Lecture 11 FML March 30, 2021

Multiclass classification

M classes H_j, $j = 0, \dots, M-1$

$y \in \mathcal{Y}$

↳ observations

Bayesian Framework

- If data is coming from class j then y is a r.v. with distribution $f(y|H_j)$
- we know $f(y|H_j)$, $j = 0, \dots, M-1$,

Prior probs of class H_j: $= \Pi_j$

C_{ij} = cost of deciding class i for observation y but it comes from class j.

≥ 0 and

$C_{jj} < C_{ij}$, $i \neq j$

Decision rule

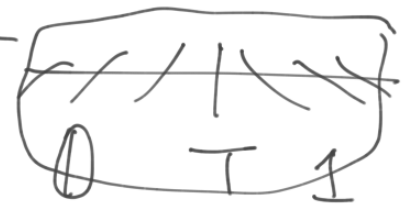
$\delta(y) \in 0, \dots, M-1$.

If $y \in Y_j$ then

$\delta(y) = H_j$

Y_j are disjoint

$Y_j = \{y : \delta(y) = j\}$



Bayes' Risk

$$R(\delta) = \sum_{j=0}^{M-1} \sum_{i=0}^{M-1} c_{ij} \cdot P[y \in Y_i \mid H_j] \pi_j$$

$y \in \mathbb{R}^d$ then $Y_j \subseteq \mathbb{R}^d$.

$\bigcup_j Y_j \subseteq \mathbb{R}^d$

— we want to choose decision rule δ s.t. $R(\delta)$ is minimized

$$R(\delta) = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} c_{ij} P(y_i \mid H_j) \pi_j$$

$$= \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} c_{ij} \int_{y_i} f(y \mid H_j) \pi_j \, dy$$

$P(H_j)$

we know $f(y \mid H_j)$, $P(H_j) = \pi_j$

$f(y) = \sum_j f(y \mid H_j) \pi_j$

$P(H_j \mid y) f(y) = P(y, H_j)$

$f(y)$ ← prob y

$$R(\delta) = \sum_i \sum_j c_{ij} \int_{\mathcal{Y}_i} f(y|H_j)\pi_j \, dy$$

$$= \sum_i \sum_j c_{ij} \int_{\mathcal{Y}_i} P(H_j|y) f(y) \, dy$$

$$= \sum_i \int_{\mathcal{Y}_i} \left[\sum_j c_{ij} P(H_j|y) \right] f(y) \, dy$$

$$= \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} C_i(y) f(y) \, dy$$

$$C_i(y) = \sum_{j=0} c_{ij} P(H_j|Y=y)$$

Average cost of δ given observation y.

- finding decision rule δ minimizing $R(\delta)$ is equivalent to finding regions $\mathcal{Y}_i$ which will minimize $R(\delta)$.

$$\mathcal{Y}_i = \{y : C_i(y) = \min_{j=0,\ldots,M-1} C_j(y)\}$$

Optimal decision rule (Bayes decision rule)

$$\delta_B(y) = \underset{0 \le i \le M-1}{\operatorname{argmin}} C_i(y)$$

Special cases:

$c_{ij} = 1\{i \neq j\}$

$c_{jj} = 0$.

This gives Bayes rule that minimizes prob of error

$$C_i(y) = \sum_{j=0}^{M-1} c_{ij} P(H_j | Y=y)$$

$$= \sum_{j \neq i} P(H_j | Y=y)$$

$$= 1 - P[H_i | Y=y]$$

$\pi_j = P(H_j)$ ← a priori prob of H_j

$P(y|H_j)$ $j \dots$

$P(H_j|y)$

$$= \frac{P(H_j, y)}{P(y)} = \frac{P(y|H_j)\,\pi_j}{P(y)}$$

Bayes rule

$$\delta_B(y) = \arg\min_i C_i(y)$$

MAP

$$= \arg\min_i (1 - P[H_i|Y=y])$$

$$\text{a posteriori} = \arg\max_i P(H_i|Y=y)$$

$$f_{MAP}(y) = \arg\max_i P[H_i | Y = y]$$

$$= \arg\max_i \frac{P[H_i, Y = y]}{P(Y = y)}$$

$$= \arg\max_i P[H_i, Y = y]$$

$$= \arg\max_i P[Y = y | H_i] \pi_i$$

→ If π_i are all same

$\therefore \pi_i = \frac{1}{M}$.

$$= \arg\max_i P[Y = y | H_i]$$

ML classifier

Ex M classes

Y vector of dim m

$\sim f[y|H_i] = N(m_i, \Lambda_i)$

MAP classifier

$$= \arg\max_i \frac{1}{(2\pi)^{m/2} \det(\Lambda_i)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y-m_i)^T \Lambda_i (y-m_i)\right) \cdot \pi_i$$

$\Lambda_i = \sigma^2 I, \forall i$

$\pi_i = 1/M$

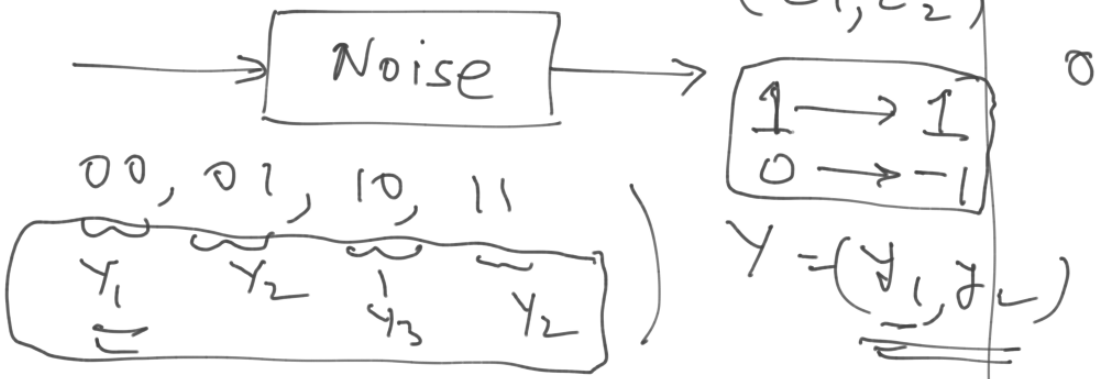
$$\arg\max_i \left(\exp\left(-\frac{1}{2\sigma^2}(y-m_i)^T(y-m_i)\right)\right)$$

$$= \arg\min_i (y-m_i)^T(y-m_i)$$

$$= \arg\min_i \|y-m_i\|^2$$

Min distance classifier

ex communication



$z = (z_1, z_2)$

$Y = (Y_1, Y_2)$

$z_i = y_i + N_i, \; i = 1, 2.$

$N_i \to N(0, \sigma^2)$

Receiving z, we want to know what was transmitted

$P[z|M_i]$

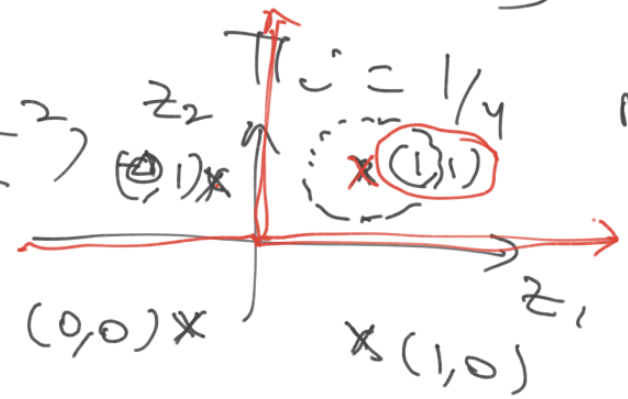
$00 \to (N_1, N_2) \to (0,0)$

$\sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$01 \to (N_1, 1 + N_2)$

cov. $\sigma^2 I$

mean $(0, 1)$



$i = 1/4$

ML classi

In ML we may not know $P(y|H_i)$

We know prior dist

$\pi_i = P(H_i)$ ← Prior.

iid observations → m sample → $(x_1, y_1), \ldots, (x_m, y_m)$

from a class $0, \ldots, M-1$

$\mathcal{H} = \{0, 1, \ldots, M-1\}$

label $\in \mathcal{H}$.

We have posterior dist (based on m iid sample)

$Q(h) = P[H_i | y_1, \ldots, y_m]$

(loss) cost fn $\ell(h)$ of $h(x), y$

$R(Q) = E_Q[\ell(h(x), y)]$

Theorem with prob $\geq 1-\delta$

$R(Q) \leq \hat{R}(Q) + [D(Q\|P) + \ln\frac{m}{\delta})/2(m-1)]^{1/2}$ — (X)

Find classifier that Q min RHS in above theorem

$$D(Q\|P) = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P(x)$$

if support of Q is in P

o.w. $= \infty$.

ML setup

X input space

Y output space

$\to k$ no of classes $x_i \in X$, $y_i \in Y$

Sample $S = (x_1, y_1), \ldots, (x_m, y_m)$

iid $\sim D$, D not known.

Assumption $\exists f$:

$y_i = f(x_i)$

$y = \{1, 2, \ldots, k\}$

H Hypothesis set.

- We estimate f from some function $h \in H$

$h : X \to Y$

$$R(h) = \underset{D}{E}[\, 1\{h(x) \neq f(x)\}]$$

= expected error for $h \in H$.

generalization error

- Based on m samples we find $h \in \mathcal{H}$ which minimizes $R(h)$.
- We don't know $R(h)$ because we do not know the dist D of (x, y).
- From the sample we can find out empirical error:

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq y_i\}$$

For the binary classification using SVM:

$$y = \{-1, 1\}$$

$$h(x) = \text{sgn}(w \cdot x + b)$$

(w, b) define a hyperplane

optimal sol. using dual problem

$$= \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b\right)$$

$$h(x) = \arg\max_{y \in \{-1,1\}} (y \cdot (w \cdot x + b))$$

$y = 1$: +ve

$y = -1$: -ve

$h(x) = \text{sgn}(w \cdot x + b)$

$= \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x + b)\right)$

$= \underset{y \in \{-1,1\}}{\text{argmax}} \left(y (w \cdot x + b)\right)$

$= \underset{y \in \{-1,1\}}{\text{argmax}} \left(y \cdot \left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x + b)\right)\right)$

In kernel SVM

$K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$

$\Phi : \mathcal{X} \longrightarrow \mathbb{H}$ Hilbert space

$K(x, x') = \langle \Phi(x), \Phi(x') \rangle$

replace

$x_i \cdot x \longrightarrow K(x_i, x)$

$= \langle \Phi(x_i), \Phi(x) \rangle$

we absorb b in kernel.

$h(x) = \underset{y \in \{-1,1\}}{\text{argmax}} \left(y \sum_{i=1}^{m} \alpha_i y_i \langle \Phi(x_i), \Phi(x) \rangle\right)$

$= \underset{y \in \{-1,1\}}{\text{argmax}} \left(\left\langle \sum_{i=1}^{m} y \alpha_i y_i \Phi(x_i), \Phi(x) \right\rangle\right)$

$\hookrightarrow w_y$

α_i picked = optimal from dual optimization problem

$h(x) = \underset{y \in \{-1,1\}}{\text{argmax}} \{\langle w_y, \Phi(x) \rangle\}$

For k class classifier

$h(x) = \underset{y \in \{1, \ldots, k\}}{\text{arg max}} \; h(x, y)$ — score function.