Bayes Decision Theory (BDT)

- BDT is a framework for making optimal decisions on problems involving uncertainty.
- Statistical approach to pattern classification.

Framework

1) <u>World</u> has states / classes, drawn from a r.v. Y.

   e.g. $Y \in \{H, T\}$, $Y \in \{A, B, C, D, F\}$, $Y \in \{ok, flu\}$

   <u>prior</u>: $p(Y)$ - prior probability of a state occurring in the world.

2) <u>observer</u> measures features / observations from r.v. X.

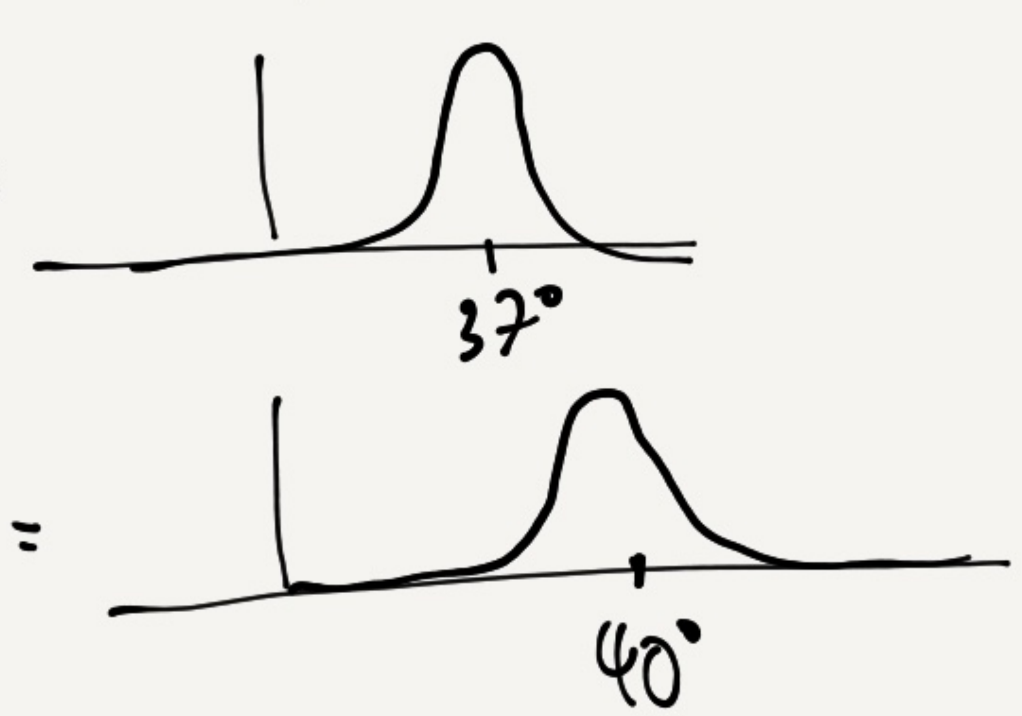
   - <u>class conditional densities</u> (CCDs)

     $p(X|Y)$ - observations conditioned on the state y.

   e.g. $Y \in \{ok, flu\}$, X = temperature

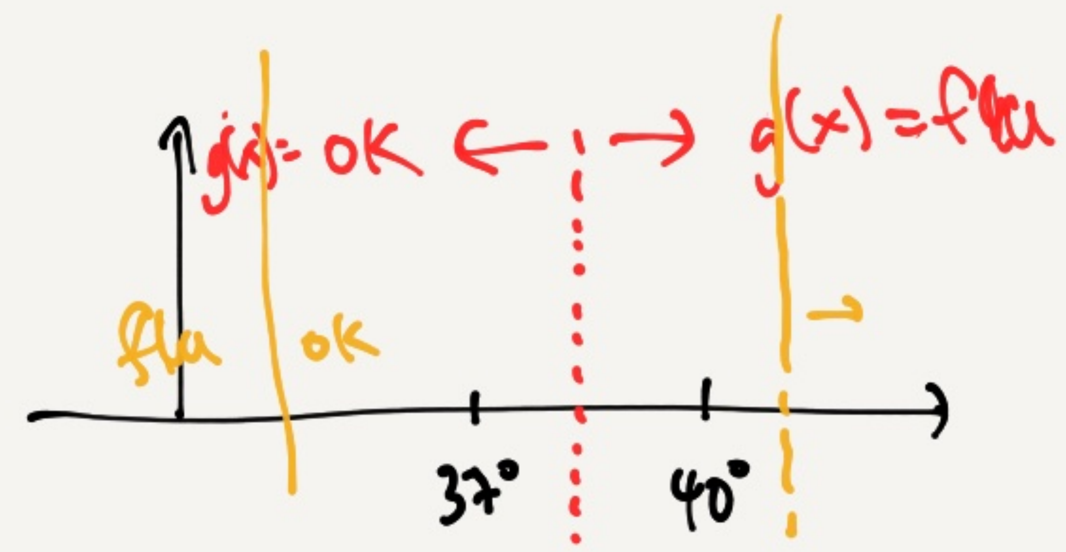
$p(X|ok) =$

$p(x|flu) =$



3) <u>decision function</u>. - use observation to make a decision about the state of the world.

   $g(x): X \to Y$



4) <u>Loss function</u> - penalty for deciding the wrong y or making the wrong decision.

   $L(g(x), y) \geq 0$, e.g. 0-1 loss function

   $$L(g(x), y) = \begin{cases} 0, & g(x) = y \\ 1, & \text{otherwise.} \end{cases}$$

<u>Goal:</u> Find the optimal $g^*(x)$ for the given assumptions. (loss function, CCD, prior)

CS5487 Lecture Notes (2020)
Dr. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Bayes Decision Rule (BDR)

Risk - expected value of the loss function

$$\text{Risk} = E_{X,Y}\left[L(g(X), Y) \right]$$

$$= \sum_y \int_x \underbrace{p(x,y)}_{p(y|x)p(x)} L(g(x), y)\, dx$$

$$= \int_x \sum_y p(y|x)\, p(x)\, L(g(x), y)\, dx$$

$$= \int_x p(x) \underbrace{\left[\sum_y p(y|x) L(g(x), y) \right]}_{\text{conditional Risk } R(x) \text{ of a particular } x.} dx$$

$$= E_X\left[R(x) \right] \leftarrow \text{expectation of conditional risk.}$$

Minimizing the Risk can be achieved by minimizing the conditional risk for each x.

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

For an x

$$g^*(x) = y^* = \underset{j \in Y}{\operatorname{argmin}} \; R(x)$$

$$= \underset{j \in Y}{\operatorname{argmin}} \sum_y p(y|x) L(j, y)$$

our decision.

$$g^*(x) = \underset{j \in Y}{\operatorname{argmin}} \; E_{Y|X}\left[L(j, y) \right]$$

conditional expectation of the loss function

$\Uparrow$

"Bayes Decision Rule"

0-1 loss function for Classification

Classes: $y \in \{1, \dots, C\}$

$$L(g(x), y) = \begin{cases} 1, & g(x) \neq y \\ 0, & \text{otherwise} \end{cases}$$

← misclassification by $g(x)$.

Conditional Risk:

$$R(x) = E_{y|x}[\underbrace{L(g(x), y)}_{\text{indicator function}}]$$

expectation of indicator is the prob of the thing happening

$$= \underbrace{P_r(g(x) \neq y \mid x)}_{\text{probability of an error given } x.}$$

$$= \sum_{y \neq g(x)} p(y|x) = 1 - \underbrace{p(y = g(x) \mid x)}_{\text{prob. of correct}}$$

Thus minimizing $R(x)$ is equivalent to minimizing the probability of making an error.

BDR

$$y^* = \arg\min_j 1 - p(y=j|x)$$

← minimize conditional risk

$$y^* = \arg\max_j p(y=j|x)$$

← choose j that has highest posterior probability.

Bayes Rule

$$y^* = \arg\max_j p(x|y=j)\, p(y=j)$$

$$\boxed{g^*(x) = \arg\max_j \log p(x|y=j) + \log p(y=j)}$$

Simple example

2-class problem $\{0, 1\}$

$$g(x) = \arg\max_j \log p(x|y=j) + \log p(y=j)$$

pick class 0 if:

$$\log p(x|0) + \log p(0) > \log p(x|1) + \log p(1)$$

$$\log p(x|0) - \log p(x|1) > \log p(1) - \log p(0)$$

$$\underbrace{\frac{p(x|0)}{p(x|1)}}_{\text{likelihood ratio}} > \frac{p(1)}{p(0)} = T$$

Summary

- for 0-1 loss function
  - → BDR is MAP (pick maximum posterior)
  - → conditional risk = prob. of error for x
  - → Risk = prob of error.
  - → BDR minimizes prob. of error (no other decision rule is better!)

Caveat: all the modeling assumptions are correct. (CCD & prior are correct)

This is a generative classification model

- CCD are learned from data, decision rule is computed from CCDs.

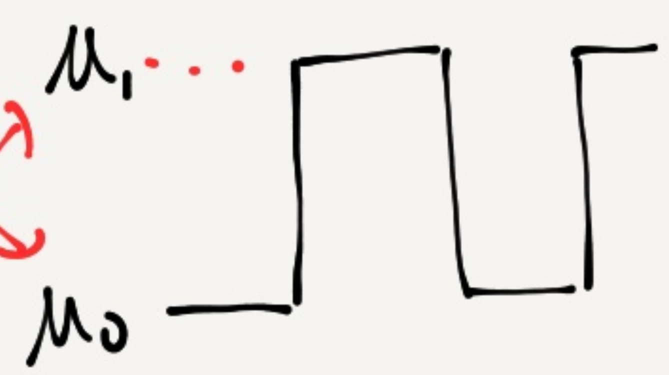
Example: Noisy Channel

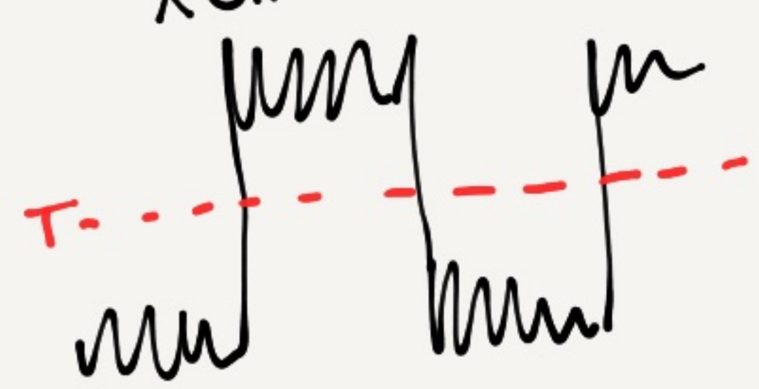
bit $Y \in \{0,1\} \rightarrow$ [transmitter] $\rightarrow$ [noisy channel] $\rightarrow$ [reciever] $\rightarrow$ decode

$X \in \mathbb{R}$

$$y = \begin{cases} 0, & x < T \\ 1, & x \geq T \end{cases}$$

transmitted values for y





What is threshold T?

Given measurment X, recover bit Y.

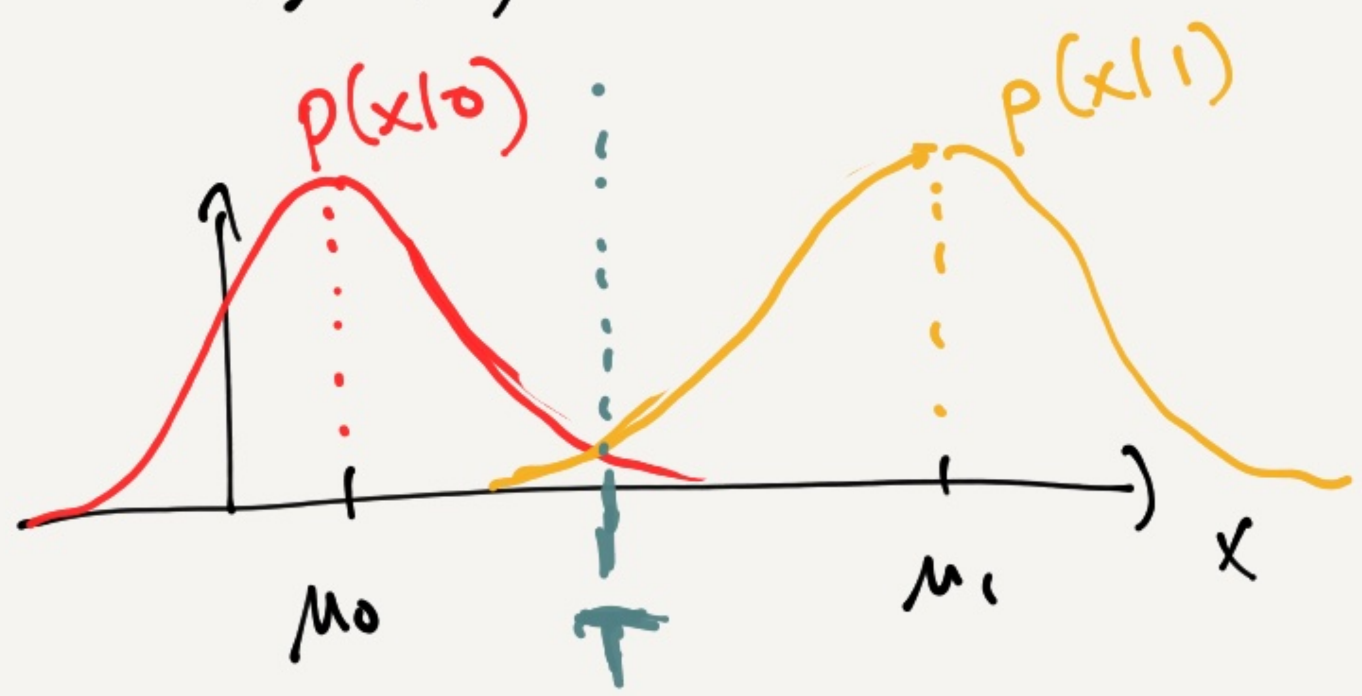
- prior: $p(y=0) = p(y=1) = 1/2$
- CCD: Assume Gaussian additive noise

$$X = \mu_Y + \epsilon \quad , \quad \epsilon \sim N(0, \sigma^2)$$

mes. xmitted value channel noise

$$p(x|y=0) = N(x|\mu_0, \sigma^2)$$
$$p(x|y=1) = N(x|\mu_1, \sigma^2)$$

Assume $\mu_0 < \mu_1$



BDR w/ 0-1 loss

$$y^* = \underset{j}{\arg\max} \log p(x|j) + \log p(j)$$

$$= \underset{j}{\arg\max} -\frac{1}{2\sigma^2}(x-\mu_j)^2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 + \log\frac{1}{2}$$

constant constant

$$= \underset{j}{\arg\max} -(x-\mu_j)^2$$

$$= \underset{j}{\arg\max} -(x^2 - 2x\mu_j + \mu_j^2)$$

constant

$$= \underset{j}{\arg\min} \mu_j^2 - 2x\mu_j$$

Choose $y^* = 0$ when

$$\mu_0^2 - 2x\mu_0 < \mu_1^2 - 2x\mu_1$$

$$2x\mu_1 - 2x\mu_0 < \mu_1^2 - \mu_0^2$$

$$2x(\mu_1 - \mu_0) < \mu_1^2 - \mu_0^2$$

$$\Rightarrow \boxed{x < \frac{\mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)} = \frac{\mu_1 + \mu_0}{2}}$$

threshold is the midpoint between μ_1 & μ_0

Assumptions are explicit

1) 0-1 loss BDR
2) uniform classpriors
3) Gaussian additive noise

What p(Y) is not uniform?

BDR:

pick 0 if:

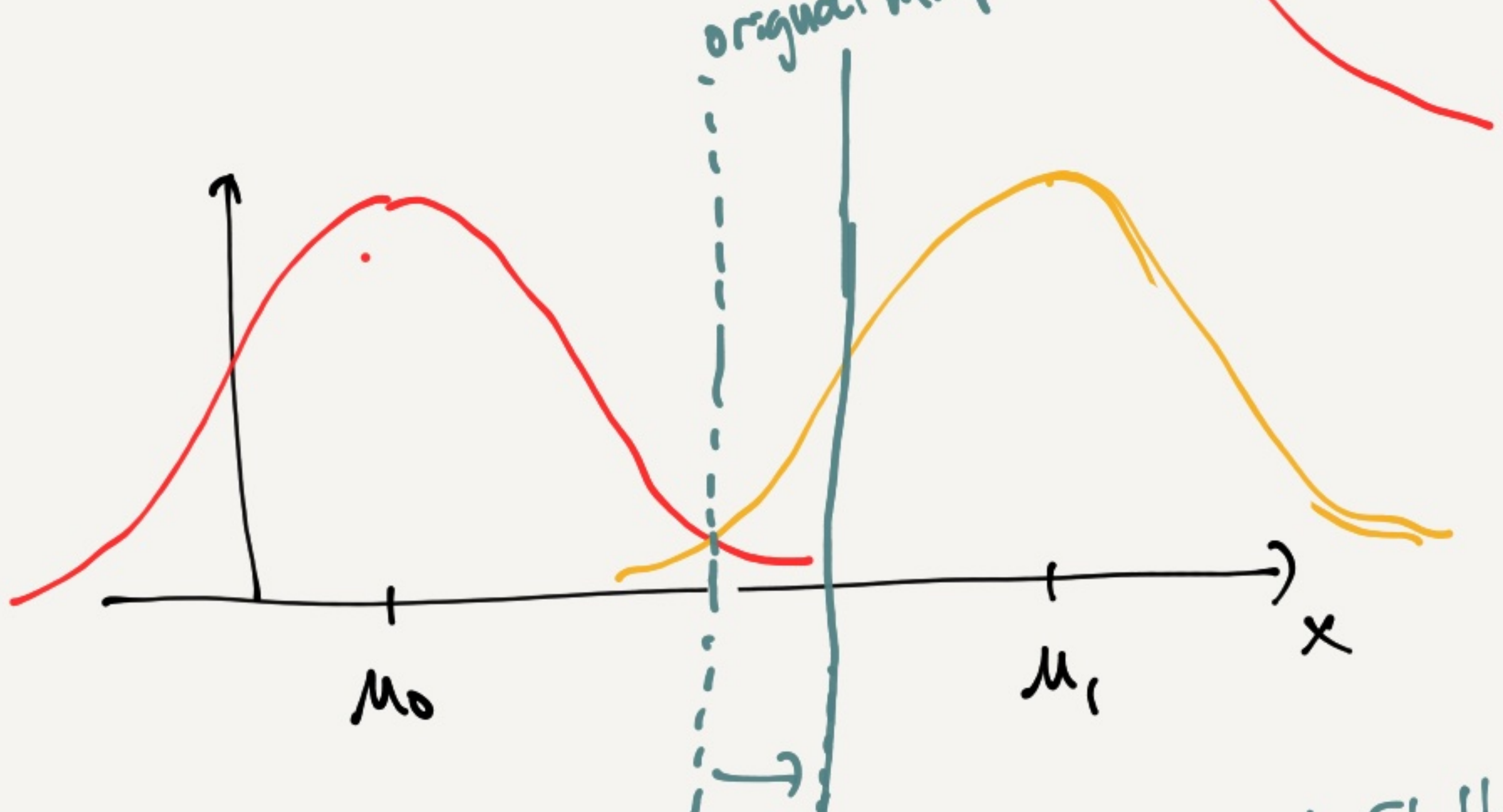
$$x < \underbrace{\frac{\mu_1+\mu_0}{2}}_{\text{same as before}} + \underbrace{\frac{\sigma^2}{\mu_1-\mu_0}} \underbrace{\log\frac{p(y=0)}{p(y=1)}}_{\text{increase threshold if } p(0)>p(1)\text{, i.e. 0 is more frequent.}}$$

$\frac{\mu_1-\mu_0}{\sigma^2}$ = normalized distance btwn means

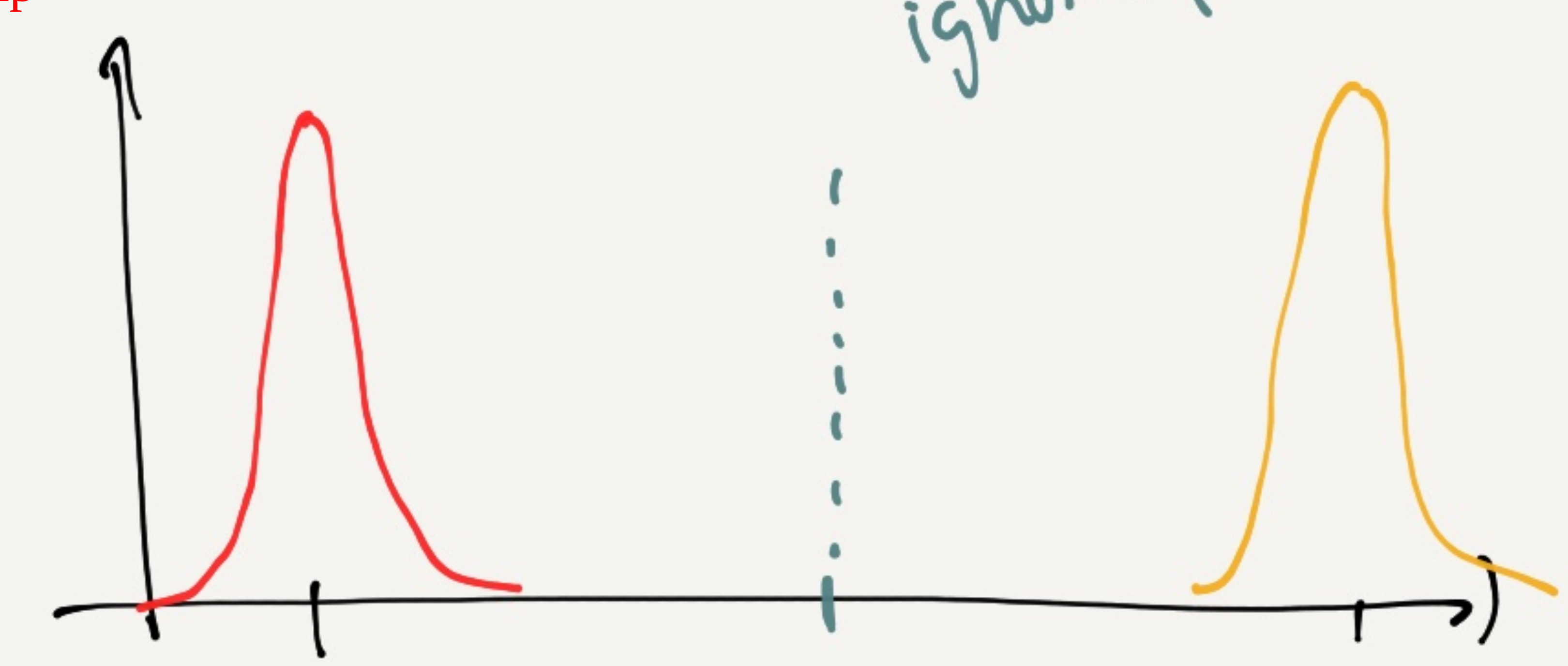
$\Rightarrow \frac{\sigma^2}{\mu_1-\mu_0} = \frac{1}{\text{normalized distance.}}$

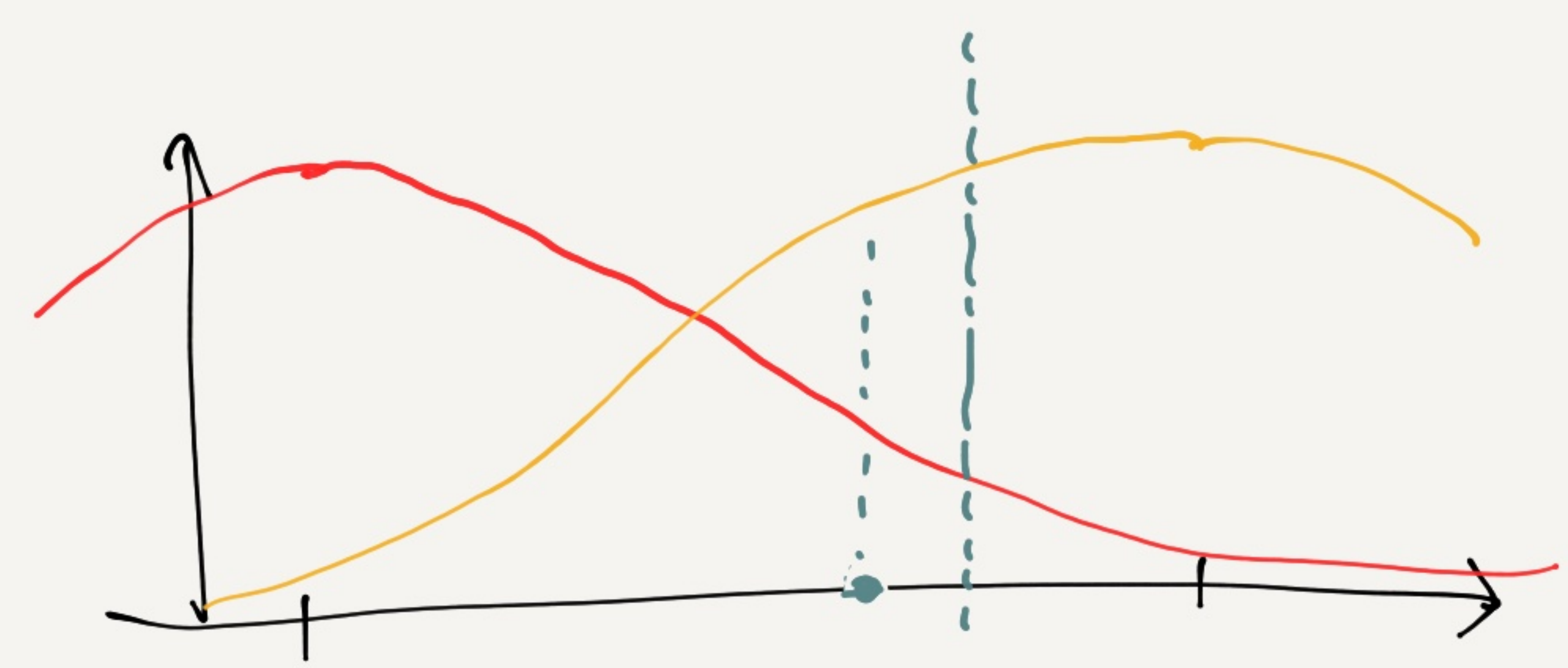
$\Rightarrow$ if means are far apart, then move T a little (ignore prior)

$\Rightarrow$ if means are close, then move T a lot (use priors)



p(0) > p(1) ⇒ shift threshold "capture" more 0's.





(Need to be very certain it is 1 to choose it)

Gaussian Classifier

$Y \in \{1, \dots, C\}$ classes

prior $p(y=j) = \pi_j$ ← can be estimated from data.

$X \in \mathbb{R}^d$

CCDs: $p(x \mid y=j) = N(x \mid \mu_j, \Sigma_j)$

BDR: $g(x) = \arg\max_j \log p(x|j) + \log p(j)$

$$= \arg\max_j -\frac{1}{2}\|x-\mu_j\|^2_{\Sigma_j} - \frac{1}{2}\log|\Sigma_j| \cancel{- \frac{1}{2}\log(2\pi)^d} + \log \pi_j$$

Special cases

i) assume $\Sigma_j = \sigma^2 I$ (shared isotropic covariances)

Define $g_j(x) = w_j^T x + b_j$

where
$$\begin{cases} w_j = \frac{1}{\sigma^2}\mu_j \\ b_j = -\frac{1}{2\sigma^2}\mu_j^T\mu_j + \log \pi_j \end{cases}$$

$\Rightarrow g^*(x) = \arg\max_j g_j(x)$

Geometric Meaning

classes i & j share a boundary if

$g_i(x) = g_j(x)$

$\Rightarrow w_i^T x + b_i = w_j^T x + b_j$

⋮

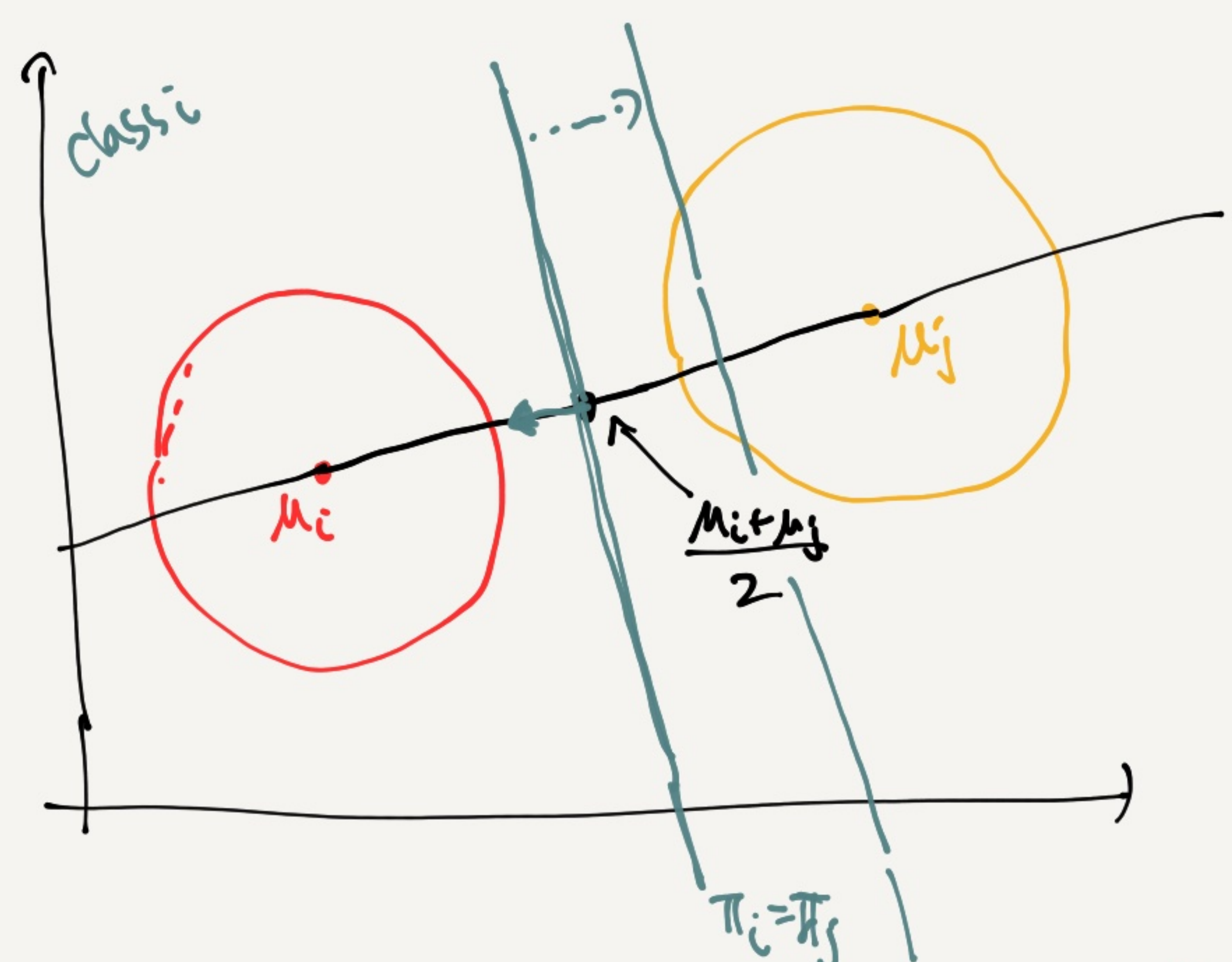
$w^T(x - x_0) = 0$ ← hyperplane passing through x_0 & normal to w.

$$\begin{cases} w = \frac{1}{\sigma^2}(\mu_i - \mu_j) \leftarrow \text{vector btwn } \mu_j \text{ \& } \mu_i \\ x_0 = \underbrace{\frac{\mu_i+\mu_j}{2}}_{\text{midpoint btwn means}} + \underbrace{(\mu_j - \mu_i)}_{\text{vector from } \mu_i \text{ to } \mu_j} \underbrace{\left[\frac{\sigma^2}{\|\mu_i-\mu_j\|^2} \cdot \log\frac{\pi_i}{\pi_j}\right]}_{\text{if } \pi_i > \pi_j \Rightarrow > 0} \end{cases}$$

class j

1 / normalized distance

priors



$\pi_i > \pi_j \Rightarrow$ move the decision boundary away from μ_i to capture more space.