Lecture 6 Bayesian Decision Theory
BDT

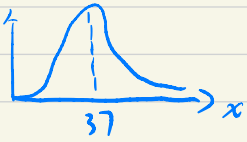
- BDT is a framework for making optimal decisions on problems involving uncertainty.

Framework

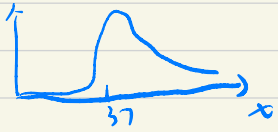
1) world has states / classes drawn from r.v. Y.
   e.g. $Y \in \{H, T\}$, $Y \in \{ok, flu, cold\}$
   prior: $P(Y)$ - prior probility of state occuring.

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 $g(x)$: $X \rightarrow Y$

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, $c(\cdot)$...)

Bayes Decision Rule (BDR)

Risk - expected value of the loss function

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

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

conditional risk $R(x)$ — function of x

$$= E_x[R(x)]$$

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

0-1 loss function and classification.

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

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

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

Conditional Risk: $R(x) = E_{y|x}[L(g(x), y)]$

indicator variable

$= P_r(g(x) \neq y \mid x)$

Probability of error given the x.

BDR.

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

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

$$y^* = \underset{j}{\operatorname{argmax}} \; pr(y = j \mid x)$$

MAP rule – Chouse the class with largest posterior.

Equvalent

$$y^* = \underset{j}{\operatorname{argmax}} \frac{p(x \mid y=j)\, p(y=j)}{p(x)} = \underset{j}{\operatorname{argmax}} \; p(x \mid y=j)\, p(y=j)$$

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

C.C.D. Prior

Example: 2-class problem (0, 1)

threshold

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

Summary

for 0-1 loss function.

- BDR is MAP rule (tell us the threshold)
- Risk = probability of error
- BDR minimize the risk. ie. the probability of an error. (nothing is better)

- Caveat: assuming the model (densities are correct)
  CCD prior

This is called a generative classification model.
model how data is generated in the world.
- CCD and prior learned from data

Example: Noisy channel.

Y = Bit {0,1} → [Trains] → [channels] → [receive] →

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

How to choose T?

assume $\mu_0 < \mu_1$

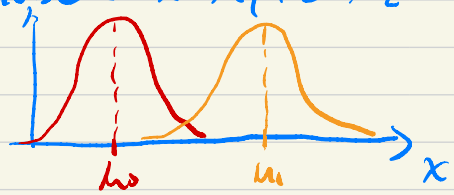
Given measurement X, recover bit Y.

Class probability. $p(Y=0) = p(Y=1) = \frac{1}{2}$

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

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

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



Assume 0-1 loss, the BDR is:

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

$$= \arg\max_j \log N(x|\mu_j, \sigma^2) + \log \frac{1}{2}$$

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

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

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

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

$$\Rightarrow \quad 2x(\mu_1 - \mu_0) < (\mu_1 - \mu_0)(\mu_1 + \mu_0)$$

$$\Rightarrow \quad x < \frac{1}{2}(\mu_1 + \mu_0)$$

intuitive threshold → halfway between μ_0 and μ_1

Assumption are explicit:

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

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

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

Pick 0 if

$$\log P(x|\mu_0, \sigma^2) + \log P(0) > \log P(x|\mu_1, \sigma^2) + \log P(1)$$

$$-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu_0)^2 + \log P(0) > -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu_1)^2 + \log P(1)$$

$$\Rightarrow \frac{1}{2\sigma^2}\left[x^2 - 2x\mu_1 + \mu_1^2 - x^2 + 2x\mu_0 - \mu_0^2\right] + \log\frac{P_0}{P(1)} > 0$$

$$\Rightarrow \frac{1}{2\sigma^2}\left[(\mu_1-\mu_0)(\mu_1+\mu_0) - 2x(\mu_1-\mu_0)\right] + \log\frac{P(0)}{P(1)} > 0$$

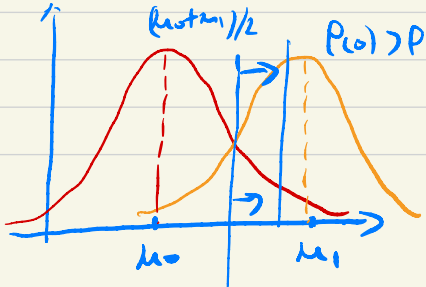
$$\Rightarrow \frac{2\sigma^2}{\mu_1-\mu_0}\log\frac{P(0)}{P(1)} + (\mu_1+\mu_0) - 2x > 0$$

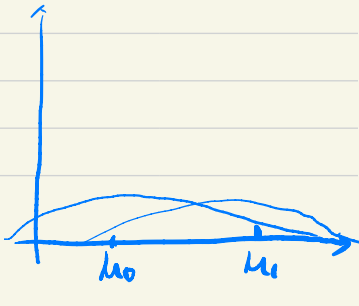
$$\Rightarrow x < \underbrace{\frac{1}{2}(\mu_1+\mu_0)}_{\text{Same as before}} + \underbrace{\frac{\sigma^2}{(\mu_1-\mu_0)}\log\frac{P(0)}{P(1)}}$$

"normalized distance between means"

$P(y=0) > P(y=1) \Rightarrow \log\frac{P(0)}{P(1)} > 1 \Rightarrow > 0$

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





normalized distance is large ⇒ ignore the priors

normalized distance is small ⇒ use priors.

Gaussian classifier

$Y \in \{1, \cdots, C\}$ C classes $p(y=j) = \pi_j$

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

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

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, 等方共分散)

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

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

Geometric meaning

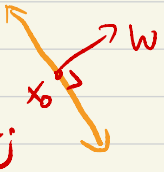
classes i and j share a boundary if $g_i(x) = g_j(x)$

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

tutorial

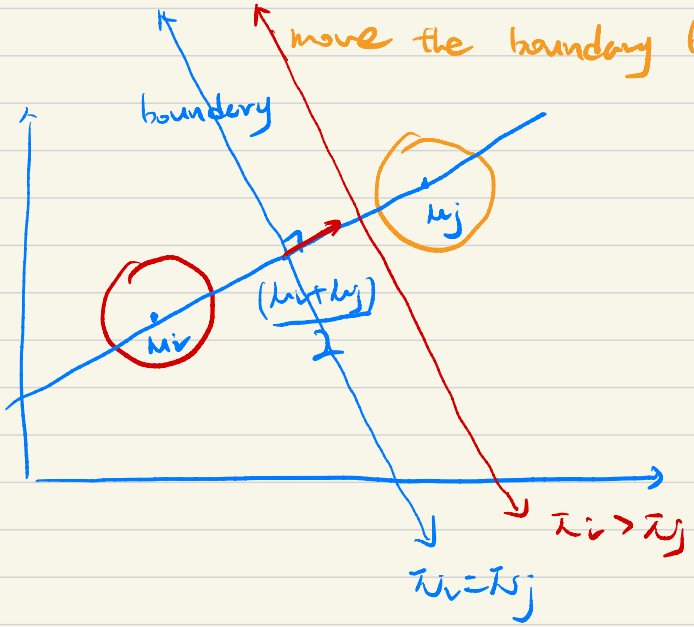
hyperplane normal to w goes through x_0

$\Rightarrow w^T(x - x_0) = 0$



$$\begin{cases} w = \frac{1}{\sigma^2}(\mu_i - \mu_j) \leftarrow \text{vector from } \mu_i \text{ to } \mu_j \\ x_0 = \underbrace{\frac{\mu_i + \mu_j}{2}}_{\text{midpoint between } \mu_i \text{ and } \mu_j} + \underbrace{(\mu_j - \mu_i)}_{\text{from } \mu_i \text{ to } \mu_j} \underbrace{\left[\frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{\pi_i}{\pi_j}\right]}_{\pi_i > \pi_j \text{ positive}} \end{cases}$$

move the boundary based on priors.



analogous to the 1-D version (noisy channel)

$\Rightarrow$ hyperplane $\sim$ high-dim threshold.