Bayes Classifier:

1. Prior probabilities of class 0 and class 1

P_0 P_1

$$P(Y=0) \quad , \quad P(Y=1)$$

2. Likelihood / Class conditional densities.

$f_0(x)$ $f_1(x)$

$$P(X \mid Y=0) \quad , \quad P(X \mid Y=1)$$

Class conditional densities.

Distribution of features

given a class.

3. Loss function.

$$L(Y, h(x))$$

↓ ↓

True Label — Predicted Label

$$L(Y, h(x)) = 1 \quad \text{if } Y \neq h(x)$$

$$= 0 \quad \text{otherwise.}$$

(Zero - one Loss function)

Perfect information case

Bayes classifier

$$P(Y=1|x) \quad , \quad P(Y=0|x)$$

$$q_1(x) \qquad\qquad q_0(x)$$

$$q_1(x) = \frac{p_1 f_1(x)}{p_0 f_0(x) + p_1 f_1(x)}$$

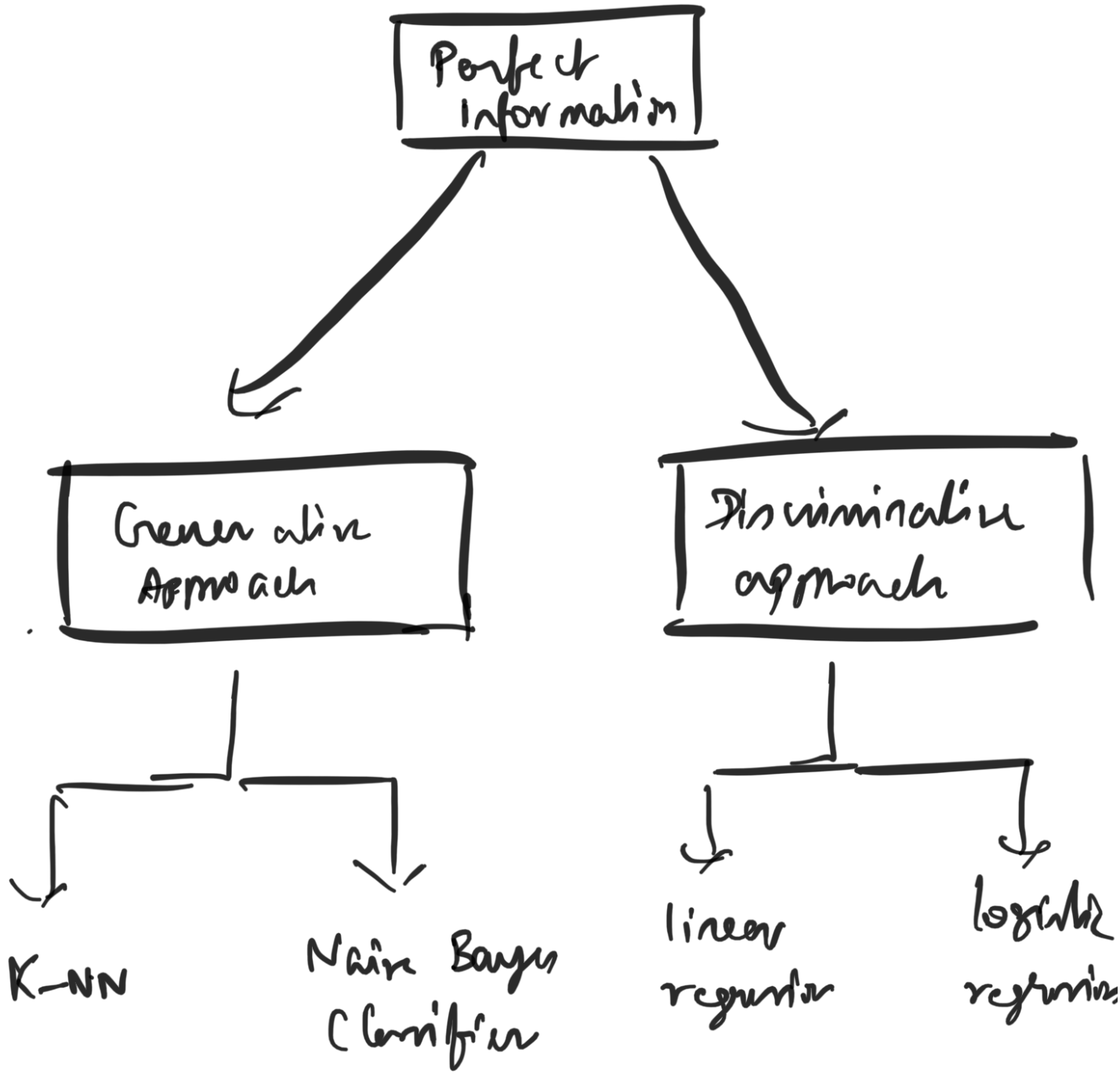
$$q_0(x) = \frac{p_0 f_0(x)}{p_0 f_0(x) + p_1 f_1(x)}$$

if $q_1(x) > q_0(x)$

$$h_B(x) = 1$$

else $q_1(x) \leq q_0(x)$

$h_B(x) = 0$



Generative Approach

1. Estimate prior probabilities from data
2. Estimate class conditional densities from data.
3. Using these estimated quantities, if we implement a Bayes Classifier, then we call the approach Generative.

Discriminative approach.

Directly approximate posterior from the data rather than estimating

Prior and class conditional densities

$$R(h) = \underset{X,Y}{E}\left[L(Y, h(X)) \right]$$

↳ Risk of a classifier.

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

A surrogate function to be optimized

$$h^* = \min_h \frac{\sum_{i=1}^{n} L(Y_i, h(X_i))}{n}$$

(Loss function)

↑

Empirical risk minimization.

Data

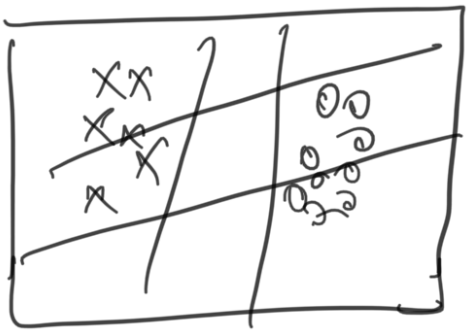
h → 5ft than Adult
less Kid

5ft Adult ✗

5 ft	Kid			
6.ft	Adult	6 ft	Adult	✓
4.5 ft	Kid	4.5ft	Kid	✓
4.5 ft	Adult	4.5ft	Kid	✗

$\mathcal{H} = \{ w^t x :$ Linear in the features$\}$

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



$\mathcal{X}$ feature space.

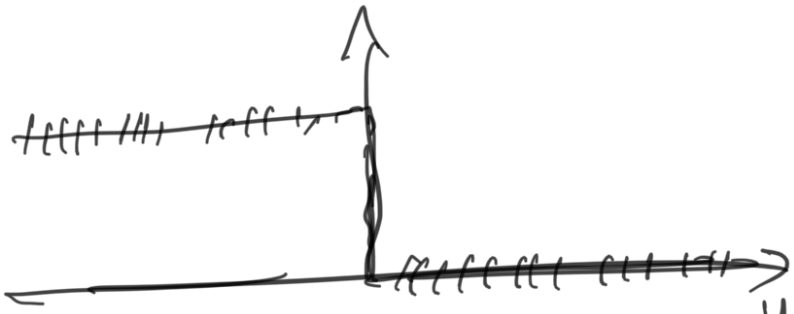
$\longrightarrow$ Classifier

1. $\underline{E} \longrightarrow$ average over the data given to us.

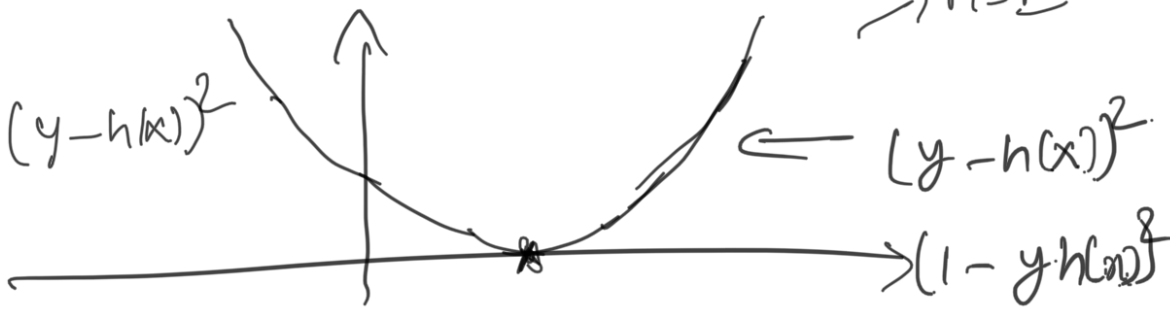
2. Instead of considering all possible classifiers, we will consider a

restricted hypothesis space.

3. Nature of Loss function



→ MSE



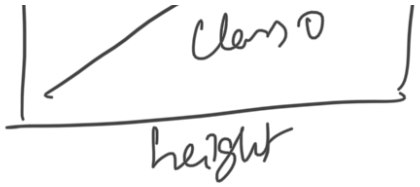
$y \quad y h(x)$

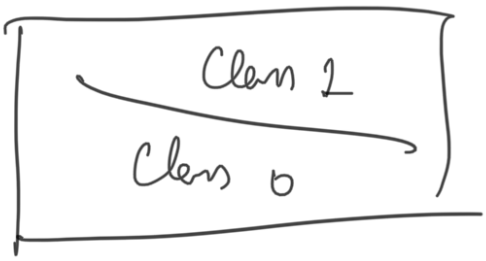
$h(x)$ → classifier prediction

y → true label

2.

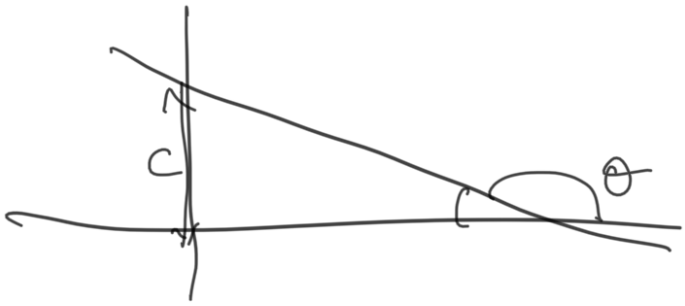
2





$\{ w_1 x_1 + w_2 x_2 + \ldots + w_d x_d \}$

$\{ mx + c \}$.



$h : \mathcal{X} \longrightarrow \mathbb{R}$

$h(x) \longrightarrow$ Score.

$h(x) > 0$ Class 1

$h(x) < 0$ class -1

$h(x) > \frac{1}{2}$ Class 1

$h(x) < \frac{1}{2}$ Class 0

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

Objective :

$$\min_{w} \frac{\sum_{i=1}^{n} (w^t x_i - y_i)^2}{n}$$

$$X = \begin{pmatrix} x_1^t \\ x_2^t \\ \\ \\ \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \\ \\ \end{pmatrix}$$

$$\begin{pmatrix} x_n^T \end{pmatrix} \qquad \begin{pmatrix} y_n \end{pmatrix}$$

$$\underset{n \times d \quad d \times 1}{Xw} = \begin{pmatrix} x_1^T w \\ x_2^T w \\ \\ x_n^T w \end{pmatrix}_{n \times 1} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \\ y_d \end{pmatrix}_{n \times 1}$$

$$\frac{1}{n} \| Xw - Y \|^2 \iff \sum_{i=1}^{n} \frac{(w^T x_i - y_i)^2}{n}$$

$$f(w) = \frac{1}{n} \| Xw - Y \|^2$$

$$= \frac{1}{n} (Xw - Y)^T (Xw - y)$$

$$= \frac{1}{n} (w^T X^T - y^T)(Xw - y)$$

$$= \frac{1}{n} (w^T X^T X w - y^T X w - w^T X^T y + y^T y)$$

$$= \frac{1}{n}\left(w^T \underbrace{(X^T X)}_{d \times d} w - \underbrace{2 (y^T X)}_{1 \times n \; n \times d} w + y^T y \right)$$

$(X^T X)$: $d \times d$ → Gram matrix.

→ quadratic

$d = 1$

$$f(w) = \underbrace{w^T A w} + b^T w + c \qquad f : \mathbb{R}^d \to \mathbb{R}$$

$$A = \underset{d \times d}{X^T X} \qquad c = \underset{1 \times 1}{y^T y}$$

$$b = \underset{1 \times d}{-2 y^T X}$$

$$\nabla f(w^*) = 0$$

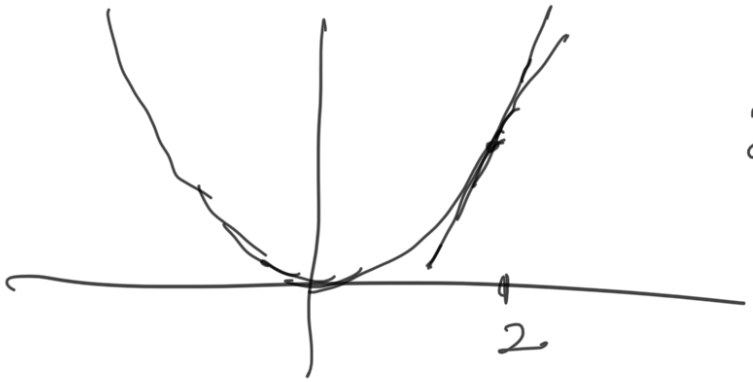
$$2Aw + b = 0$$

$$w = -\frac{1}{2} A^{-1} b$$

$$\boxed{w^* = \underbrace{(X^T X)^{-1}}_{d \times d} y^T X}$$

we found the minima of the loss function

Gradient descent Algorithm:



$y = x^2$

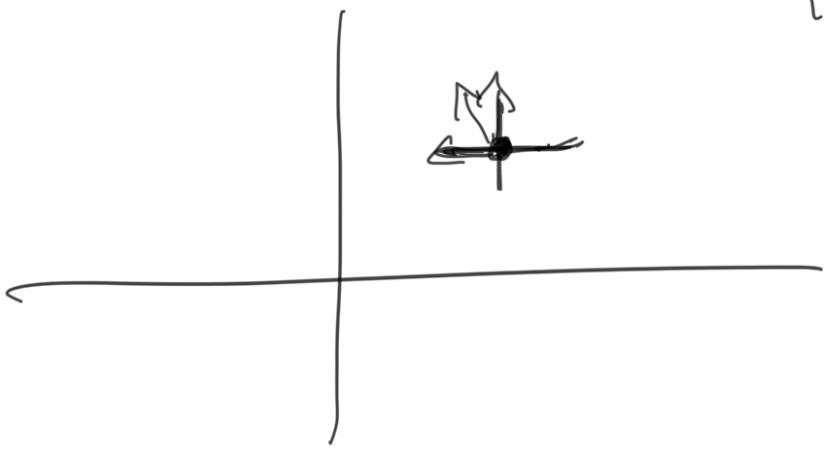
$2x > 0$

$w_0 \longrightarrow w_1$

$$w_1 = w_0 - \alpha \left(\frac{dl}{dw}\right)$$

Slope

$$\begin{pmatrix} w_1(1) \\ w_1(2) \\ \\ w_1(d) \end{pmatrix} = \begin{pmatrix} w_0(1) \\ w_0(2) \\ \\ w_0(d) \end{pmatrix} - \alpha \begin{pmatrix} \frac{\partial f}{\partial w_0(1)} \\ \frac{\partial f}{\partial w_0(2)} \\ \\ \frac{\partial f}{\partial w_0(d)} \end{pmatrix}$$



$\frac{\partial f}{\partial x} > 0$

$\frac{\partial f}{\partial y} < 0$

$$f(w) = \|Xw - y\|^2 = w^T A w + b^T w + c$$

$$\nabla f(w) = 2Aw + b$$

$$w_{t+1} = w_t - \alpha (2A w_t + b)$$

$$w_t \to w^*$$

Stochastic gradient descent:

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

$$= \frac{1}{n} \sum_{i=1}^{n} L_i$$

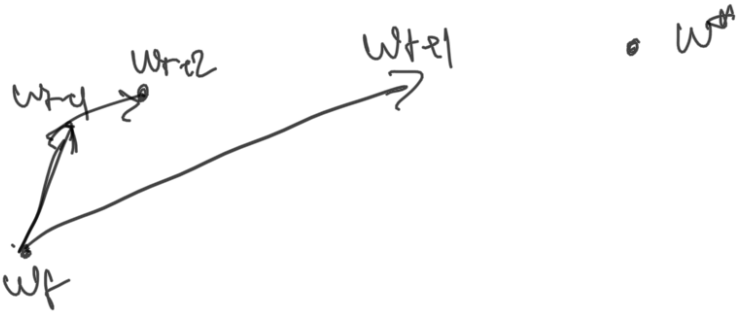
$$L_i = (w^T x_i - y_i)^2$$

$$\boxed{w_{t+1} = w_t - \alpha_t \boxed{(w_t^T x_i - y_i) x_i}}$$

$$w_t \to w^*$$

error: (scalar

$$w_{t+1} = w_t - \alpha_t \underbrace{\left(w_t^T x_i \right)}_{\text{Prediction}} - \underbrace{y_i}_{\text{target}} x_i$$



Linear regression based classification

1. Loss function — MSE

2. $f(w) = \sum_{n=1}^{n} (w^T x_i - y_i)^2$

Algorithm for finding best w:

1. $\nabla f(w^*) = 0$

$$w^* = (X^T X)^{-1} X^T y$$

$d \times d \qquad O(d^2)$

2. Iterative way:

$$w_{t+1} = w_t - \alpha_t (A w_t + b).$$

3. Stochastic gradient descent

$$w_{t+1} = w_t - \alpha (w_t^T x_i - y_i) x_i$$

4. Mini batch gradient descent: