Bayesian Decision Theory

- Assume decision problem posed in probabilistic terms.
- Assume all relevant prob. values known
- Decision / Hypothesis testing / classification } decide among set of mutually exclusive alternatives
- Assume set of hypotheses, categories, "states of nature" which partition sample space.
- Each performance of experiment $\Rightarrow$ one and only one hypothesis or category correct.
- Y a r.v. corresponding to correct hypothesis

  $Y \in \mathcal{Y}$ e.g. $\mathcal{Y} = \{1, .., m\}$.
- $P_Y(y)$ = prob. of hypothesis y : <u>a prior prob.</u> of y
- observe feature vector $\underline{X} \in \mathcal{X}$. Based on observed value $\underline{x}$ of $\underline{X}$, make decision on hypothesis.

- Conditional PMF $P_{\underline{X}|Y}(\underline{x}|y)$ or density $f_{\underline{X}|Y}(\underline{x}|y)$
  likelihood

- Experiment performed $\rightarrow \omega \in \Omega \rightarrow Y = y$ correct hypothesis or category or state of nature
  $\underline{X} = \underline{x}$ feature
  $\downarrow$
  $\alpha(\underline{x})$ decision based on $\underline{x}$ (not on y)

- $\alpha(\cdot)$ decision rule

  Require $\alpha(\underline{x}) \in \mathcal{Y} \quad \forall \underline{x}$

- Pay a cost if you make decision $\alpha(\underline{x})$ when true state is y

- Loss function $L(\alpha(\underline{x}), y)$

  e.g. all wrong answers penalized same:

  0-1 loss fn

  $L(\alpha(\underline{x}), y) = 0$ if $\alpha(\underline{x}) = y$ (correct decision)

  $L(\alpha(\underline{x}), y) = 1$ if $\alpha(\underline{x}) \neq y$

  But could have if $\mathcal{Y} = \{0, 1\}$

  $L(\alpha(\underline{x}) = 1, y = 0) = 1$ false positive

  $L(\alpha(\underline{x}) = 0, y = 1) = 100$ false negative

- **Risk function** combines loss function, decision rule, probabilities

$$(1) \qquad R(\alpha) = \sum_{y} \int_{\underline{x}} L(\alpha(\underline{x}), y)\, P_{Y|\underline{X}}(y|\underline{x})\, f_{\underline{X}}(\underline{x})\, d\underline{x}$$

- Choose decision rule to minimize risk

$$\hat{\alpha} = \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}}\ R(\alpha) \qquad \underline{\text{Bayes decision}}$$

$$R(\hat{\alpha}) = \underline{\text{Bayes risk}}$$

- (1) can be minimized by

$$(2) \quad \text{choosing } \hat{\alpha}(\underline{x}) = \underset{\alpha(\underline{x}) \in \mathcal{Y}}{\operatorname{argmin}} \sum_{y} L(\alpha(\underline{x}), y)\, P_{Y|\underline{X}}(y|\underline{x})$$

- Suppose we have 0-1 loss function:

$$\text{then} \quad \sum_{y} L(\alpha(\underline{x}), y)\, P_{Y|\underline{X}}(y|\underline{x})$$

$$= \sum_{y \neq \alpha(\underline{x})} P_{Y|\underline{X}}(y|\underline{x})$$

$$= 1 - P_{Y|\underline{X}}(y = \alpha(\underline{x})|\underline{x})$$

$$(2) \iff \hat{\alpha}(\underline{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ \underbrace{P_{Y|\underline{X}}(y|\underline{x})}_{\text{a posteriori prob. of } y \text{ given } \underline{x}}$$

Max. a posteriori prob. (MAP) rule

Note $\quad P_{Y|\underline{X}}(y|\underline{x}) = \dfrac{P_Y(y)\, f_{\underline{X}|Y}(\underline{x}|y)}{f_{\underline{X}}(\underline{x})}$ by Bayes

For given $\underline{x}$, $\quad \arg\max \; P_{Y|\underline{X}}(y|\underline{x})$

$$\Leftrightarrow \arg\max \; P_Y(y)\, f_{\underline{X}|Y}(\underline{x}|y)$$

If $P_Y(y)$ same for all $y \in \mathcal{Y}$: equiprob. hypotheses

$$\text{MAP} \Leftrightarrow \arg\max_{y \in \mathcal{Y}} f_{\underline{X}|Y}(\underline{x}|y)$$

<u>Maximum Likelihood</u> (ML) Rule

<u>Two category classification</u>

Suppose $\mathcal{Y} = \{1, 2\}$.

$L(\alpha(\underline{x}), y)$ can be expressed as 2×2 matrix.

$$L_{a,i} = L(\alpha(\underline{x}) = a, y = i) \qquad \begin{array}{l} a = 1, 2 \\ i = 1, 2 \end{array}$$

Expected loss for decision $\alpha(\underline{x}) = 1$ is

$$L_{1,1}\, P_{Y|\underline{X}}(1|\underline{x}) + L_{1,2}\, P_{Y|\underline{X}}(2|\underline{x})$$

For $\alpha(\underline{x}) = 2$:

$$L_{2,1}\, P_{Y|\underline{X}}(1|\underline{x}) + L_{2,2}\, P_{Y|\underline{X}}(2|\underline{x})$$

$$\hat{\alpha}(\underline{x}) = \underset{a=1,2}{\operatorname{argmin}} \left[L_{a,1} P_{Y|\underline{X}}(1|\underline{x}) + L_{a,2} P_{Y|\underline{X}}(2|\underline{x}) \right]$$

i.e.

$$L_{1,1} P_{Y|\underline{X}}(1|\underline{x}) + L_{1,2} P_{Y|\underline{X}}(2|\underline{x}) \underset{\hat{\alpha}(\underline{x})=1}{\overset{\hat{\alpha}(\underline{x})=2}{\gtrless}} L_{2,1} P_{Y|\underline{X}}(1|\underline{x}) + L_{2,2} P_{Y|\underline{X}}(2|\underline{x})$$

or $$(L_{1,2} - L_{2,2}) P_{Y|\underline{X}}(2|\underline{x}) \underset{\hat{\alpha}(\underline{x})=1}{\overset{\hat{\alpha}(\underline{x})=2}{\gtrless}} (L_{2,1} - L_{1,1}) P_{Y|\underline{X}}(1|\underline{x})$$

Assuming $L_{1,2} - L_{2,2} > 0$ (wrong decision has larger loss)

$$\frac{P_{Y|\underline{X}}(2|\underline{x})}{P_{Y|\underline{X}}(1|\underline{x})} \underset{\hat{\alpha}(\underline{x})=1}{\overset{\hat{\alpha}(\underline{x})=2}{\gtrless}} \frac{L_{2,1} - L_{1,1}}{L_{1,2} - L_{2,2}}$$

By Bayes,

$$\Lambda(\underline{x}) = \underbrace{\frac{f_{\underline{X}|Y}(\underline{x}|2)}{f_{\underline{X}|Y}(\underline{x}|1)}}_{\substack{\text{likelihood ratio} \\ \text{LR}}} \underset{\hat{\alpha}(\underline{x})=1}{\overset{\hat{\alpha}(\underline{x})=2}{\gtrless}} \underbrace{\frac{P_Y(1)(L_{2,1} - L_{1,1})}{P_Y(2)(L_{1,2} - L_{2,2})}}_{\substack{T_L \\ \text{threshold} \\ \text{indep of } \underline{x}}}$$

e.g. 0-1 loss function:

$$T_L = \frac{P_Y(1)}{P_Y(2)}. \qquad \underline{\text{MAP rule}}$$

if $P_Y(1) = P_Y(2) = 1/2$: (can be used when $P_Y(1)$, $P_Y(2)$ unknown)

$$T_L = 1. \qquad \underline{\text{ML Rule}}$$

Note that for MAP rule:

$$P_Y(1) \nearrow \Rightarrow T_L \nearrow \Leftrightarrow |\{ \underline{x} : \alpha(\underline{x}) = 1 \}| \nearrow$$

i.e. more certain $Y = 1$ initially

$\Rightarrow$ need stronger evidence to change mind.

<u>Multicategory case</u> $\mathcal{Y} = \{1, \dots, m\}$

Discriminant functions $g_i(\underline{x})$, $i = 1, \dots, m$.

$$\hat{\alpha}(\underline{x}) = i \quad \text{if} \quad g_i(\underline{x}) \geq g_j(\underline{x}) \quad \forall j \neq i$$

For general case with risks,

$$g_i(\underline{x}) = -\sum_{y} L(i, y)\, P_{Y|\underline{X}}(y|\underline{x})$$

For 0-1 loss function,

$$g_i(\underline{x}) = P_{Y|\underline{X}}(i|\underline{x})$$

Note if we replace $g_i(\underline{x})$ by $f(g_i(\underline{x}))$ where f is monotonically increasing, resulting decision is unchanged.

Thus, for 0-1 loss function, can use

$$g_i(\underline{x}) = P_{Y|\underline{X}}(i|\underline{x})$$

$$g_i(\underline{x}) = f_{\underline{X}|Y}(\underline{x}|i)\, P_Y(i)$$

$$g_i(\underline{x}) = \ln f_{\underline{X}|Y}(\underline{x}|i) + \ln P_Y(i)$$

- Divides $\mathcal{X}$ into <u>decision regions</u> $\mathcal{R}_1, \dots, \mathcal{R}_m$

  If $g_i(\underline{x}) \geq g_j(\underline{x})$ $\forall j \neq i$, then $\underline{x} \in \mathcal{R}_i$