Bayes Classifier

2 - class problems.

Ex: D - Default

R - Return

Feature:

Ex: H - High risk loaners

L - Low risk loaners.

$\mathcal{X}$ - Feature space.

$x \in \mathcal{X}$ is called a feature.

Prior probabilities:

$P(C_0)$ → probability of class 0

$P(C_1)$. → probability of class 1

$Y = 0$ → class 0

$Y = 1 \longrightarrow$ class 1.

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

$\downarrow$ $\downarrow$

P_0 P_1

Likelihood or the class - conditional densities:

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

$\downarrow$ $\downarrow$

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

This gives distribution of the feature conditioned on a class.

Ex: $P(H \mid D)$, $P(L \mid D) \rightarrow$ class 0

$P(H \mid R)$, $P(L \mid R) \rightarrow$ class 1

Posterior probabilities:

$P(Y = 1 \mid X = x)$ $P(Y = 0 \mid X = x)$

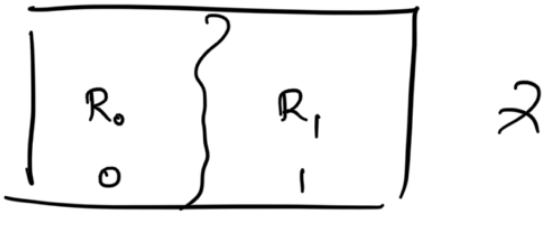
$q_1(x)$ ↓ $q_0(x)$ ↓

Ex: $P(D|H)$, $P(R|H)$
$P(D|L)$, $P(R|L)$
} Can be computed using Bayes rule.

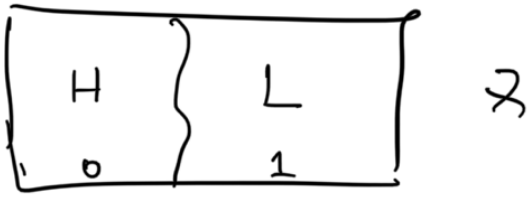
$$q_0(x) = \frac{P_0 f_0(x)}{P_0 f_0(x) + P_1 f_1(x)} \qquad q_1(x) = \frac{P_1 f_1(x)}{P_0 f_0(x) + P_1 f_1(x)}$$

Bayes classifier rule:

$$h_B(x) = 1 \quad \text{if} \quad q_1(x) > q_0(x)$$
$$= 0 \quad \text{if} \quad q_1(x) < q_0(x)$$



Ex:



Bayes Rule:

$$\text{Posterior} = \frac{\text{likelihood}}{\text{evidence}} \times \text{Prior}$$

likelihood — How likely is the feature given a class.

Prior — What is the prior probabilities for each class

Posterior — what is the posterior probability of a class given a feature value

evidence — How likely is the feature.

$$err(h_B) = P(y=1, h_B(x)=0) + P(y=0, h_B(x)=1).$$

$$= P(y=1)\, P(h_B(x)=0 \mid y=1) + P(y=0)\, P(h_B(x)=1 \mid y=0)$$

$$= P(y=1)\, P(x \in R_0 \mid y=1)$$

$$+ P(Y=0)\ P(x \in R_1 \mid Y=0)$$

What is a classifier?

A classifier is a rule or a mapping from feature space $\mathcal{X} \longrightarrow Y = \{0, 1\}$.

$$err(h) = P(h(x) = 0, Y=1) + P(h(x)=1, Y=0)$$

h_B is the optimal classifier because it has the least error.

Height of the population is your feature

Adult, child } Two classes.

$P(\text{Adult}) = 0.5$

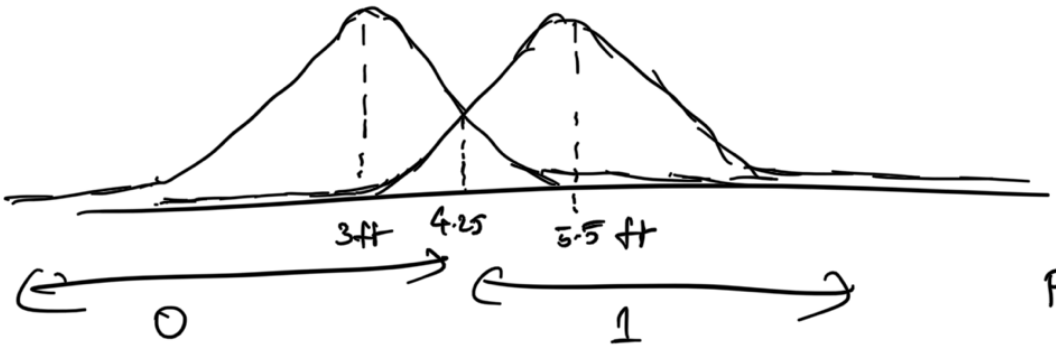
$\mu_1 = 5.5$ ft $\qquad \sigma_1 = 1$

$\mu_0 = 3$ ft $\qquad \sigma_0 = 1$

$$P(\text{child}) = 0.5$$

$$f_0(x) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$$

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right).$$



$P(Y=1 \mid h) \rightarrow q_1(h)$

$P(Y=0 \mid h) \rightarrow q_0(h)$

$$\frac{q_1(x)}{q_0(x)} > 1 \implies \text{class Adult}$$

$$\frac{q_1(x)}{q_0(x)} < 1 \implies \text{class child.}$$

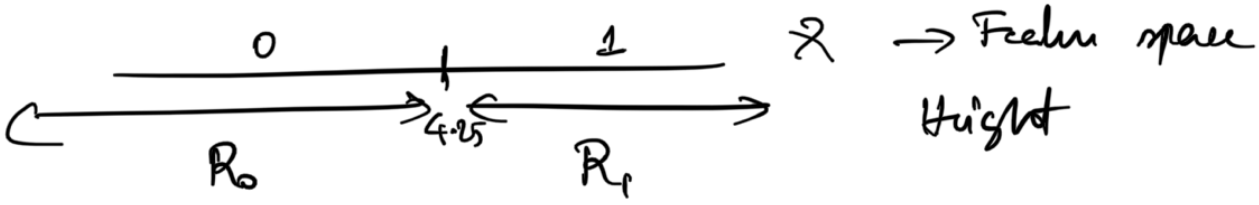
$$q_1(x) = \frac{P_1 f_1(x)}{}$$

$P_0 f_0(x) + P_1 f_1(x)$

$$q_0(x) = \frac{P_0 f_0(x)}{P_0 f_0(x) + P_1 f_1(x)}$$

$$\frac{q_1(x)}{q_0(x)} = \frac{P_1 f_1(x)}{P_0 f_0(x)} = \frac{f_1(x)}{f_0(x)} > 1$$

$\Rightarrow$ Class 1



$$R_0 = \{ x : x < 4.25 \}$$

$$R_1 = \{ x : x \geq 4.25 \}$$

$$h_B(x) = 0 \quad \text{if} \quad x \in R_0$$

$$= 1 \quad \text{if} \quad x \in R_1$$

When can more errors weights matter?.

Loss function :

$L : Y \times Y \longrightarrow \mathbb{R}$

(True, Prediction)

$L(0, 0)$ → Loss incurred when classifier says 0 class and the true class is 0

$L(0, 1)$ → Loss incurred when classifier says 1 class and the true class is 0.

$L(1, 0)$ → "

$L(1, 1)$ → "

Zero / one loss function

$L(0, 0) = 0$

$L(1, 0) = 1$

$L(0, 1) = 1$

$L(1, 1) = 0$

$\underset{x,y}{E}\left[L(y, h(x))\right]$ → Expected loss or Risk of a

Classifier.

$$R(h) = \underset{x,y}{E}\left[L(y, h(x)) \right].$$

$$= L(0,0)\, P(y=0, x \in R_0)$$

$$+ L(0,1)\, P(y=0, x \in R_1)$$

$$+ L(1,0)\, P(y=1, x \in R_0)$$

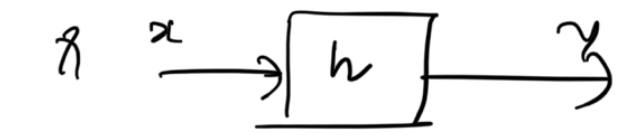
$$+ L(1,1)\, P(y=1, x \in R_1)$$

Enemy aircraft problem:

$$L(0,1) = 1$$

$$L(1,0) = 10$$

What is a classifier?



Feature space

Label space.

How do we measure the performance of a classifier?

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

h_B^* = Bayes classifier is the one that minimizes the risk.

Perfect information

1. prior probabilities
2. Class conditional densities
3. Loss function $\longrightarrow$ Zero / one loss function

↓

Bayes classifier

↓

optimal.

Generative

1. Naive Bayes
2. K - Nearest Neighbours

Discriminative

1. Logistic regression
2. SVM (Support vector machines).
3. Linear discriminant analysis

Data: $(x_i, y_i)_{i=1}^{n}$ ⟶ Dataset.

Ex! $n = 6$

1. $x_1 = 7, \quad y = 1$
2. $x_2 = 3, \quad y = 0$
3. $x_3 = 3.5, \quad y = 0$
4. $x_4 = 5, \quad y = 0$
5. $x_5 = 6, \quad y = 1$

6. $x_6 = 2, \quad y = 0$

What is Generative approach for classifier design?

1. Estimate prior probabilities from data
2. Estimate class conditional densities from data.

K - Nearest neighbour classifier:

$x \rightarrow$ feature

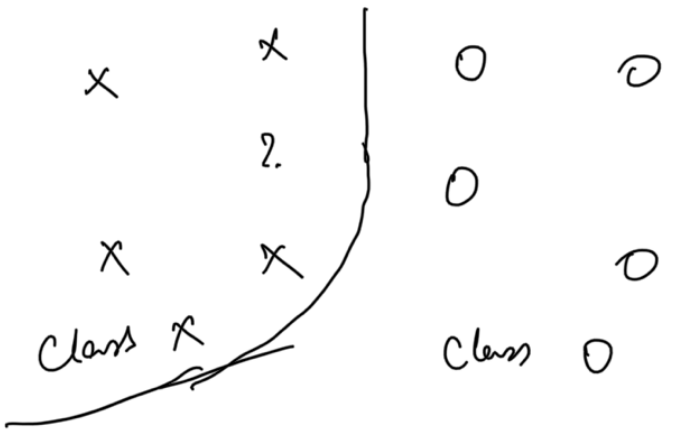
What is the label corresponding to this?

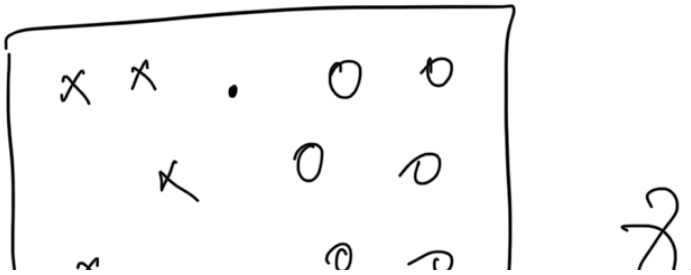
1. Given a feature x find the K closest prototype to this example, or this feature
2. The class label will then correspond

to majority of the labels in the K prototypes.

Two important components:

1. distance metric.
2. How many prototypes to choose?





Voronoi regions
or voronoi diagram

1. Distance metric
   — Euclidean metric

$$x = (x_1, x_2, \ldots, x_d)$$

$$y = (y_1, y_2, \ldots, y_d)$$

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \ldots + (x_d-y_d)^2}$$

2. $K \rightarrow$ odd number

   $K = 3, 5$ etc.

$K =$ nn Classifier error cannot be more than $2 *$ Bayes error.

Ex: Height, weight

(7.ft, 10 kg) $x = (x_1, x_2)$

$y = (y_1, y_2)$

$$d(x, y) = \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2}} + \sqrt{\frac{(x_2 - y_2)^2}{\sigma_2^2}}$$

1.	(0.1, 100)	(−0.3, −3)
2.	(0.2, 101)	(−0.2, −2)
3.	(0.3, 104)	(−0.1, 1)
4.	(0.7, 99)	(0.3, −4)
5.	(0.6, 105)	(0.2, 2)

(0.15, 102)

$$\sigma_i^2 = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{n} \qquad \mu_i = \sum_{i=1}^{n} \frac{x_i}{n}$$

$h(x) \longrightarrow \{0, 1\}$

$h(x) \longrightarrow$ Score.

$h : \mathbb{R}^2 \to \mathbb{R}$

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

$+1 \quad h(x) > 0$

$-1 \quad h(x) < 0$

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

Example:

$H \longrightarrow D$

1. $L \longrightarrow D$

$H \longrightarrow D$

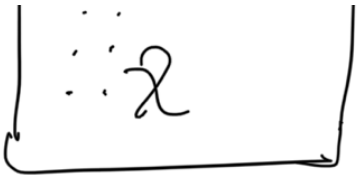
2. $L \longrightarrow R$

$H \longrightarrow R$

3. $L \longrightarrow D$

$H \longrightarrow R$

4. $L \longrightarrow R$



$\mathcal{H}$ = { Hypothesis space }

$\mathcal{H}$ = { $w^t x$, $w \in \mathbb{R}^d$ }.

$w^t x = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_d x_d$

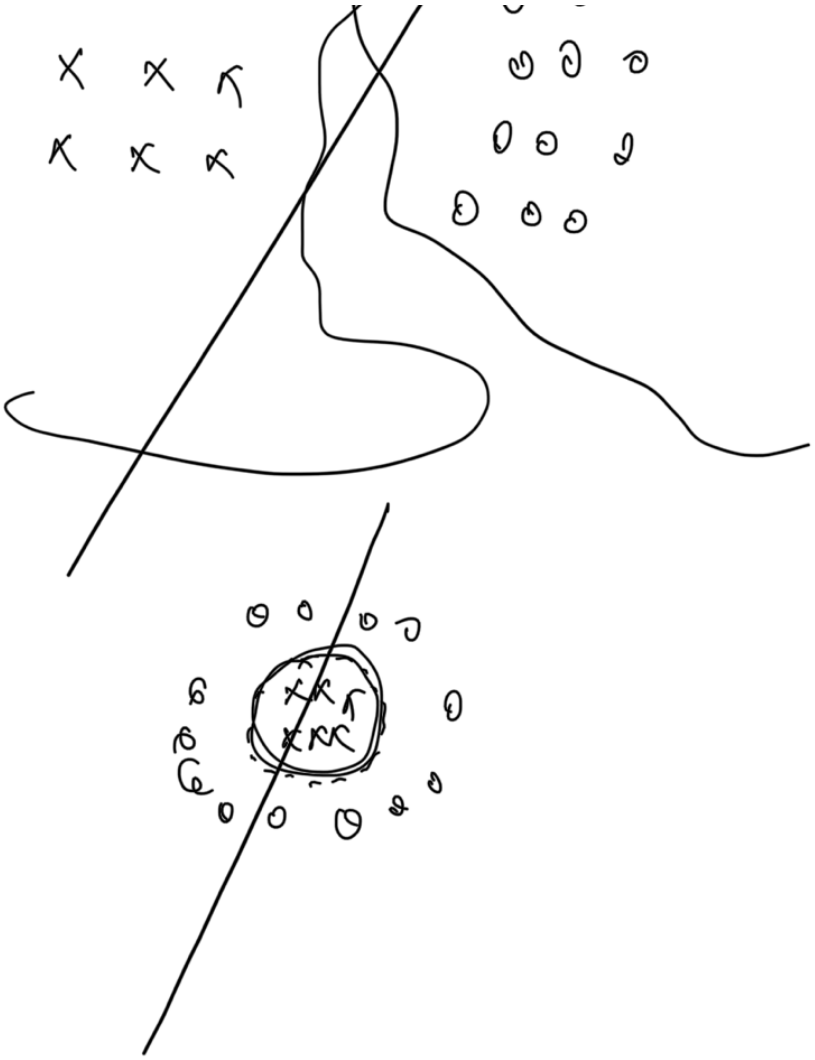
Discriminative Modeling:

1. $h : \mathcal{X} \rightarrow \mathbb{R}$
2. We are going to restrict the space of hypothesis from all possible mapping from $\mathcal{X} \rightarrow \mathbb{R}$ to a restricted class

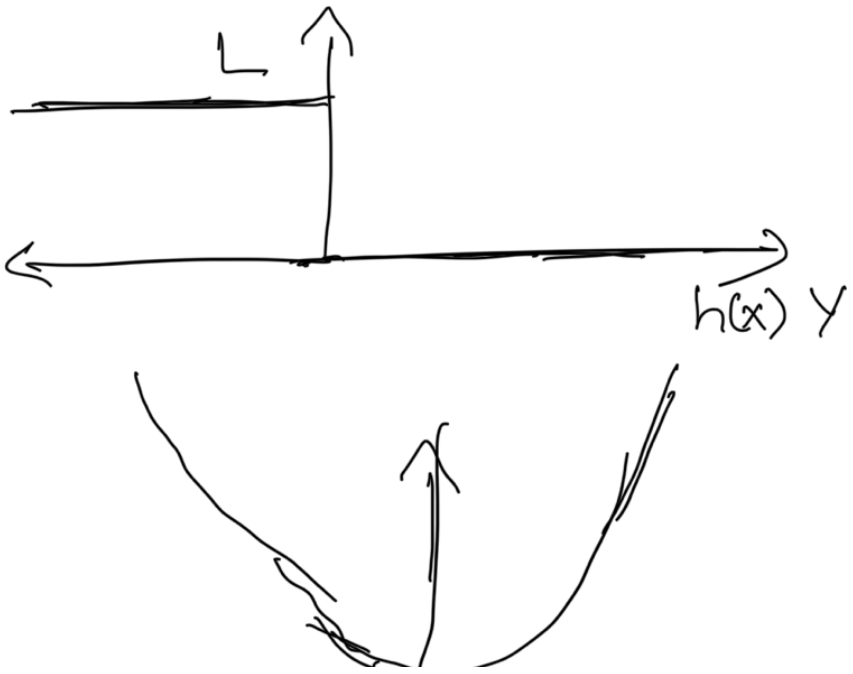
$\{ w^t x \}$ $w \in \mathbb{R}^d$

$\mathcal{X} = \mathbb{R}^2$

$\mathbb{R} \times \mathbb{R}$



$$R(h) = E_{x,y}\left(L(h(x), y)\right)$$



→ Squared Loss function


h(x) Y