$= -\frac{1}{N}\sum_{i=1}^{N} [y_i \log p(y_i=1 \mid \cdots) + (1-y_i) \log (1 - p(y_i=1 \mid \cdots)]$$

⇒ Binary Cross Entropy

$$= -\frac{1}{N}\sum_{i=1}^{N} [y_i \log \sigma(\bar{w}^T \bar{x}_i) + (1-y_i) \log (1 - \sigma(\bar{w}^T \bar{x}_i)]$$

$\log p(y \mid \cdots) = y \log p(y=1 \mid \cdots) + (1-y) \log (1 - p(y=1 \mid \cdots))$

$y=0$ $\log p(y=0 \mid \cdots)$ $0 + \log (1 - p(y=1 \mid \cdots)) = \log p(y=0 \mid \cdots)$

$y=1$ $\log p(y=1 \mid \cdots)$ $\log p(y=1 \mid \cdots) + 0$

Gradient Descent

$$J'(w) = \frac{dJ(w)}{dw} = \lim_{\varepsilon \to 0} \frac{J(w+\varepsilon) - J(w)}{\varepsilon}$$

For small ε, $J'(w) \approx \frac{J(w+\varepsilon) - J(w)}{\varepsilon}$

Approximate J around w $\quad J(w+\varepsilon) = J(w) + \varepsilon J'(w)$ (1)

Reduce J, how to change w according to (1)?

* $J'(w) > 0$, $\varepsilon < 0$, decrease w
* $J'(w) < 0$, $\varepsilon > 0$, increase w

$$w \leftarrow w - \alpha J'(w)$$

α: step size / learning rate

$\bar{w} = (w_0, w_1, \cdots w_D)^T \qquad \bar{x} = (1, x_1, x_2, \cdots, x_D)^T$

$$J(\bar{w}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \sigma(\bar{w}^T \bar{x}_i) + (1 - y_i) \log (1 - \sigma(\bar{w}^T \bar{x}_i)]$$

i-th example $\bar{x}_i = (x_{i,0}, x_{i,1}, \cdots, x_{i,D})$

$$z = \bar{w}^T \bar{x} = w_0 x_0 + w_1 x_1 + \cdots + w_D x_0$$

$$\frac{\partial z}{\partial w_1} = x_1 \qquad \frac{\partial \bar{w}^T \bar{x}}{\partial w_j} = x_j$$

$$\frac{d\sigma(z)}{dz} = -\frac{1}{(1+e^{-z})^2}\frac{d(1+e^{-z})}{dz} \qquad \sigma(z) = \frac{1}{1+e^{-z}}$$

$$= \frac{1}{(1+e^{-z})^2}(-e^{-z})$$

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

$$= \sigma(z)\left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= \sigma(z)(1-\sigma(z))$$

$$\frac{\partial}{\partial w_j}\left[y \log \sigma(z) + (1-y) \log (1-\sigma(z)) \right]$$

$$= \frac{\partial}{\partial z}\left[y \log \sigma(z) + (1-y) \log (1-\sigma(z)) \right] \frac{\partial z}{\partial w_j}$$

$$= \left[y \frac{1}{\sigma(z)} \frac{d\sigma(z)}{dz} + (1-y) \frac{1}{1-\sigma(z)} \frac{d(1-\sigma(z))}{dz} \right] x_j$$

$$= \left[y \frac{1}{\sigma(z)} \sigma(z)(1-\sigma(z)) - (1-y) \frac{1}{1-\sigma(z)} \sigma(z)(1-\sigma(z)) \right] x_j$$

$$= \left[y(1-\sigma(z)) - (1-y)\sigma(z) \right] x_j$$

$$= \left[y - y\sigma(z) - \sigma(z) + y\sigma(z) \right] x_j$$

$$= \left[y - \sigma(z) \right] x_j$$

Update Rule

observed

predicted

$$w_j \leftarrow w_j + \alpha \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)] x_{i,j}$$

$x_{i,j} > 0$, $y_i > \hat{y}_i$ increase $\hat{y}_i$, increase w_j

$[y_i - \sigma(\cdots)] x_{ij} > 0$

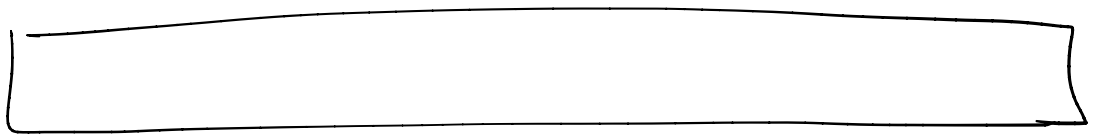
$y_i < \hat{y}_i$ Decrease $\hat{y}_i$, decrease w_j

$[y_i - \sigma(\cdots)] x_{ij} < 0$

$x_{i,j} < 0$, $y_i > \hat{y}_i$, increase $\hat{y}$, decrease w_j

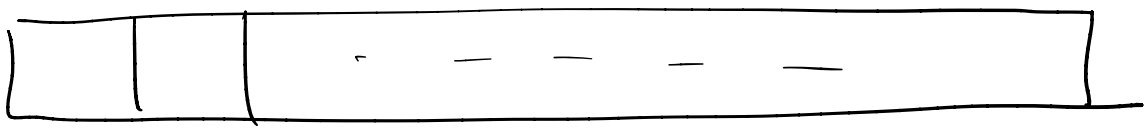
$[y_i - \sigma(\cdots)] x_{ij} < 0$

Stochastic Gradient Descent



$N = 1,000,000$

Random Reshuffle



$|B| = 100$

B_1 B_2

10,000 updates

epoch 1

epoch 2

Softmax Regression

$\{\bar{x}_i, y_i\}_{i=1}^{N}$ $\quad y_i = \{1, 2, \dots C\}$

$\bar{x}$ which class?

10

$\bar{w}_1 \bar{x}$ $\quad e^{\bar{w}_1^T \bar{x}}/Z = P(Y=1|\dots) \in [0,1]$

$\bar{x}$ $\quad \bar{w}_2 \bar{x}$ $\quad e^{\bar{w}_2^T \bar{x}}/Z = P(Y=2|\dots) \in [0,1]$

$\vdots$ $\qquad\qquad \vdots$

$\bar{w}_C \bar{x}$ $\quad e^{\bar{w}_C^T \bar{x}}/Z = P(Y=C|\dots) \in [0,1]$

$$Z = \sum_{c=1}^{C} e^{\bar{w}_c^T \bar{x}}$$

$C=2$, equivalent to logistic Regression

$\bar{w}_c = (w_{c,0}, \dots w_{c,D})$

$\bar{W} = (\bar{w}_1, \dots \bar{w}_C)$

Newton's Method

Find w s.t. $f(w) = 0$

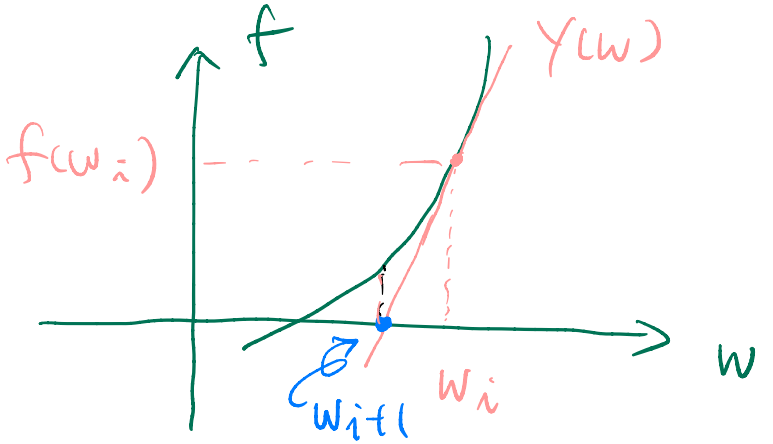
* pick w_0

* Iterate:

$$w_{i+1} \leftarrow w_i - \frac{f(w_i)}{f'(w_i)}$$

Task: find w s.t. $J'(w) = 0$

Let $f(w) = J'(w)$



Approximate f using linear func

$$Y(w) = f'(w_i)\, w + f(w_i) - f'(w_i)\, w_i$$

$$Y(w_i) = f'(w_i)\, w_i + b = f(w_i) = 0$$

$$b = f(w_i) - f'(w_i)\, w_i$$

$$w_{i+1} \leftarrow w_i - \frac{J'(w_i)}{J''(w_i)} \leftarrow \text{2nd order}$$

GD

$$w_{i+1} \leftarrow w - \alpha J'(w_i)$$

$$\bar{w} = [w_0, w_1, \cdots w_D]^T$$

$$\nabla J = \left[\frac{\partial J}{\partial w_0}, \frac{\partial J}{\partial w_1}, \cdots \frac{\partial J}{\partial w_D}\right]^T$$

Gradient vector

Hessian Matrix

$$H = \begin{bmatrix} \frac{\partial^2 J}{\partial w_0 \partial w_0} & \cdots & \frac{\partial^2 J}{\partial w_D \partial w_0} \\ \vdots & & \\ \frac{\partial^2 J}{\partial w_0 \partial w_D} & \cdots & \frac{\partial^2 J}{\partial w_D \partial w_D} \end{bmatrix}$$

Newton's update Rule

$$\bar{w} \leftarrow \bar{w} - H^{-1} \nabla J$$

$O((D+1)^{2.7})$

Best Matrix Inversion Algo

$[f_0(\bar{w}), \cdots, f_D(\bar{w})]^T$

Jacobian Matrix

$$\begin{bmatrix} \frac{\partial f_0}{\partial w_0} & \cdots & \frac{\partial f_0}{\partial w_D} \\ \vdots & & \\ \frac{\partial f_D}{\partial w_0} & \cdots & \frac{\partial f_D}{\partial w_D} \end{bmatrix}$$

Hessian matrix is Jacobian matrix of Gradient Vector

Taylor expansion

$$J(w+\varepsilon) = J(w) + J'(w)\varepsilon + \frac{1}{2!}J''(w)\varepsilon^2 + \frac{1}{3!}J'''(w)\varepsilon^3 + \cdots$$

1st order approximation

$$J(w+\varepsilon) \approx J(w) + J'(w)\varepsilon \Rightarrow \text{Gradient Descent Algo}$$

2nd order approximation

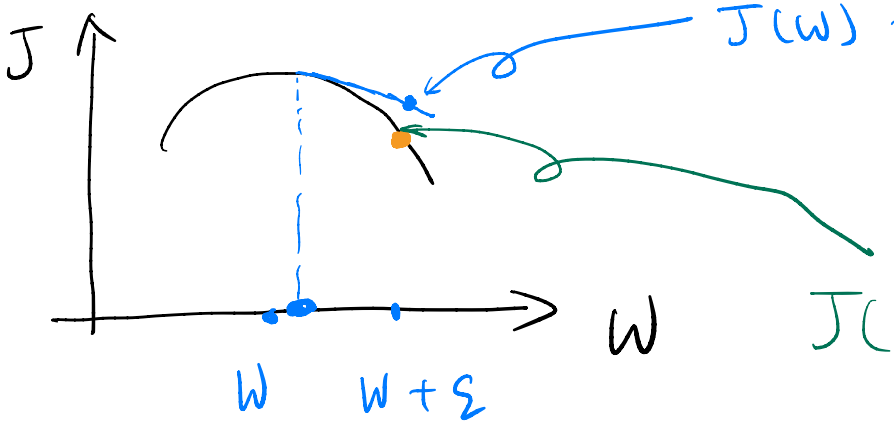
$$J(w+\varepsilon) \approx J(w) + J'(w)\varepsilon + \frac{1}{2}J''(w)\varepsilon^2 = f(\varepsilon)$$

$$\frac{df}{d\varepsilon} = J'(w) + J''(w)\varepsilon = 0$$

$$\Rightarrow \varepsilon = -\frac{J'(w)}{J''(w)}$$

Newton's method

Advantage of Newton's Method



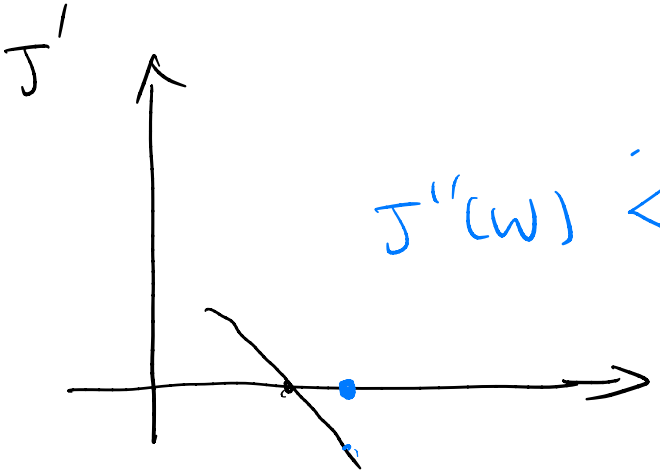
$J(w) + J'(w)\varepsilon$ over-estimate

$J(w+\varepsilon)$

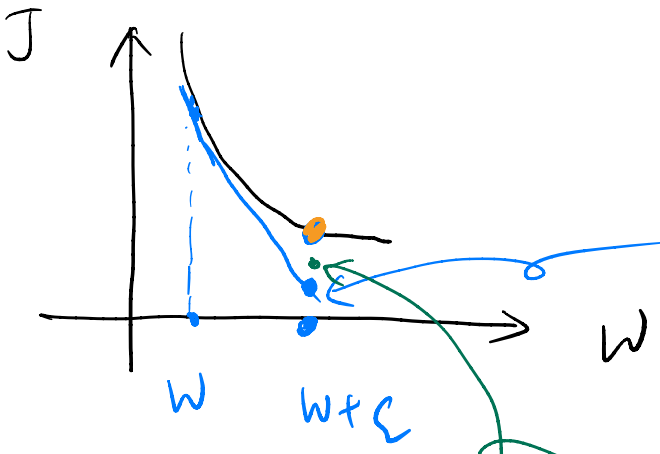
$J(w) + J'(w)\varepsilon + \frac{1}{2} J''(w)\varepsilon^2$

< 0

correct the above over-estimation



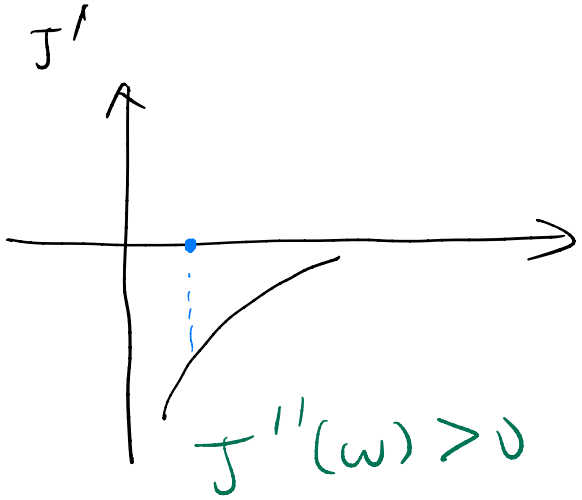
$J''(w) < 0$



$J(w) + J'(w)\,\varepsilon$

under-estimate

$J(w+\varepsilon)$



$J(w) + J'(w)\,\varepsilon + \frac{1}{2} J''(w)\,\varepsilon^2$

Corrects the above

under estimation

Optimization Approach

$D = \{\bar{x}_i, y_i\}_{i=1}^{N} \qquad y_i \in \{-1, +1\}$

w_0

Linear classifier: $\hat{y} = \text{sign}(\underbrace{w^T \bar{x} + b}_{z}) = \begin{cases} +1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$

zero/one loss

$$L_{0/1}(y_i, \hat{y}_i) = \begin{cases} 1 & \hat{y}_i \neq y_i \\ 0 & \hat{y}_i = y_i \end{cases}$$

$$L(\bar{w}, b) = \frac{1}{N} \sum_{i=1}^{N} L_{0/1}(y_i, \hat{y}_i)$$ Empirical loss

Objective $\arg\min_{\bar{w}, b} L(\bar{w}, b)$

$$\hat{Y} = \text{sign}(\underbrace{W^T \bar{X} + b}_{z}) = \begin{cases} +1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$$

$Y \in \{-1, +1\}$

$$L_{0/1}(Y_i, \hat{Y}_i) = \begin{cases} 1 & \hat{Y}_i \neq Y_i \\ 0 & \hat{Y}_i = Y_i \end{cases}$$

observed, predicted

$$\mathbb{1}(Y_i z_i \leq 0)$$

$$= \begin{cases} 1 & \left\{ Y_i = 1,\ z_i \leq 0,\ \hat{Y}_i = 0, -1 \right\} \quad Y_i \neq \hat{Y}_i \\ & \left\{ Y_i = -1,\ z_i \geq 0,\ \hat{Y}_i = 1, 0 \right\} \\ 0 & \end{cases}$$

$$Y_i = \hat{Y}_i$$

$$= L_{0/1}(Y_i, \hat{Y}_i)$$

$$L(\bar{w}, b) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i z_i \leq 0)$$

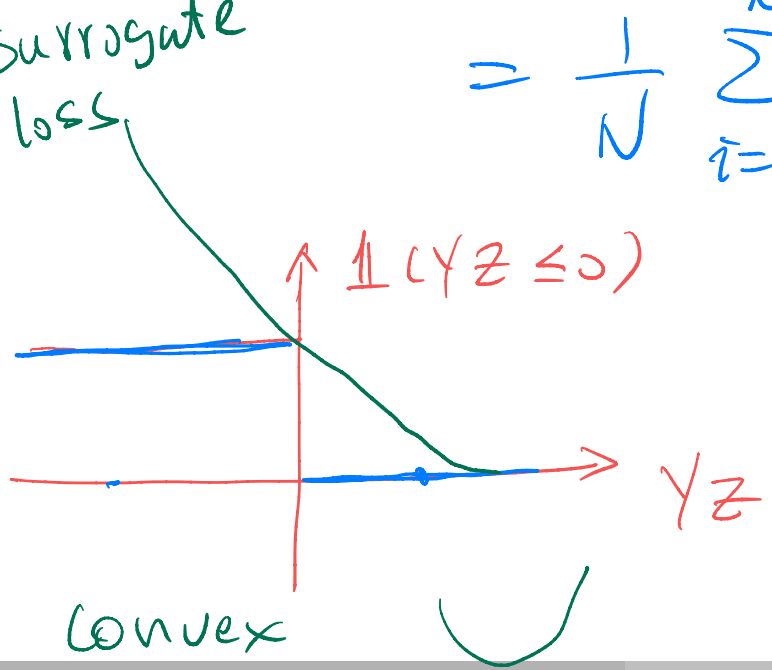
$\bar{w} = [w_1, \cdots, w_D]^T$

$$\frac{\partial L}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbb{1}(y_i z_i \leq 0)}{\partial w_j}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathbb{1}(y_i z_i \leq 0)}{\partial y_i z_i} \frac{\partial y_i z_i}{\partial w_j}$$

$= 0$

cannot do GD on empirical loss



Want

$$\min \frac{1}{N} \sum_{i=1}^{N} L_{0/1}(y_i, \hat{y}_i)$$

training error

Test error

Do

$$\min \frac{1}{N} \sum_{i=1}^{N} L_{\text{surrogate}}(y_i, z_i)$$

loss

loss