1. Bayes Classifier

(i) Probability of Class 0, 1

(ii) conditional densities

$\downarrow$

Posterior and implement Bayes classifier (Perfect information).

2. Data — $(X_i, Y_i)_{i=1}^{n}$

Generative

$\downarrow$

Estimates priors and conditional densities.

Discriminative.

$\downarrow$

Posterior estimation using ERM.

(i) K-NN

(ii) Naive Bayes

(i) Linear regression for classification

(ii) logistic regression

Discriminative:

$$R(h) = \mathop{E}_{\substack{X, Y \\ \in \mathbb{R}^d,\ \in\{0,1\}}} \left[L(Y, h(X)) \right] \leftarrow \text{Risk of a classifier.}$$

$$h^* = \min_h E_{X,Y} \left[L(Y, h(X)) \right].$$

$$D = \{ (x_i, y_i)_{i=1}^n \}.$$

$$\hat{R}(h) = \frac{\sum_{i=1}^{n} L(Y_i, h(x_i))}{n}.$$

(ii) $\mathcal{H} = \{ w^t x : w \in \mathbb{R}^d \}$

$$h(x) = w^t x.$$

(iii) classification based on linear regression

$$L(Y, h(x)) = \mathbb{1}_{\{Y \neq h(x)\}}.$$

Indicator loss function.

$$L_{lin}(Y, h(x)) = (Y - h(x))^2$$

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - w^t x_i)^2.$$

$$w^* = \min_{w \in \mathbb{R}^d} J(w).$$

(a) $\nabla J(w) = 0 \implies w^* = (X^T X)^{-1} X^T y.$

$$X = \begin{pmatrix} x_1^t \\ x_2^t \\ \vdots \\ x_n^t \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \begin{matrix} \to +1 \text{ or } -1 \\ \to +1 \text{ or } -1 \\ \\ +1 \text{ or } -1 \end{matrix}$$

(b) $w_0 = \begin{pmatrix} 0 \\ \vdots \end{pmatrix}$

(o)

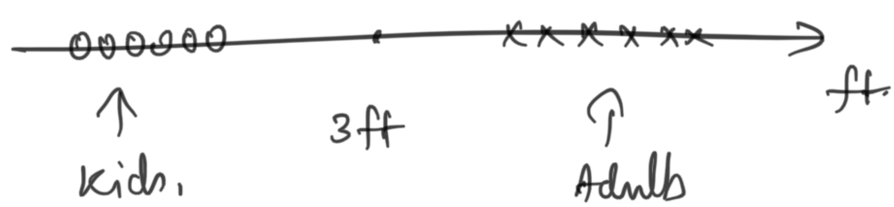
$$w_{n+1} = w_n - \alpha \nabla J(w_n)$$

(i) Batch gradient

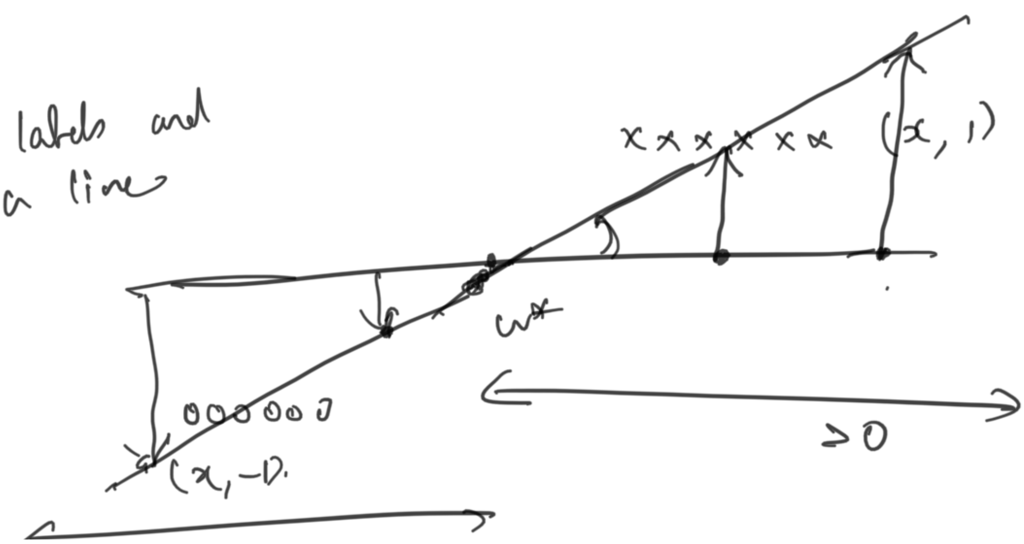
(ii) Stochastic gradient

(iii) mini batch gradient.

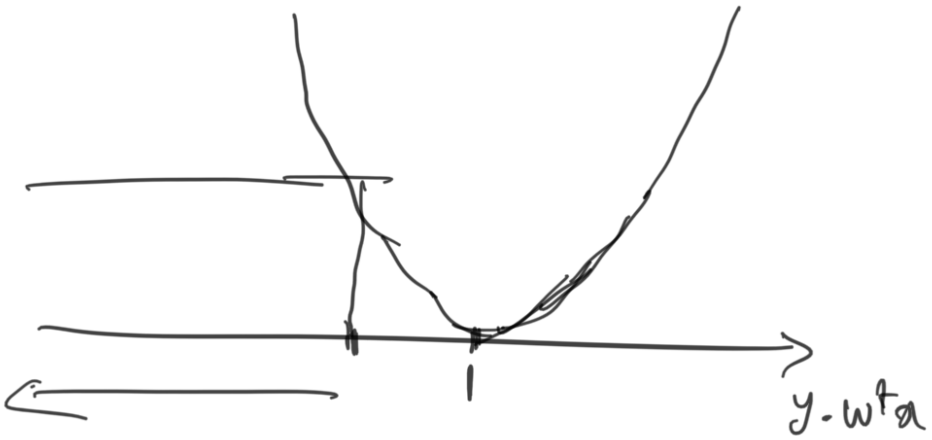
1.

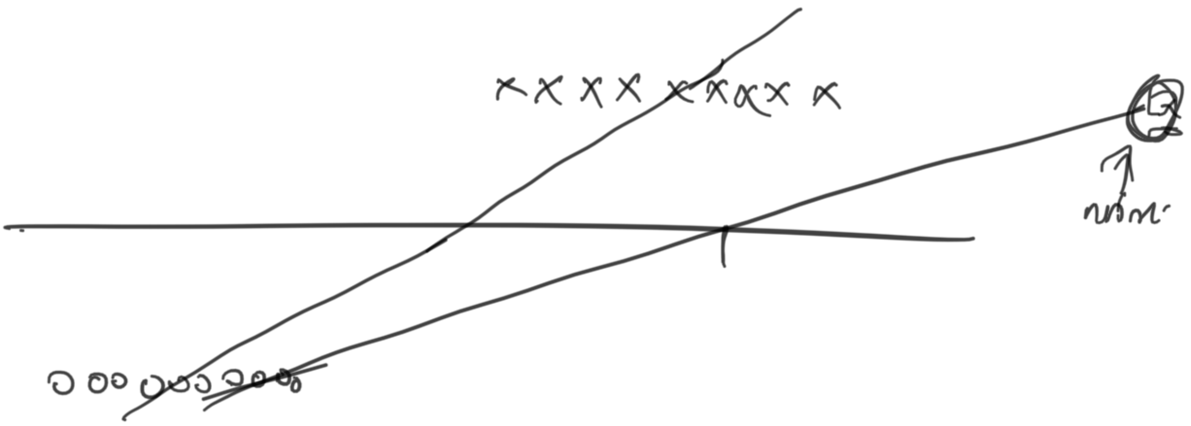


with y labels and fitted a line



$$h(x) = \begin{cases} +1 & \text{if } w^{*t}x > 0 \\ -1 & \text{if } w^{*t}x < 0 \end{cases}$$

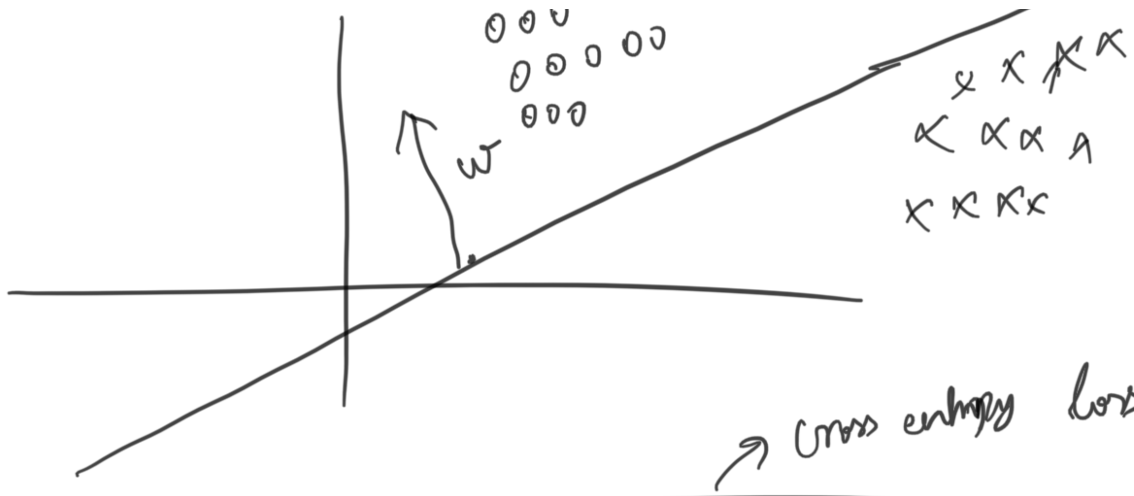




<u>Dis adv:</u>
Linear regression for classification is not suitable when there is outlier in the present

<u>Adv:</u> Simplicity and closed form answer.

$\{ w^t x \}$.



cross entropy loss

$$\boxed{L(y, h(x)) = y \log h(x) + (1-y) \log (1-h(x))}$$

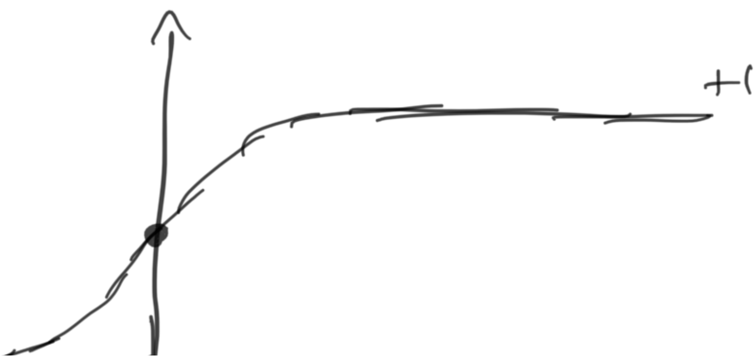
$$h(x) = \frac{1}{1+e^{-(w^t x)}} = \sigma(w^t x).$$

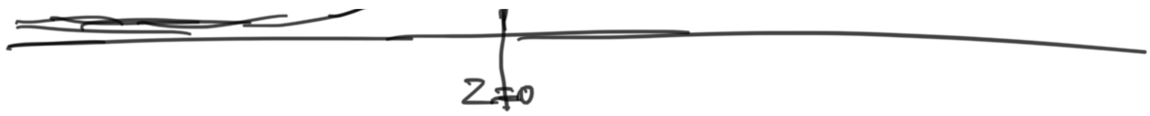
$$h : \overset{\mathcal{X}}{\mathbb{R}^d} \longrightarrow \mathbb{R}$$

↓ score function.

$$h : \mathbb{R}^d \longrightarrow [0 \quad 1]$$

$\sigma \longrightarrow$ Sigmoid function $\quad \sigma(z) = \frac{1}{1+e^{-z}}$,

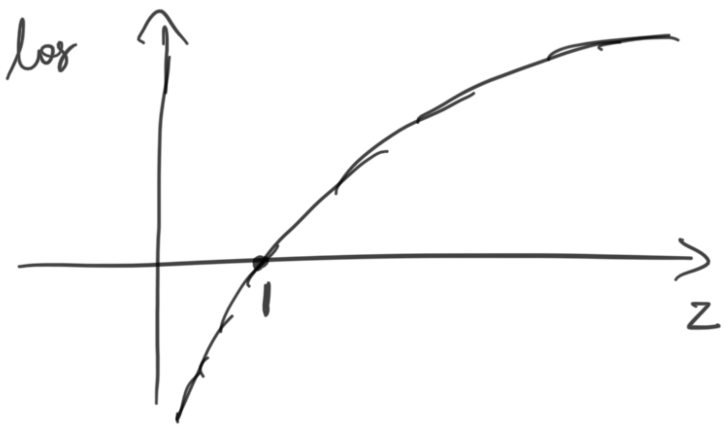


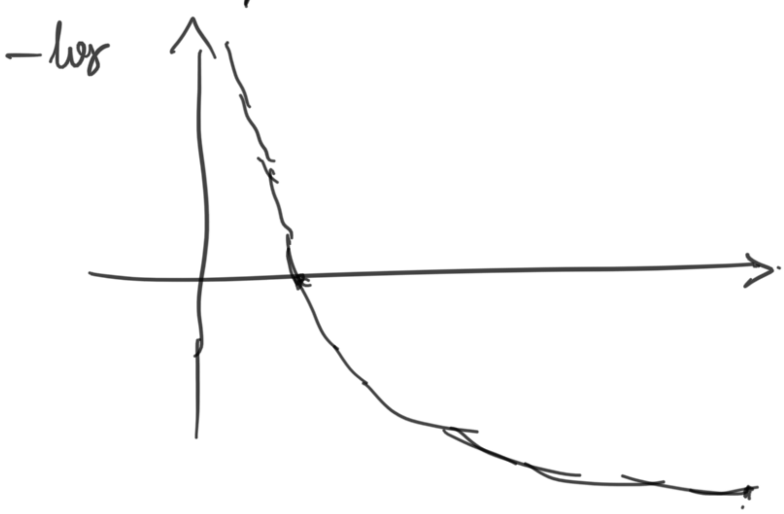


$$L(y, h(x)) = \underline{y \log h(x)} + \underline{(1-y)\log(1-h(x))}$$

$$h(x) = \frac{1}{1+e^{-w^t x}}$$

$$\log h(x) = -\log(1+e^{-w^t x})$$





$x \in$ class 1

$w^t x >>>>> 0$

$x \notin$ class 1

$w^t x << < 0$

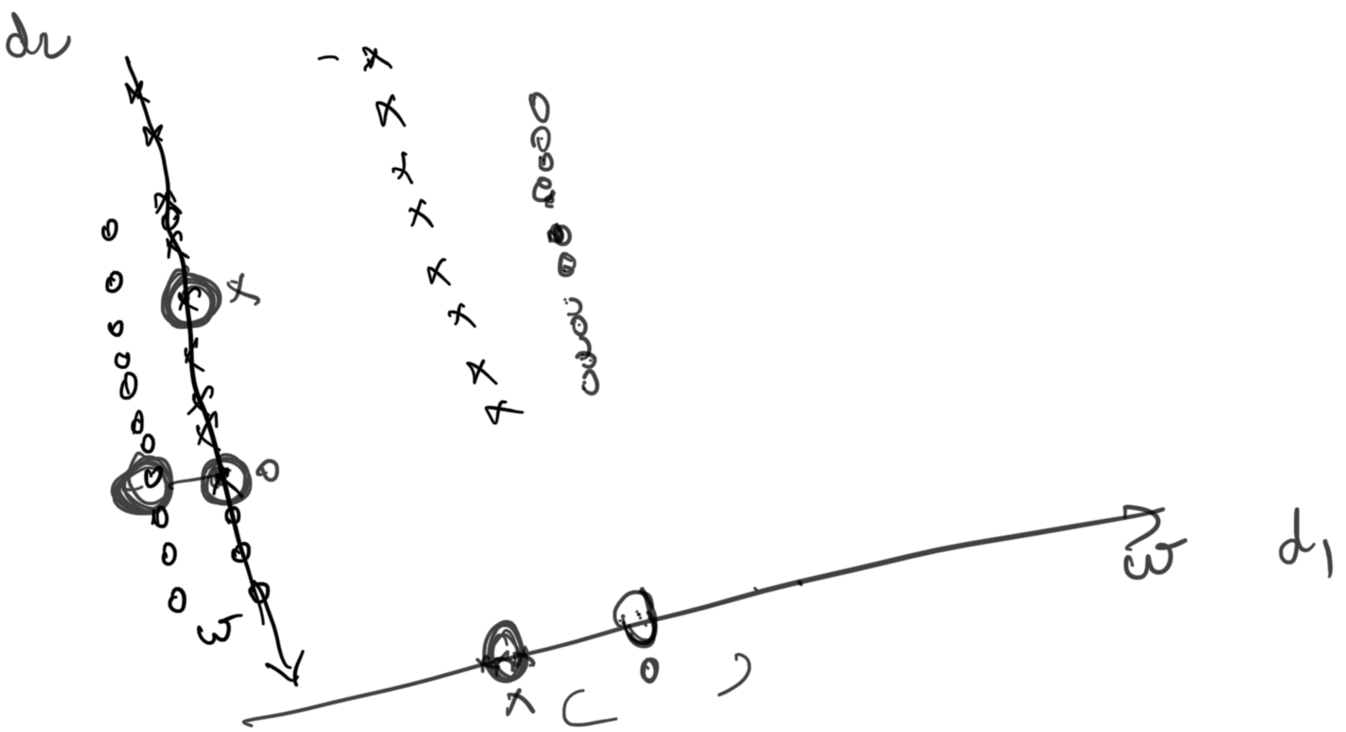
$$L(y, h(x)) =$$

$$-y \log(1+e^{-w^t x}) + (1-y) \log \frac{e^{-w^t x}}{\quad}$$

$$1+e^{-w^t x}$$

$$-\left[y \log\left(1+e^{-w^t x}\right) + (1-y) \log\left(1+e^{w^t x}\right)\right]$$

FLDA : Fisher linear discriminant analysis.



$$C_0 = \{ i : y_i < 0 \}$$

$$C_1 = \{ i : y_i > 0 \}$$

$w^T x_i$

$w^t x_i$

$w_1 x_1 + w_2 x_2 + \dots + w_d x_d$

$$m_0 = \sum_{i \in C_0} w^t x_i \qquad m_1 = \sum_{i \in C_1} w^t x_i$$

$$\overline{\quad\quad} \qquad \frac{\quad}{|C_0|} \qquad\qquad \frac{\quad}{|C_1|}.$$

$$s_0^2 = \frac{\sum_{i \in C_0} (w^t x_i - m_0)^2}{|C_0|} \longrightarrow \text{Variance of class } 0$$

$$s_1^2 = \frac{\sum_{i \in C_1} (w^t x_i - m_1)^2}{|C_1|} \longrightarrow \text{Variance of class } 1$$

$$J(w) = \frac{(m_1 - m_0)^2}{s_0^2 + s_1^2}$$

$$= \frac{w^t S_B w}{w^t S_W w}$$

$$S_B = \underset{d \times 1}{(M_1 - M_0)} \; \underset{1 \times d}{(M_1 - M_0)^T} \longleftarrow d \times d \text{ Matrix between class}$$

$$M_1 = \frac{\sum_{i \in C_1} x_i}{|C_1|} \qquad M_0 = \frac{\sum_{i \in C_0} x_i}{|C_0|}$$

$M_1 \,,\, M_0 \in \mathbb{R}^d$

$$S_W = \sum_{i \in C_0} (x_i - M_0)(x_i - M_0)^T + \sum_{i \in C_1} (x_i - M_1)(x_i - M_1)^T$$

$$J(w) = \frac{w^t S_B w}{w^t S_W w}$$

$$w^* = \max_{w} J(w)$$

$$\boxed{w^* \propto S_W^{-1}(M_1 - M_0)}$$

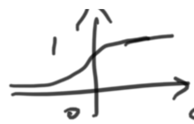
Duda & Hart
Pattern classification
chapter 3

Logistic regression:

$h : \mathbb{R}^d \longrightarrow \mathbb{R} \xrightarrow{\sigma} [0 \;\; 1]$

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

$$h(x) = w^t x$$



sum

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

$$D = \{(x_i, y_i)_{i=1}^{n}\}$$

Christopher Bishop
Chapter 3/4
Linear models.

$$P\left(Y \begin{smallmatrix}=0\\=-1\end{smallmatrix} \mid x\right) = 1 - \frac{1}{1+e^{-w^tx}} = 1-\sigma(w^tx)$$

$$= \frac{e^{-w^tx}}{1+e^{-w^tx}} = \frac{1}{1+e^{w^tx}}$$

$\{0, 1\}$
$\{-1, 1\}$

$$\boxed{\underbrace{P(Y=y_1 \mid x_1)}_{\text{Posterior}} = \frac{1}{1+e^{-y_1 w^t x_1}}}$$

$$P(Y=1 \mid x_1) = \frac{1}{1+e^{-w^tx_1}}$$

$$P(Y=-1 \mid x_1) = \frac{1}{1+e^{w^tx_1}}$$

$$P(D) = \prod^{n} P(y_i \mid \quad)$$

$$P(D) = \prod_{i=1}^{n} P(y_i \mid x_i)$$

$$J(w) = \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i w^t x_i}}$$

$$w^* = \max_{w} J(w) = \max_{w} P(D)$$

$$\tilde{J}(w) = \log J(w) = \log \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i w^t x_i}}$$

$$= \sum_{i=1}^{n} -\log\left(1 + e^{-y_i w^t x_i}\right)$$

$$l(w) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i w^t x_i}\right) = -\tilde{J}(w)$$

Logistic regression:

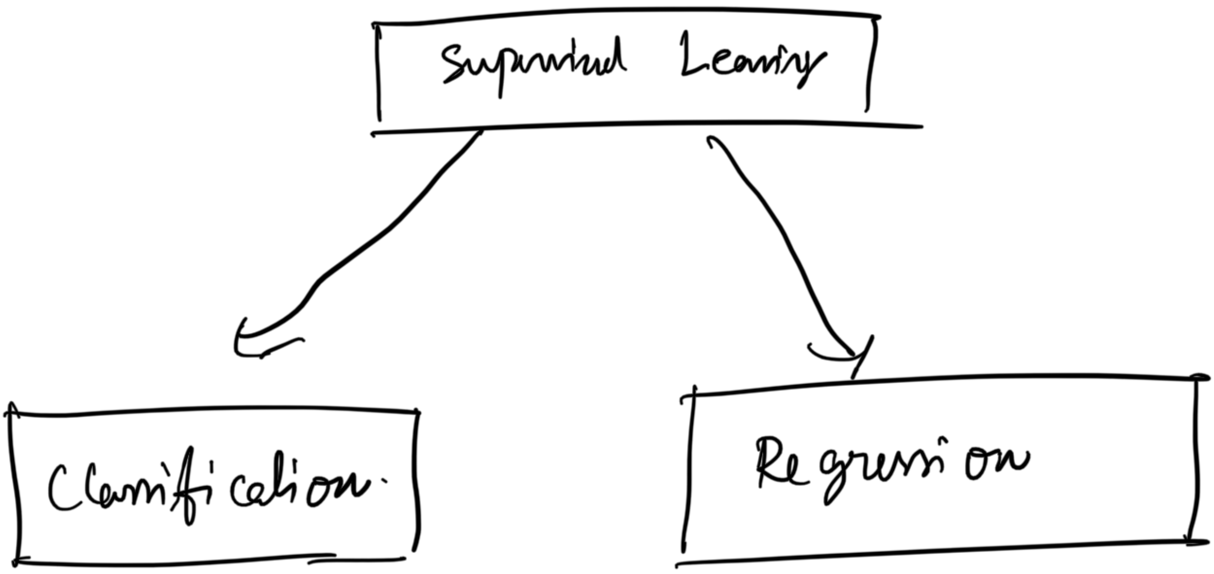
$$\boxed{l(w) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i w^t x_i}\right) \quad y_i \in \{-1, 1\}}$$

$$L(w) = -\sum_{i=1}^{n} y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))$$

$$L(w) = -l(w) \qquad y_i \in \{0, 1\}$$

Classifiers : (Discriminative)

1. Linear regression — closed form, iterative
2. Logistic regression — iterative.
3. FLDA — closed form



$y = \{0, 1\}$ $\qquad$ $y \in \mathbb{R}$

(x_1, y_1) $\qquad$ $x_1, y_1, x_2, y_2, x_3, y_3,$

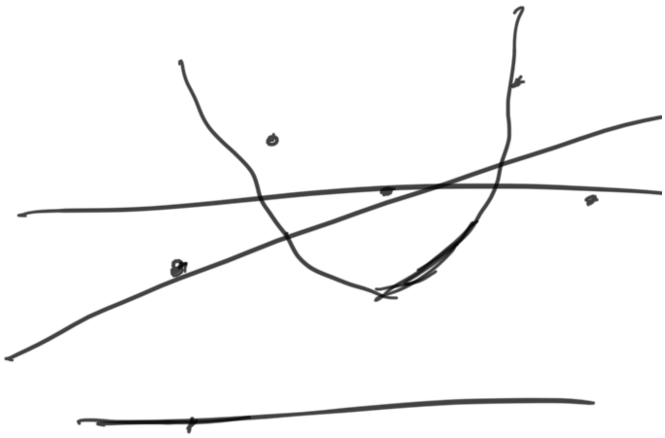
(x_2, y_2) $\qquad$ $x_4, y_4, x_5, y_5 \in \mathbb{R}^1$

(x_3, y_3)

(x_4, y_4)

(x_5, y_5)

(x, y)



$$\hat{y} = \underline{w_1} x + \underline{w_0} \longrightarrow \text{Model } M=1$$

$$\hat{y} = w_2 x^2 + w_1 x + w_0 \rightarrow M=2$$

$$\hat{y} = w_3 x^3 + w_2 x^2 + w_1 x + w_0$$

$M=3$.

$$y = w_0 \qquad M=0$$

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \Longleftarrow J(w)$$