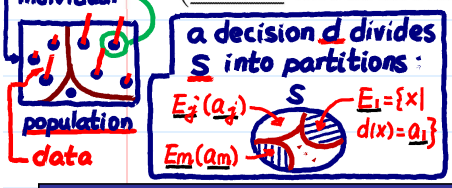
• Application of Decision Theory: Classification

cf. Estimation, Testing — Data

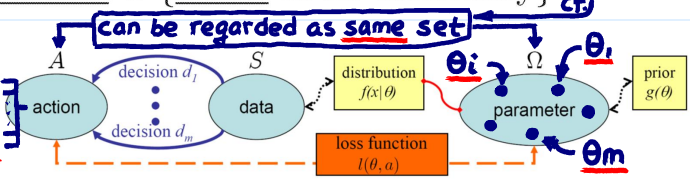
A population consists of m classes, on the basis of some measurements, we wish to classify an individual as belonging to one of the classes. For example:

- (physiological measurements), a patient $\in$ {having disease, not having}
- (radar measurements), an aircraft $\in$ {enemy, friendly, unknown}
- (sound measurements), an utterance $\in$ {words in its dictionary}

collection of all individuals; individual; population; data; data; individual; classes; Estimation (LNp.14); cf.

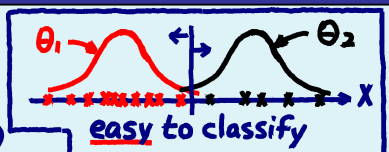


a_1: $\in$ class θ_1
a_m: $\in$ class θ_m



Formulation of Classification Problem (2nd Ed., TBp.581)

- Data: X ($\in S$) = the measurements of one particular individual
- $\Omega = \{\theta_1, \theta_2, \ldots, \theta_m\}$: class membership. → θ_i: 一個平行世界 (unknown); usually, finite classes
- Bayesian: r.v. $\Theta = \theta_i$ ($\in \Omega$), if membership in ith class, $i = 1, 2, \ldots, m$
- $\pi_1, \pi_2, \ldots, \pi_m$: prior probabilities; e.g. may use proportions of individuals in each classes
  $\pi_i = P(\Theta = \theta_i)$, $i = 1, 2, \ldots, m$, with $\sum_{i=1}^m \pi_i = 1$.
- $f(x|\theta_i)$: assumed known (in practice, the key to the problem) → Why?
- l_{ij}: loss in classifying a member of class i to class j — $i=j$ → correct classification; $i \neq j$ → misclassification; $l_{ij} = l(\theta_i, a_j)$: $m \times m$ matrix



Theorem 10.8 (Bayes rule for Classification, 2nd Ed., TBp.581)

Q: How should we assign prior in classification?
Ans. If available, the proportion of individuals of each classes in the population is a good choice (check LNp.29)

- Observe $X = x$, posterior distribution of Θ is

$$h(\theta_i|x) = P(\Theta = \theta_i \mid X = x) = \frac{\pi_i f(x|\theta_i)}{\sum_k \pi_k f(x|\theta_k)}$$

By Thm10.1 (LNp.12); update; joint, e.g., $P(X=x, \Theta=\theta_i)$; marginal, e.g., $P(X=x)$

- posterior risk of assigning an observation x to the jth class is

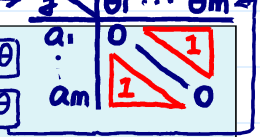
$$\mathrm{PR}(j|x) = \sum_{i=1}^m l_{ij} h(\theta_i|x) = \frac{\sum_{i=1}^m l_{ij} \pi_i f(x|\theta_i)}{\sum_k \pi_k f(x|\theta_k)}$$

average loss of a_j over θ_i's with posteriors as weights

- Bayes rule $d(x)$: choose the value of j for which $\mathrm{PR}(j|x)$ is a minimum. → map to a_j

Example 10.10 (0-1 loss, 2nd Ed., TBp. 581-582)

Q: Why not use L^2-norm? cf.



- 0-1 loss function: (L^0-norm (LNp.14))

$$l_{ij} = \begin{cases} 0, & i = j \leftarrow \text{correct classification} \\ 1, & i \neq j \leftarrow \text{misclassification} \end{cases}$$

cf. quantitative θ; qualitative θ

average loss of θ_i over X with $X|\theta_i$ as weights

- For any classification rule $d(X)$, the risk function is

Then, if take Frequentist approach, can use minimax to choose a d

$$R(i,d) = E_X[l(\theta_i, d(X))] = \sum_{j=1}^m l_{ij} P_{\theta_i}(d(X) = j) = \sum_{j \neq i} P_{\theta_i}(d(X) = j) \quad (*)$$
$$= 1 - P_{\theta_i}(d(X) = i),$$

θ_i; $f(x|\theta_i)$; E_j (LNp.29); a_j; a_i; θ_i

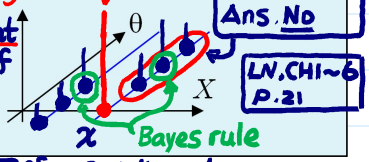
which is the probability of misclassification in class i.

$\sum_{j=1}^m P_{\theta_i}(d(X)=j) = 1$

- Posterior risk of assigning an observation x to the jth class:

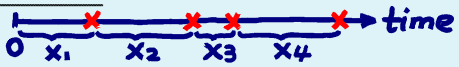
$$\mathrm{PR}(j|x) = \sum_{i=1}^{m} l_{ij}\, h(\theta_i|x) = \sum_{i\neq j} h(\theta_i|x) = \frac{\sum_{i\neq j} \pi_i f(x|\theta_i)}{\sum_{k=1}^{m} \pi_k f(x|\theta_k)} = P_{\Theta|X}(\Theta \neq \theta_j|\, x) = 1 - P_{\Theta|X}(\Theta = \theta_j|\, x) = 1 - h(\theta_j|x)$$

- $\mathrm{PR}(j|x)$ is minimized for that value of j such that $P_{\Theta|X}(\Theta = \theta_j|\, x)$ is maximized, i.e., for the class that has maximum posterior probability.



Example 10.11 (waiting times between emissions, 2nd Ed., TBp. 582)

- Data: $\mathbf{X} = (X_1, X_2, \ldots, X_n) = n$ waiting times between emissions of alpha particles from a radioactive substance.



- On the basis of $\mathbf{X}$, a decision is to be made as to whether to classify the particles as coming from substance I or substance II.
- Statistical Modeling:
  - $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim E(\theta_1)$ if particles came from substance I
  - $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim E(\theta_2)$ if particles came from substance II
  - prior: $\pi_1 = \pi_2 = 1/2$
- loss function: 0-1 loss