- BDT is a framework for making OPT decisions on problems involving uncertainty

Framework

1) world has states/classes drawn from r.v. Y

eg. $Y \in \{H, B\}$, $Y \in \{ok, flu, cold\}$

prior: $P(Y)$ - prior prob. of state occurring.

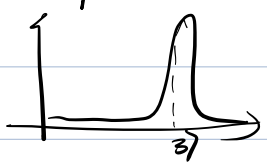
2) observer measures features/observations from r.v. X

class - conditional density (CCD)

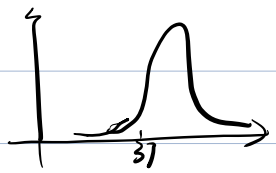
$P(X|Y)$ - observations conditioned on class/state

eg. X - temperature

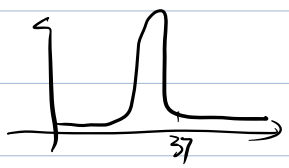
$P(X|ok)$



$P(X|flue)$



$P(X|cold$



3) Decision Function - use observation to make a decision about the state.

$g(x): X \to Y$

(4) loss function - penalty for deciding the wrong Y (wong decision)

$L(g(x), Y) \geq 0$

eg. 0-1 loss function: $L(g(x), y) = \begin{cases} 0, & g(x) = Y \\ 1 & \text{Otherwise.} \end{cases}$

Goal: Find the optimal decision function $g^*(x)$
for given assumptions (loss, prior, CCD)

Bayes Decision Rule (BDR)

Risk - expected value of the loss function

$$\text{Risk} = E_{X,Y}[L(g(x), Y)] = \sum_y \int_x \underbrace{p(x,y)}_{p(y|x)p(x)} L(g(x), y)\, dx$$

$$= \int_x p(x) \left[\sum_y p(y|x) L(g(x), y) \right] dx$$

$\sum_y p(y|x) L(g(x), y)$: Conditional Risk $R(x)$ (function of x)

$$= E_x[R(x)]$$ ← expectation of Conditional Risk

Since $L(g(x), y) \geq 0$, then minimizing risk $\Leftrightarrow$ minimizing conditional risk $R(x)$, $\forall x$

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

$$= \underset{j}{\operatorname{argmin}}\ E_{Y|x}[L(j, y)]$$

↓ function of x

0-1 loss function & classification

$y \in \{1, 2, \dots, c\}$

$g(x) \in \{1, 2, \dots, C\}$

In this case $L(g(x), y)$ ⇒ indicator variable

Conditional Risk $R(x) = E_{Y|X}\, I\, L(g(x), y)]$

$= Pr\,(g(x) \neq y \mid x)$ ← prob. of error, given x

BDR

$$y^* = \arg\min_j R(x) = \arg\min_j Pr\,(j \neq y \mid x)$$

$$= \arg\min_j \; 1 - Pr\,(y = j \mid x)$$

$$y^* = \arg\max_j Pr(y = j \mid x)$$ MAP rule

choose the class w/ largest posterior.

Equivalently,

$$y^* = \arg\max_j \frac{P(x \mid y=j) \cdot P(y=j)}{P(x)} = \arg\max_j P(x \mid y=j)\, P(y=j)$$

$$= \arg\max_j \underbrace{\log P(x \mid y=j)}_{\text{CCD}} + \underbrace{\log P(y=j)}_{\text{prior (typically estimated from data)}}$$

Example

2-class problem (0, 1)

pick 0 if $P(x \mid 0) \cdot P(0) > P(x \mid 1)\, P(1)$

$$\Rightarrow \underbrace{\frac{P(x \mid 0)}{P(x \mid 1)}}_{\text{likelihood ratio test}} > \underbrace{\frac{P(1)}{P(0)}}_{\text{threshold}} = T$$

Summary

for 0-1 loss function:

- BDR is MAP rule (tells the threshold)
- Risk = prob. of error.

- BDR minimizes the risk, i.e. the prob. of error
(nothing is better)

Caveat: assume the model (density is correct)
C.C.D. & prior — best you can do given them

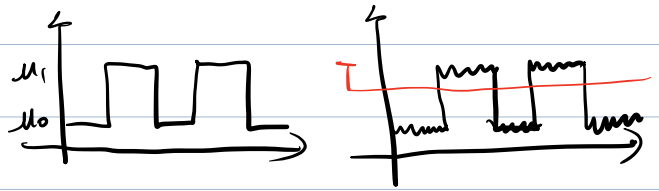
This is called a generative classification model
- model how data is generated in the world
  - CCD & prior learned from data. — assume they are from some existing density, i.e.;

Example: noisy channel.

Y = Bit $\{0, 1\}$ → transmit → channel → receiver → decoder

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

How to choose the threshold T



Given measurement X, recover bit Y

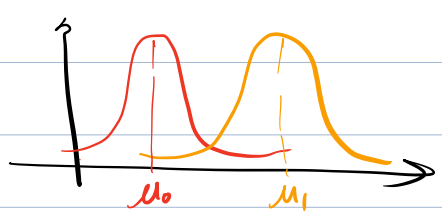
class prob. $p(Y=0) = p(Y=1) = \frac{1}{2}$

CCD: Gaussian additive noise (assume $\mu_1 > \mu_0$)

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

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

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



Assume 0-1 loss, the BDR is

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

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

$$= \arg\max_j -\frac{1}{2\sigma^2}(x^2 - 2x\mu_j + \mu_j^2)$$

$$= \arg\min_j -2x\mu_j + \mu_j^2$$

pick 0 when $-2x\mu_0 + \mu_0^2 \le -2x\mu_1 + \mu_1^2$

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

intuitively halfway between the two

assumptions are explicit:

1) 0-1 loss, BDR
2) uniform class prior
3) Gaussian noise (iid), additive, same for each bit.

What if $p(Y)$ is not uniform?

eg. channel coding: 7 ⇒ 1111110

BDR: pick 0 if $x < \underbrace{\frac{\mu_1+\mu_0}{2}}_{\text{midpoint}} + \frac{\sigma^2}{\mu_1-\mu_0}\log\underbrace{\frac{P(y=0)}{P(y=1)}}$

eg. $P(y=0) > P(y=1) \Rightarrow \log(>1) \Rightarrow >0$

increase if 0 is more frequent.

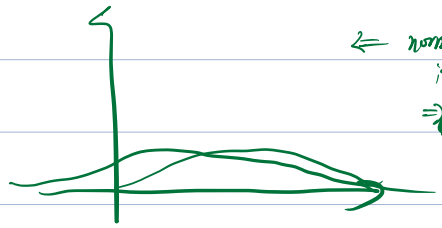
make sense since expecting to see more 0

$\frac{1}{\text{"normalized distance between means"}}$

eg:

⇐ normalized distance is large ⇒ ignore priors

(unlikely we have confusion since they center around the mean)



← normalized distance is small
⇒ vote the prior

Gaussian Classifier

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

$p(y=j) = \pi_j$

$X \in \mathbb{R}^d$, CCDs are m.v. Gaussian.

$p(x|y=j) = \mathcal{N}(x|\mu_j, \Sigma_j)$

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

$= \arg\max_j \underbrace{-\frac{1}{2}\|x-\mu_j\|^2_{\Sigma_j} - \frac{1}{2}\log|\Sigma_j| + \log\pi_j}$

$g_j(x)$ = discriminant function for class j

Special case: $\Sigma_j = \sigma^2 I$ (shared isotropic covariances)

(tutorial 1)

$g_j(x) = w_j^T x + b_j$ ← discriminant is linear.

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

Geometric meaning

classes i & j share a boundary if $g_i(x) = g_j(x)$

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

tutorial

$\Rightarrow \quad w^T(x - x_0) = 0$ ← hyperplane normal to w, goes through x_0

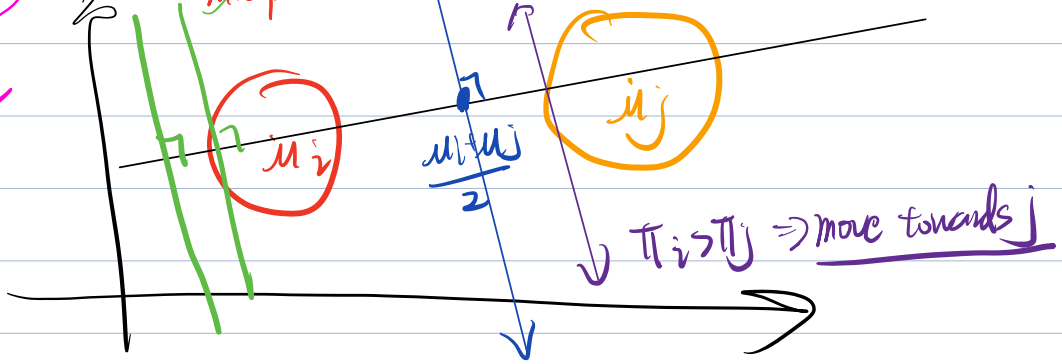
$$w = \frac{1}{\sigma^2}(\mu_i - \mu_j)$$

$$x_0 = \left(\frac{\mu_i + \mu_j}{2}\right) + (\mu_j - \mu_i)\left[\frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{\pi_i}{\pi_j}\right]$$

parameters determine x_0 boundary

midpoint

from $j \to i$



★ analogous to 1-D version (noisy channel)

⇒ hyperplane ~ high-dim.