Lecture 6 - Bayesian Decision Theory (BDT)

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

Framework

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

   e.g. $Y \in \{H, T\}$, $Y \in \{ok, flu, cold\}$

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

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

   class-conditional density (CCD):

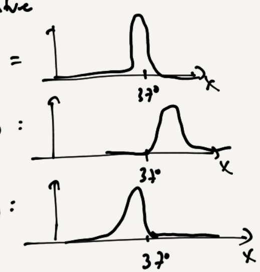
   $p(X|Y)$ - observations conditioned on the class/state.

   e.g. X = temperature

   $p(X|ok) =$

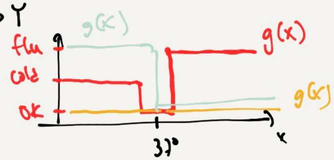
   $p(X|flu) :$

   $p(X|cold) :$



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

   $g(x) : X \to Y$



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

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

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

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

Bayes Decision Rule (BDR)

Risk - expected value of the loss function

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

r.v.

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

(function of x)

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

Since $L(g(x),y) \geq 0$, then minimizing the risk is equivalent to minimizing the conditional risk $R(x)$ for each x.

For an x,

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

$$= \operatorname*{argmin}_j \underbrace{E_{Y|X}[L(j,y)]}_{\text{conditional expectation of loss.}}$$

"Bayes Decision Rule" - will give the minimum risk!

0-1 loss function & Classification

$y \in \{1, 2, \ldots, C\}$

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

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

Conditional Risk: $R(x) = E_{Y|X}[\underbrace{L(g(x),y)}_{\text{indicator variable}}]$

$$= \Pr(g(x) \neq y \mid x) \leftarrow \text{probability of error given the x.}$$

BDR

$$y^* = \operatorname*{argmin}_{j \in Y} R(x) = \operatorname*{argmin}_{j \in Y} \Pr(j \neq y \mid x)$$

$$= \operatorname*{argmin}_j 1 - \Pr(y = j \mid x)$$

$$\boxed{y^* = \operatorname*{argmax}_j \Pr(y = j \mid x)}$$

MAP rule - choose the class w/ largest posterior.

Equivalently

$$y^* = \operatorname*{argmax}_j \frac{p(x|y=j)p(y=j)}{p(x)} = \operatorname*{argmax}_j p(x|y=j)p(y=j)$$

$$\boxed{y^* = \operatorname*{argmax}_j \underbrace{\log p(x|y=j)}_{\text{CCD}} + \underbrace{\log p(y=j)}_{\text{prior}}}$$

Example: 2-class problem (0,1)

$$\text{pick 0 if } \underbrace{p(x|0)}_{\text{CCD}}\underbrace{p(0)}_{\text{prior}} > p(x|1)p(1) \Rightarrow \underbrace{\frac{p(x|0)}{p(x|1)}}_{\text{likelihood ratio test}} > \frac{p(1)}{p(0)} = \underset{\text{threshold}}{T}$$

Summary

for 0-1 loss function:

- BDR is MAP rule (tells us the threshold)
- Risk = probability of error
- BDR minimizes the risk, i.e. the prob. of error (nothing is better)
- Caveat: assuming the model (densities are correct) (CCD, prior)

This is called a <u>generative</u> classification model.

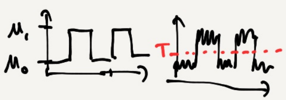
- model how data is generated in the world.
  - CCD & prior learned from data.

Example: Noisy Channel

y = Bit $\{0,1\}$ → [trans.] → [channel] → [reciever] → decoder $y = \begin{cases} 0, & x < T \\ 1, & x \geq T \end{cases}$

how to choose T?

(assume $\mu_0 < \mu_1$)



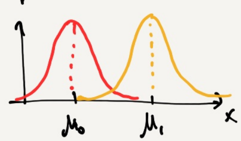
Given measurement X, recover bit Y

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

CCD: Gaussian additive noise: $X = \mu_Y + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

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

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



Assume 0-1 loss, the BDR is:

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

$$= \underset{j}{\operatorname{argmax}} \log N(x|\mu_j, \sigma^2) + \log \tfrac{1}{2}$$

$$= \underset{j}{\operatorname{argmax}} -\frac{1}{2\sigma^2}(x-\mu_j)^2 \underbrace{- \tfrac{1}{2}\log\sigma^2 - \tfrac{1}{2}\log 2\pi + \log\tfrac{1}{2}}_{\text{constant wrt } j}$$

$$= \underset{j}{\operatorname{argmax}} -\frac{1}{2\sigma^2}(x^2 - 2x\mu_j + \mu_j^2)$$

$$= \underset{j}{\operatorname{argmin}} -2x\mu_j + \mu_j^2$$

Hence, pick 0 when $-2x\mu_0 + \mu_0^2 < -2x\mu_1 + \mu_1^2$

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

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

assume $\mu_0 < \mu_1$

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

$$\Rightarrow \boxed{\text{pick 0 when } x < \frac{\mu_1 + \mu_0}{2}}$$

intuitive threshold → halfway between μ_0 & μ_1

Assumption are explicit:

1) 0-1 loss, BDR
2) uniform class prior ($p(Y=0) = \frac{1}{2}$)
3) Gaussian noise (iid), additive, same for each bit.

what if $p(Y)$ is not uniform? PS 6-3

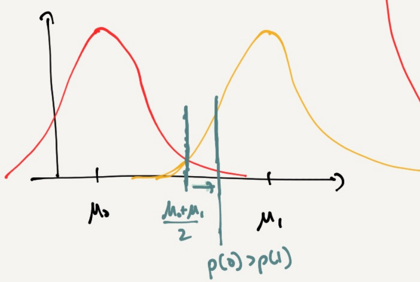
e.g. channel coding: $7 \Rightarrow 1111110$

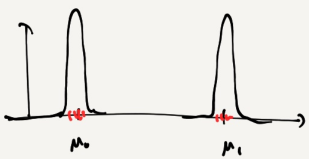
BDR: pick 0 if $x < \underbrace{\frac{\mu_1+\mu_0}{2}}_{\text{same as before}} + \underbrace{\frac{\sigma^2}{\mu_1-\mu_0}}_{} \underbrace{\log \frac{p(y=0)}{p(y=1)}}_{}$

$p(y=0) > p(y=1) \Rightarrow \log(>1) \Rightarrow > 0$

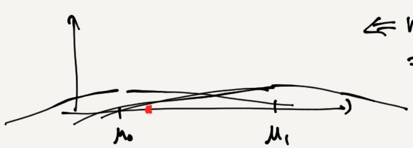
- increase the threshold if 0 is more frequent (predict more 0), vice versa

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





$\Leftarrow$ normalized distance is large $\Rightarrow$ ignore the priors



$\Leftarrow$ normalized distance is small $\Rightarrow$ use the priors.

Gaussian classifier

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

$p(y=j) = \pi_j$

$x \in \mathbb{R}^d$, CCDs are m.v. Gaussians

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

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

$$= \arg\max_j \underbrace{-\frac{1}{2}\|x-\mu_j\|^2_{\Sigma_j} - \frac{1}{2}\log|\Sigma_j| + \log \pi_j}_{g_j(x) \,=\, \text{discriminant function for class } j.}$$

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

⋮ (tutorial)

$g_j(x) = w_j^T x + b_j$ $\Leftarrow$ discriminant is linear function.

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

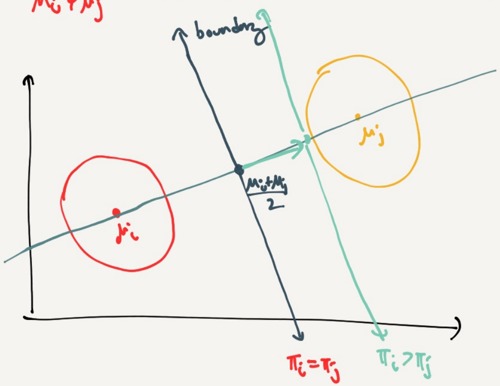
classes i & j share a boundary if $g_i(x) = g_j(x)$

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

⋮ tutorial

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

$$\begin{cases} w = \frac{1}{\sigma^2}(\mu_i - \mu_j) & \leftarrow \text{vector from } \mu_j \text{ to } \mu_i \\ x_0 = \underbrace{\left(\frac{\mu_i + \mu_j}{2}\right)}_{\text{midpoint btwn } \mu_i \,\&\, \mu_j} + \underbrace{\underbrace{(\mu_j - \mu_i)}_{\text{from } \mu_i \text{ to } \mu_j} \left[\frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \underbrace{\log \frac{\pi_i}{\pi_j}}_{\pi_i > \pi_j \Rightarrow \text{positive}}\right]}_{\text{moves the boundary based on priors.}} \end{cases}$$



* analogous to the 1-D version (noisy channel)

$\Rightarrow$ hyperplane ~ high-dim threshold.