Model: $f(\underset{\sim}{x}) = \lambda_1 f_1(\underset{\sim}{x}, \underset{\sim}{\theta}_1) + \cdots + \lambda_k f_k(\underset{\sim}{x}, \underset{\sim}{\theta}_k)$

Ex: Mixtures of multivariate normals

$$f_j(\underset{\sim}{x}, \underset{\sim}{\mu}_j, C_j) = \frac{1}{(2\pi)^{p/2}|C_j|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_j)^T C_j^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_j)\right)$$

$\hat{\Delta}_{ij}$ = est'd prob that $\underset{\sim}{x}_i$ belongs to cluster j.

Given $\{\hat{\Delta}_{ij}\}$ we have $\hat{\underset{\sim}{\mu}}_j = \dfrac{\sum_{i=1}^n \hat{\Delta}_{ij} \underset{\sim}{x}_i}{\sum_{i=1}^n \hat{\Delta}_{ij}}$ ← estimate of # of $\underset{\sim}{x}_i$ in cluster j.

$$\hat{C}_j = \frac{1}{\sum_{i=1}^n \hat{\Delta}_{ij}} \sum_{i=1}^n \hat{\Delta}_{ij} (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_j)(\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_j)^T$$

⇒ repeat until convergence.

Toy example: p=2, k=2

$C_1 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ $C_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$

n=400 (200 each from the 2 ppl'n)

EM R function: very slow & inefficient
- use k-means clustering to give initial values

$\hat{\underset{\sim}{\mu}}_1 = \begin{pmatrix} 0.19 \\ -0.20 \end{pmatrix}$ $\hat{C}_1 = \begin{pmatrix} 1.16 & 0.62 \\ 0.62 & 1.76 \end{pmatrix}$

$\hat{\underset{\sim}{\mu}}_2 = \begin{pmatrix} 3.25 \\ -3.38 \end{pmatrix}$ $\hat{C}_2 = \begin{pmatrix} 0.84 & -0.74 \\ -0.74 & 1.52 \end{pmatrix}$

- after 50 iterations of EM

$\hat{\underset{\sim}{\mu}}_1 = \begin{pmatrix} -0.01 \\ -0.07 \end{pmatrix}$ $\hat{C}_1 = \begin{pmatrix} 0.97 & 0.93 \\ 0.93 & 1.82 \end{pmatrix}$

$\hat{\underset{\sim}{\mu}}_2 = \begin{pmatrix} 2.98 \\ -3.03 \end{pmatrix}$ $\hat{C}_2 = \begin{pmatrix} 1.15 & -1.14 \\ -1.14 & 2.00 \end{pmatrix}$

Classification:
- regression with a categorical (or group) response.

So far: Unsupervised learning → given multivariate data $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$ from some ppl'n which may consist of sub-ppl'ns
- no info about sub-ppl'ns.

Tools: PCA, ICA } Identify interesting low dimensional projections

Cluster Analysis: identify clusters within data

Supervised learning: For each observation $\underset{\sim}{x}_i$ we have a group/cluster/sub-ppl'n identifier g_i.
⇒ observe $(g_1, \underset{\sim}{x}_1), \cdots, (g_n, \underset{\sim}{x}_n)$

- Assume that $g_1, \dots, g_n$ take k known values/variables.

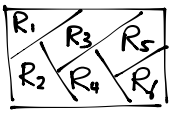
Ex: ① Species identification/classifications
- given measurements $\underset{\sim}{X}_i$, determine species

② spam filter
- given attributes (features) of an e-mail, determine whether it is "good" or spam

Model: $(G, \underset{\sim}{X})$ has a joint distribution described by $P(G=j) = \lambda_j$ for $j = 1, \dots, k$.
and conditional on $G=j$, $\underset{\sim}{X}$ has density $f_j(\underset{\sim}{x})$
Given $\underset{\sim}{X} = \underset{\sim}{x}$, determine conditional dist'n of G i.e. $P(G=j \mid \underset{\sim}{X} = \underset{\sim}{x})$
Marginal density of $\underset{\sim}{x}$:

$$f(\underset{\sim}{x}) = \lambda_1 f_1(\underset{\sim}{x}) + \cdots + \lambda_k f_k(\underset{\sim}{x})$$

Data problem: Given data $\overbrace{(g_1, \underset{\sim}{X}_1), \dots, (g_n, \underset{\sim}{X}_n)}^{\text{training data}}$. Determine a classification rule with small error rate.



← space of $\underset{\sim}{X}$ (feature space)

$R_1, \dots, R_k$ are disjoint regions of $\mathbb{R}^p$

Rule: If $\underset{\sim}{x} \in R_j$ then classify $\underset{\sim}{x}$ as belonging to group j.

Choose $R_1 \cdots R_k$ to minimize classification error rate.

$$\text{error rate} = P(\text{error}) = \sum_{j=1}^k P(G=j)\, P(\text{error} \mid G=j) = \sum_{j=1}^k \lambda_j P(\underset{\sim}{X} \notin R_j \mid G=j)$$
$$= \sum_{j=1}^k \lambda_j \left\{ 1 - \int_{R_j} f_j(\underset{\sim}{x})\, d\underset{\sim}{x} \right\}$$
$$= 1 - \underbrace{\sum_{j=1}^k \lambda_j \int_{R_j} f_j(\underset{\sim}{x})\, d\underset{\sim}{x}}_{P(\text{correct classification})}$$

Optimal classification rule: (Bayes classifier)
Classify $\underset{\sim}{x}$ to group j if for all $i \neq j$, $\lambda_j f_j(\underset{\sim}{x}) > \lambda_i f_i(\underset{\sim}{x})$

$$\frac{f_j(\underset{\sim}{x})}{f_i(\underset{\sim}{x})} > \frac{\lambda_i}{\lambda_j} \quad \text{for all } i \neq j$$

Proof for $k=2$: $R_2 = R_1^c$

Want to maximize $\lambda_1 \int_{R_1} f_1(\underset{\sim}{x})\, d\underset{\sim}{x} + \lambda_2 \int_{R_1^c} f_2(\underset{\sim}{x})\, d\underset{\sim}{x}$

$$= \lambda_1 \int_{R_1} f_1(\underset{\sim}{x})\, d\underset{\sim}{x} + \lambda_2 \left(1 - \int_{R_1} f_2(\underset{\sim}{x})\, d\underset{\sim}{x}\right)$$
$$= \int_{R_1} \underbrace{[\lambda_1 f_1(\underset{\sim}{x}) - \lambda_2 f_2(\underset{\sim}{x})]}_{\geq 0}\, d\underset{\sim}{x} + \lambda_2$$

- maximized over R_1 for $R_1 = \{\underset{\sim}{x} : \lambda_1 f_1(\underset{\sim}{x}) - \lambda_2 f_2(\underset{\sim}{x}) > 0\}$
$= \{\underset{\sim}{x} : \lambda_1 f_1(\underset{\sim}{x}) > \lambda_2 f_2(\underset{\sim}{x})\}$

Note: In this proof, we don't use the fact that $\lambda_1 + \lambda_2 = 1$
In general, result holds for any non-negative $\lambda_1 \cdots \lambda_k \Rightarrow$ minimizing weighted error rate.
- useful if different misclassifications have different costs

-avoid identifying good emails as spam, for example.

Application: $f_1(\underline{x}) \cdots f_k(\underline{x})$ are multivariate normal with different means $\underline{\mu}_1, \cdots, \underline{\mu}_k$ but equal covariance matrices (call it C)

Then $f_j(\underline{x}) = \frac{1}{(2\pi)^{p/2}|C|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_j)^T C^{-1}(\underline{x}-\underline{\mu}_j)\right\}$

optimal classification rule: $\underline{x} \to j$ if $\lambda_j f_j(\underline{x}) > \lambda_i f_i(\underline{x})$ for all $i \neq j$
or $\ln(f_j(\underline{x})/f_i(\underline{x})) > \ln(\lambda_j/\lambda_i)$

$$\ln f_j(\underline{x}) - \ln f_i(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{\mu}_j)^T C^{-1}(\underline{x}-\underline{\mu}_j) + \frac{1}{2}(\underline{x}-\underline{\mu}_i)^T C^{-1}(\underline{x}-\underline{\mu}_i)$$
$$= [\underline{x}^T C^{-1}\underline{\mu}_j - \tfrac{1}{2}\underline{\mu}_j^T C^{-1}\underline{\mu}_j - (\underline{x}^T C^{-1}\underline{\mu}_i - \tfrac{1}{2}\underline{\mu}_i^T C^{-1}\underline{\mu}_i)]$$

Define discriminant scores:
linear function of $\underline{x}$: $d_j(\underline{x}) = \underline{x}^T C^{-1}\underline{\mu}_j - \frac{1}{2}\underline{\mu}_j^T C^{-1}\underline{\mu}_j + \ln(\lambda_j)$
$\Rightarrow$ classify $\underline{x} \to j$ if $d_j(\underline{x}) > d_i(\underline{x})$ for all $i \neq j$

Special case: $k=2$
Suffices to look at $d_2(\underline{x}) - d_1(\underline{x}) = \beta_0 + \underline{x}^T\underline{\beta}_1$ ← depends on $C, \underline{\mu}_1, \underline{\mu}_2$ and λ_1, λ_2

Binary regression models
Ex: logistic regression $\theta(\underline{x}) = P(\text{group } 2 \mid \underline{x})$

$$\ln\left(\frac{\theta(\underline{x})}{1-\theta(\underline{x})}\right) = \beta_0 + \underline{x}^T\underline{\beta}$$

-suggests close relationship between linear discriminant analysis + binary regression models.

How to apply to data?
Given $(g_1, \underline{x}_1), \cdots, (g_n, \underline{x}_n)$ can estimate $\underline{\mu}_1, \cdots, \underline{\mu}_k$ and C as follows:

$$\hat{\underline{\mu}}_j = \sum_{i=1}^n \underline{x}_i I(g_i = j) / \sum_{i=1}^n I(g_i = j)$$
$I(g_i=j) = \begin{cases} 1 & \text{if } g_i = j \\ 0 & \text{o.w.} \end{cases}$

$$\hat{C} = \frac{1}{n-k}\sum_{i=1}^n (\underline{x}_i - \hat{\underline{\mu}}_j(i))(\underline{x}_i - \hat{\underline{\mu}}_j(i))^T$$

$\frac{1}{n-k}$ can also put here as $\frac{1}{n}$

$\hat{\underline{\mu}}_j(i)$: group associated with $\underline{x}_i$

Question: ① How to estimate $\lambda_1 \cdots \lambda_k$?
· assume $\lambda_1, \cdots, \lambda_k$ are known (often true) or est'd "out of sample"
- allows biased samplings
- estimate $\lambda_1 \cdots \lambda_k$ from sample proportions

$$\hat{\lambda}_j = \frac{1}{n}\sum_{i=1}^n I(g_i = j)$$

- need to assume that $(g_1, \underline{x}_1) \cdots (g_n, \underline{x}_n)$ is some sort of representation sample from population.

② How to estimate error rate for a given classifier estimated from $(g_1, \underline{x}_1) \cdots (g_n, \underline{x}_n)$?

Naive method: Resubstitution estimate for each x_i in sample, we obtain classification $\hat{g}_1, \cdots, \hat{g}_n$

error rate $= \frac{1}{n} \sum_{i=1}^{n} I(g_i \neq \hat{g}_i)$ = proportion of misclassified observations.

Resubstitution error rate tends to be biased downwards.

How to fix this?

- cross-validation: leave out observations from sample, estimate classifier from remaining observed and look at error rate for observations not used to estimate the classifier
  e.g. leave-one-out cross-validation.