Bayesian Decision Theory (BDT)

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

Framework

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

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

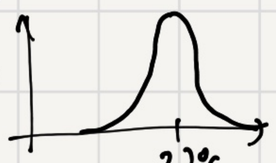
   prior: $p(Y)$ = prior prob. of state occurring.

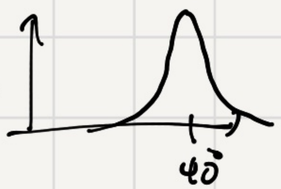
2) Observer measures observations/features from r.v. X

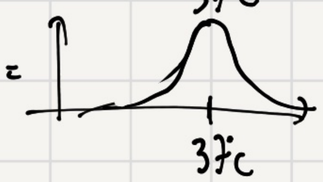
   class conditional density (CCD)

   $p(X|Y)$ = distribution of features for a particular class/state.

e.g. x_1 = temperature

$p(x_1|ok)$ = 

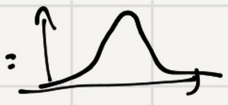
$p(x_1|flu)$ = 

$p(x_1|cold)$ = 

x_2 = quantity of snot

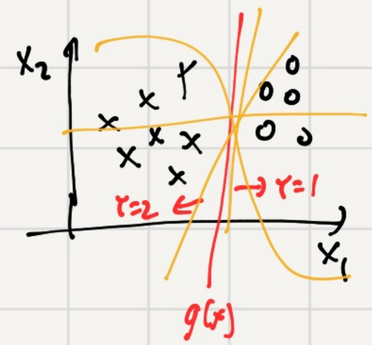
$p(x_2|ok)$ = 

$p(x_2|flu)$ = 

$p(x_2|cold)$ = 

3) Decision Function - use features to infer the state.

$g(x): X \to Y$



4) Loss function - penalty for deciding the wrong Y (making the wrong decision)

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

0-1 loss function.

($g(x)$: prediction, y: true)

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

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

Bayes Decision Rule (BDR)

Risk - expected value of the loss

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

$$= \sum_{y \in Y} \int_X p(x, y) L(g(x), y)\, dx$$

($\sum_{y \in Y} \int_X$: sum over classes, integrate over x; $p(x,y) = p(y|x)p(x)$)

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

($p(x)$ pulled out of the inner sum; dx: integrate over x)

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

$\left[\sum_y p(y|x) L(g(x), y) \right]$: conditional risk $R(x)$ (given observation x)

$E_X[R(x)]$: expectation of cond. risk.

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

Given an x,

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

(value of $g(x)$: j)

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

$E_{Y|x}[L(j,y)]$: conditional exp. of loss

Bayes' Decision Rule

0-1 loss function & classification

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

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

$\Leftarrow$ misclassified sample (x, y)

Conditional Risk:

$$R(x) = E_{Y|x}[L(g(x), y)] = P_r(g(x) \neq y \mid x) = 1 - P(g(x) = y \mid x)$$

$L(g(x), y)$: indicator variable

$P_r(g(x) \neq y \mid x)$: prob. of misclassifying x. (prob. of error)

$P(g(x) = y \mid x)$: prob. correct classification of x. $p(y = g(x) \mid x)$

BDR:

$$g^*(x) = y^* = \underset{j \in Y}{\text{argmin}}\ \underbrace{1 - p(y = j \mid x)}_{R(x)}$$

($j = g(x)$)

$$\boxed{g^*(x) = \underset{j \in Y}{\text{argmax}}\ p(y = j \mid x)}$$

$\Leftarrow$ MAP rule: choose j w/ largest posterior probability.

Equivalently, $g^*(x) = \underset{j}{\text{argmax}} \dfrac{p(x|y=j)p(y=j)}{p(x)}$

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

Example: 2-class problem $y \in \{0, 1\}$

pick 0 if $p(x|0)p(0) > p(x|1)p(1) \Rightarrow \dfrac{p(x|0)}{p(x|1)} > \dfrac{p(1)}{p(0)} = T$

$\dfrac{p(x|0)}{p(x|1)}$: likelihood ratio test

$\dfrac{p(1)}{p(0)}$: threshold

Summary: for 0-1 loss:

- BDR is MAP rule
- Risk = prob. of error
- BDR minimizes risk (best you can do)
- caveat: assume densities are correct

generative classification model

Example: Noisy Channel

Bit $Y \in \{0,1\}$ → Trans → Channel → Rec. → Decoder

Assume $\mu_0 < \mu_1$

Noise

$X \in \mathbb{R}$

$$Y = \begin{cases} 0, & X < T \\ 1, & X > T \end{cases}$$

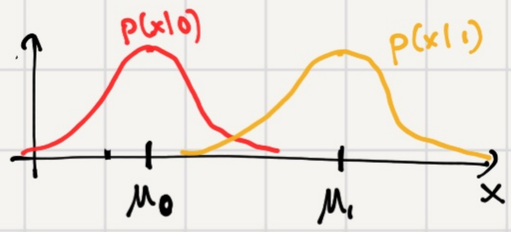
what is best T?

Goal: given X, recover bit Y

Assume: $p(Y=0) = p(Y=1) = \frac{1}{2}$

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

$$\Rightarrow \begin{cases} p(x|0) = N(x|\mu_0, \sigma^2) \\ p(x|1) = N(x|\mu_1, \sigma^2) \end{cases}$$



BDR w/ 0-1 loss

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

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

$\frac{-1}{2\sigma^2}$: scalar; the rest: constant

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

x^2: constant

$$y^* = \arg\min_j \mu_j^2 - 2x\mu_j$$

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

$$2x(\mu_1 - \mu_0) < \mu_1^2 - \mu_0^2 \quad \Leftarrow a^2 - b^2 = (a-b)(a+b)$$

$$\boxed{x < \frac{\mu_0 + \mu_1}{2}}$$

intuitive threshold → halfway btwn means

Assumptions
1) 0-1 loss (MPE)
2) uniform prior
3) Gaussian iid noise, additive

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

e.g. channel coding: 7 → 1111110

↓ PS6-3

BDR: pick 0 if:

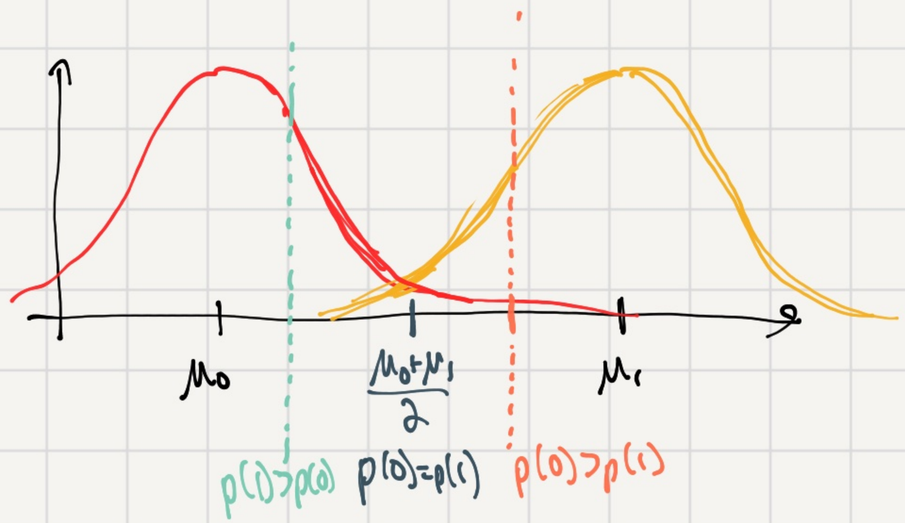
$$x < \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{p(y=0)}{p(y=1)}$$

same as before

"normalized distance btwn means"

- if means are far apart, then ignore priors
- if means are close, use prior.

- if $p(y=0) = p(y=1) \Rightarrow$ same as before
- if $p(y=0) > p(y=1)$ $\Rightarrow \log \frac{p(0)}{p(1)} > 0 \Rightarrow$ increase threshold (capture more space for bit 0)
- if $p(1) > p(0)$ $\Rightarrow \log \frac{p(0)}{p(1)} < 0 \Rightarrow$ decrease threshold



Gaussian Classifier

$Y \in \{1, \dots, C\}$, C classes, $p(Y=j) = \pi_j$ ← different mean, cov, & prior for each class

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

CCD are Gaussian: $p(x \mid y=j) = \mathcal{N}(x \mid \mu_j, \Sigma_j)$

BDR & 0-1 loss: $g^*(x) = \underset{j}{\text{argmax}} \log p(x \mid y=j) + \log p(j)$

$$g^*(x) = \underset{j}{\text{argmax}} \underbrace{-\frac{1}{2}\|x-\mu_j\|^2_{\Sigma_j^{-1}} - \frac{1}{2}\log|\Sigma_j| + \log \pi_j}_{g_j(x)}$$

Special case: Assume: $\Sigma_j = \sigma^2 I$ (shared isotropic cov.)

$$g_j(x) = -\frac{1}{2\sigma^2}\|x-\mu_j\|^2 - \frac{1}{2}\log|\sigma^2 I| + \log \pi_j$$

$$= -\frac{1}{2\sigma^2}(x^T x - 2x^T\mu_j + \mu_j^T\mu_j) + \log \pi_j + \text{const.}$$

$$= \underbrace{-\frac{1}{2\sigma^2}(-2x^T\mu_j + \mu_j^T\mu_j) + \log \pi_j}_{\uparrow} + \text{const.}$$

$$\Rightarrow g_j(x) = \underbrace{w_j^T x + b_j}_{\text{linear discriminant function.}} \quad \text{where} \begin{cases} w_j = \frac{1}{\sigma^2}\mu_j \\ b_j = -\frac{1}{2\sigma^2}\mu_j^T\mu_j + \log \pi_j \end{cases}$$

Note: similar to 1-D case ⇒ hyperplane as high-dim version of threshold.

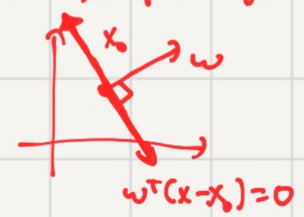
Geometric meaning

classes (i, j) share a boundary when $g_i(x) = g_j(x)$

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

tutorial

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



$$\begin{cases} w = \frac{1}{\sigma^2}(\mu_i - \mu_j) \\ x_0 = \frac{\mu_i + \mu_j}{2} + (\mu_j - \mu_i)\left[\frac{\sigma^2}{\|\mu_i - \mu_j\|^2}\log\frac{\pi_i}{\pi_j}\right] \end{cases}$$

- $\frac{\mu_i + \mu_j}{2}$: midpoint btwn means
- $(\mu_j - \mu_i)$: vector from $\mu_i \to \mu_j$
- $\frac{\sigma^2}{\|\mu_i - \mu_j\|^2}$: $= 1$ normalized distance
- $\log\frac{\pi_i}{\pi_j}$: $\pi_i > \pi_j \Rightarrow \log(\cdot) > 0$; $\pi_i < \pi_j \Rightarrow \log(\cdot) < 0$

move $x_0 \Rightarrow$ boundary