V. Classification and Clustering

V.A. Discriminant (and Classification) Analysis

So-called “discriminant analysis” has two basic goals:

1. Describe the separation of groups using p-variate observations on observations within groups
2. Prediction — allocate observations from unknown groups into one of the groups.

Often the first task is referred to as “discriminant analysis” or “canonical discriminant analysis.”

Second task called “discriminant analysis”, “classification analysis”, “allocation”, or “supervised classification.” ← machine learning term

We’ll refer to task (1) as “discriminant analysis” and task (2) as “classification analysis.”

Note that both tasks require information on group membership. "Cluster Analysis" is a different method which assumes no formal group membership, but rather looks for natural clusters of observations.

In machine learning, what we call "classification analysis" is called "supervised classification," and what we call "cluster analysis" is called "unsupervised classification" or "unsupervised clustering."

→ "Supervised/unsupervised" refers to whether or not group membership from some "training data" is given

V.A.i. Describing Group Separation

Mostly review ...

- Separation of Two Groups

$$\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

  $\mathbf{a} = \mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ contains the discriminant function coefficients for the discriminant function:

$$z = \mathbf{a}'\mathbf{x}$$

  → $z = \mathbf{a}'\mathbf{x}$ is the linear combination of the x's that maximizes the distance between transformed group means

$$\mathbf{a} = \underset{\mathbf{b} \neq \mathbf{0}}{\operatorname{argmax}} \frac{[\mathbf{b}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{b}'\mathbf{S}_{p\ell}\mathbf{b}}$$

- Separation of g groups

$$\begin{aligned}\mathbf{x}_1 &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\\ \mathbf{x}_2 &\sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\\ &\vdots\\ \mathbf{x}_g &\sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})\end{aligned}$$

  Consider $s = \min(g-1, p)$ different discriminant functions which describe the separation of the g group means in the s-space. The s functions are $\mathbf{a}_i,\ i = 1, \ldots, s$, the s eigenvectors of $\mathbf{E}^{-1}\mathbf{H}$.

  - When group means are collinear, only one discriminant function is needed to describe separation of means.

- When group mean vectors have "essential dimension" of only $r < s$, we need only r of the discriminant functions to describe group separation.
  - E.g., means lie on a 2-D disc $\Rightarrow r = 2$;
    means lie in a 3-D ellipse $\Rightarrow r = 3$; etc.
  - "Essential dimension" of group means indicated by $\lambda_1, \ldots, \lambda_s$.
    When $\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{s} \lambda_i} > c$ (where c is .80, .85, .90, etc.)
    $\Rightarrow$ "essential dimension" is r.

Goals:

1. Identify variables important to the separation of the groups
2. Project data onto lower-dimensional space that is optimal for illustrating group separation (i.e., plot $z_1 = \mathbf{a}_1'\mathbf{x}$ vs. $z_2 = \mathbf{a}_2'\mathbf{x}$)

Interpreting Discriminant Function Coefficients

- Use standardized coefficients
  - 2 group case:

$$\mathbf{a}^* = \mathbf{D}^{1/2}\mathbf{a} \qquad \text{where } \mathbf{D} = \begin{bmatrix} s_{11} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & s_{pp} \end{bmatrix}$$

  - g group case:

$$\mathbf{a}_i^* = \mathbf{D}^{1/2}\mathbf{a}_i, \ i - 1, \ldots, s$$

$$\text{where } \mathbf{D} = \frac{1}{\nu_E} \begin{bmatrix} e_{11} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & e_{pp} \end{bmatrix}$$

$$= \frac{1}{\nu_E}\text{diag}(\mathbf{E}) \quad [\text{or } \frac{1}{\nu_W} \ \text{diag}(\mathbf{W})]$$

- Don't use corr(x_i, z_1) which is proportional to the univariate t-test or F-test used to compare group means for x_i only (ignoring presence of $x_{i'}, i' \neq i$)

$\boxed{\text{ex}}$ Football helmet data (see p. IV. 11)

3 groups of subjects:

group1 = high school football players

group2 = college football players

group3 = non-football players

6 variables:

wdim = head width at widest dimension

circum = head circumference

fbeye = front-to-back measure at eye level

eyehd = eye-to-top-of-head measure

earhd = ear-to-top-of-head measure

jaw = jaw width

$$\left.\begin{matrix} p=6 \\ g=3 \end{matrix}\right\} \Rightarrow s=2$$

		coefficients in a_1	cofficients in $\underset{\sim}{a}_1^*$
↔	wdim	−.631	−.413
↔	circum	.002	.004
↔	fbeye	.004	.003
↕	eyehd	.431	$\boxed{.478}$
↕	earhd	.336	.264
↔	jaw	.551	.338

$\lambda_1 = 1.9178$ $\lambda_2 = .1159$

$$\frac{\lambda_1}{\lambda_1+\lambda_2} = .94$$

V.A.ii. Foundational Classification Tools: LDA, QDA, KNN

Object: Use measurements $\underset{p\times 1}{\mathbf{x}}$ on observations from known groups $G_1, \ldots, G_g$ in order to create classification rule for allocating or assigning observations of unknown membership to one of the groups.

$\boxed{\text{ex}}$ Use academic information from students who graduated (and who dropped out) in order to predict which applicants are likely to graduate.

$\boxed{\text{ex}}$ Use financial information obtained from loan defaulters and non-defaulters to predict whether or not a future loan applicant should be given a loan.

We'll start with 2 groups and then extend to the $g > 2$ scenario.

Two-Group Classification

Let the random vector $\mathbf{x}$ measured on each observation in group G_i $(i = 1, 2)$ have density $f_i(\mathbf{x})$. We will assign each observation $\mathbf{x}$ to "predicted group" $\hat{G}_1$ or $\hat{G}_2$.

Some terms:

- Misclassification rates

$$
\begin{aligned}
\Pr\{2|1\} &= \Pr\{\mathbf{x} \in \hat{G}_2 | \mathbf{x} \in G_1\} \\
&= \Pr\{\text{erroneously assigning an observation from } G_1 \text{ into } \hat{G}_2\} \\
\Pr\{1|2\} &= \Pr\{\mathbf{x} \in \hat{G}_1 | \mathbf{x} \in G_2\}
\end{aligned}
$$

- Prior probabilities

$$p_1 = \Pr\{\text{observation comes from group } G_1\}$$
$$p_2 = \Pr\{\text{observation comes from group } G_2\}$$

$\boxed{\text{ex}}$ We know that 95% of a class of cysts are benign (5% malignant)

$$\Rightarrow p_1 = .95$$
$$p_2 = .05$$

- Cost of misclassification

$$C\{2|1\} = \text{ cost of erroneously assigning an } \mathbf{x} \text{ from } G_1 \text{ into } \hat{G}_2$$
$$C\{1|2\} = \text{ cost of erroneously assigning an } \mathbf{x} \text{ from } G_2 \text{ into } \hat{G}_1$$

$\boxed{\text{ex}}$ Calling a malignant tumor “benign” (false negative) is worse than calling a benign tumor “malignant” (false positive)

So . . .

$$\begin{aligned}
&\Pr\{\mathbf{x} \text{ is correctly assigned to } G_1\} = \Pr\{1|1\} \cdot p_1\\
&\Pr\{\mathbf{x} \text{ is correctly assigned to } G_2\} = \Pr\{2|2\} \cdot p_2\\
&\Pr\{\mathbf{x} \text{ is incorrectly assigned to } G_1\} = \Pr\{1|2\} \cdot p_2\\
&\Pr\{\mathbf{x} \text{ is incorrectly assigned to } G_2\} = \Pr\{2|1\} \cdot p_1
\end{aligned}$$

An “optimal” classification rule minimizes the “expected cost of misclassification” (ECM) given by

$$\text{ECM} = c\{2|1\} \cdot \Pr\{2|1\} \cdot p_1 + c\{1|2\} \cdot \Pr\{1|2\} \cdot p_2$$

- ECM is minimized with the rule

$$\begin{aligned}\hat{G}_1 &= \{\mathbf{x} : p_1 f_1(\mathbf{x})\ c\{2|1\} > p_2 f_2(\mathbf{x}) c\{1|2\}\} \\ \text{and} & \\ \hat{G}_2 &= \{\mathbf{x} : p_1 f_1(\mathbf{x})\ c\{2|1\} < p_2 f_2(\mathbf{x}) c\{1|2\}\}\end{aligned} \qquad (1)$$

  Note: An observation has an increased tendency to be assigned to $\hat{G}_1$ when:

  - cost of misclassifying an observation from G_1 ($c\{2|1\}$) is large
  - density $f_1(\mathbf{x})$ is large (i.e., G_1 looks "likely")
  - large proportion of observations actually come from G_1

Note:

- Equal misclassification costs

$$\begin{aligned}\hat{G}_1 &= \{\mathbf{x} : p_1 f_1(\mathbf{x}) > p_2 f_2(\mathbf{x})\} \\ \text{and} & \\ \hat{G}_2 &= \{\mathbf{x} : p_1 f_1(\mathbf{x}) < p_2 f_2(\mathbf{x})\}\end{aligned} \tag{2}$$

- Equal misclassification costs and prior probabilities

$$\begin{aligned}\hat{G}_1 &= \{\mathbf{x} : f_1(\mathbf{x}) > f_2(\mathbf{x})\} \\ \text{and} & \\ \hat{G}_2 &= \{\mathbf{x} : f_1(\mathbf{x}) < f_2(\mathbf{x})\}\end{aligned} \tag{3}$$

Approach Based on Discriminant Function

Fisher (1936) proposed using $\mathbf{a} = \mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ to predict group membership. (Recall $\mathbf{a}$ defines a new axis with best separation of groups when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.)

Suppose we wish to assign $\mathbf{x}_0$ to $\hat{G}_1$ or $\hat{G}_2$.

- Assign $\mathbf{x}_0$ to $\hat{G}_1$ if:

  $$z_0 = \mathbf{a}'\mathbf{x}_0 \text{ is closer to } \bar{z}_1 = \mathbf{a}'\bar{\mathbf{x}}_1 \text{ than } \bar{z}_2 = \mathbf{a}'\bar{\mathbf{x}}_2$$

  or

  $$z_0 > \tfrac{1}{2}(\bar{z}_1 + \bar{z}_2)$$

So that the rule is

$$\begin{aligned} \hat{G}_1 &= \left\{ \mathbf{x} : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}\mathbf{x} > \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \right\} \\ &\text{and} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4) \\ \hat{G}_2 &= \left\{ \mathbf{x} : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}\mathbf{x} < \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \right\} \end{aligned}$$

Though originally proposed as a distribution-free approach, when

(i) f_1 and f_2 are normal densities

(ii) covariance matrices associated with f_1 and f_2 are equal ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$)

(iii) $p_1 = p_2$

(iv) $c\{1|2\} = c\{2|1\}$

then

> Fisher's rule (4) is equivalent to rule (3) and is optimal in the sense of minimizing the probability of misclassification (and ECM).

For the case of $p_1 \neq p_2$ Welch (1939) extended Fisher's rule so that our rule is:

$$\begin{aligned}
\hat{G}_1 &= \left\{\mathbf{x} : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}\mathbf{x} > \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \ln\frac{p_2}{p_1}\right\} \\
&\text{and} \qquad\qquad (5) \\
\hat{G}_2 &= \left\{\mathbf{x} : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}\mathbf{x} < \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \ln\frac{p_2}{p_1}\right\}
\end{aligned}$$

Thus when (i), (ii), and (iv) above hold, rule (5) is equivalent to rule (2) and is optimal in the sense of minimizing the probability of misclassification (and ECM).

Misclassification costs can be incorporated using rule:

$$\begin{aligned}\hat{G}_1 &= \left\{\mathbf{x} : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}\mathbf{x} > \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \ln \frac{p_2 \ c\{1|2\}}{p_1 \ c\{2|1\}}\right\}\\ \text{and} & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (6)\\ \hat{G}_2 &= \left\{\mathbf{x} : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}\mathbf{x} < \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{p\ell}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \ln \frac{p_2 \ c\{1|2\}}{p_1 \ c\{2|1\}}\right\}\end{aligned}$$

Then when (i) and (ii) above hold, rule (6) is equivalent to rule (1) and is optimal in the sense of minimizing ECM.

Notes:

1. When (i) normality and (ii) $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ do not hold, the approaches rule (4), rule (5), and rule (6) may be somewhat reasonable, but they are not optimal. For brevity, we move on to the general g-group case ($g \geq 2$) to discuss approaches when (i) and/or (ii) are invalid. (Of course, these apply to the $g = 2$ case.)

2. For simplicity of discussion, we now consider the equal cost of misclassification case. (Unequal costs substantially complicates the problem when $g > 2$.) Johnson and Wichern provide a good discussion of the unequal costs scenario.

g-group Classification ($g \geq 2$)

Linear Classification Functions ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g$)

Simple idea for $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_g$ case:

> Assign $\mathbf{x}$ to group $\hat{G}_i$ if the multivariate (Mahalanobis-like) distance between $\mathbf{x}$ and $\bar{\mathbf{x}}_i$ is minimized.

Formally,

> assign $\mathbf{x}$ to group $\hat{G}_{i'}$ if
>
> $$D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_{p\ell}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)$$
>
> is minimized when $i = i'$.

Note,

$$D_i^2(\mathbf{x}) = \underbrace{\mathbf{x}'\mathbf{S}_{p\ell}^{-1}\mathbf{x}}_{\substack{\text{same for} \\ \text{all } G_i\text{—} \\ \text{can be ignored}}} -2\, \underbrace{\bar{\mathbf{x}}_i'\mathbf{S}_{p\ell}^{-1}\mathbf{x}}_{\substack{\text{linear fcn.} \\ \text{of } \mathbf{x}}} + \underbrace{\bar{\mathbf{x}}_i'\mathbf{S}_{p\ell}^{-1}\bar{\mathbf{x}}_i}_{\substack{\text{doesn't} \\ \text{involve } \mathbf{x}\text{—} \\ \text{like an intercept} \\ \text{for } i\text{th group}}}$$

Ignoring first term of $D_i^2(\mathbf{x})$ and multiplying by $-\frac{1}{2}$, we obtain the "linear classification functions"

$$L_i(\mathbf{x}) = \underbrace{\bar{\mathbf{x}}_i' \mathbf{S}_{p\ell}^{-1}}_{\substack{\mathbf{c}_i':\\ \text{defines}\\ \text{linear comb.}}} \mathbf{x} \underbrace{- \frac{1}{2} \bar{\mathbf{x}}_i' \mathbf{S}_{p\ell}^{-1} \bar{\mathbf{x}}_i,}_{\substack{c_{0,i}:\\ \text{"intercept"}\\ \text{or}\\ \text{"constant"}}} \quad i = 1, \ldots, g$$

Our "linear classification rule" is:

Assign $\mathbf{x}$ to $\hat{G}_{i'}$ if $L_i(\mathbf{x})$ is maximized when $i = i'$.

This rule minimizes probability of misclassification when densities $f_1(\mathbf{x}), \ldots, f_g(\mathbf{x})$ are normal with equal priors $p_1 = \cdots = p_g$.

If prior probabilities are known (and unequal) and densities are normal, the linear classification functions are modified to be

$$L_i^*(\mathbf{x}) = \underbrace{\bar{\mathbf{x}}_i' \mathbf{S}_{p\ell}^{-1}}_{\substack{\mathbf{c}_i': \\ \text{defines} \\ \text{linear comb.}}} \mathbf{x} \underbrace{- \frac{1}{2} \bar{\mathbf{x}}_i' \mathbf{S}_{p\ell}^{-1} \bar{\mathbf{x}}_i + \ln p_i,}_{\substack{c_{0,i}: \\ \text{“intercept”} \\ \text{or} \\ \text{“constant”}}} \quad i = 1, \ldots, g$$

and our optimal (minimal-probability-of-misclassification) rule is:

Assign $\mathbf{x}$ to $\hat{G}_{i'}$ if $L_i^*(\mathbf{x})$ is maximized when $i = i'$.

Note:

- The $L_i(\mathbf{x})$ functions can be obtained without parametric assumption — merely minimizes a distance D_i^2. (Rule is optimal for normal data when $p_1 = \cdots = p_g$.)

- The $L_i^*(\mathbf{x})$ functions require normality for $f_1(\mathbf{x}), \ldots, f_g(\mathbf{x})$ to derive. (Rule is optimal for normal data.)

- While “discriminant analysis” requires only a few of the $s = \min(p, g-1)$ discriminant functions, “classification analysis” requires creation and use of all g classification functions.

Quadratic Classification Functions ($\mathbf{\Sigma}_i$'s not equal)

When covariance matrices are unequal, observations tend to be classified too frequently in the groups with "small" variance-covariance matrices.

Simple idea for $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_g$ not all equal:

Assign $\mathbf{x}$ to group G_i if the multivariate distance

$$D_i^2(\mathbf{x}) = \underbrace{(\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i)}_{\substack{\text{quadratic function of } \mathbf{x} \\ \text{that cannot be reduced} \\ \text{to a linear function}}}$$

Thus, rules based on $\mathbf{S}_i$ called "quadratic classification rules." Rule is reasonable but not optimal.

Assuming normality for $f_1(\mathbf{x}), \ldots, f_g(\mathbf{x})$ with prior probabilities $p_1, \ldots, p_g$, we can derive our optimal (minimal-probability-of-misclassification) rule:

Assign $\mathbf{x}$ to $\hat{G}_{i'}$ if

$$Q_i(\mathbf{x}) = \ln p_i - \frac{1}{2} \ln |\mathbf{S}_i| - \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)' \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i)$$

is maximized when $i = i'$.

Nonparametric Classification Rules

- k Nearest Neighbors Classification Rule
  - Calculate distance of observation $\mathbf{x}$ to all other points $\mathbf{x}_i \neq \mathbf{x}$:

$$(\mathbf{x}-\mathbf{x}_i)'\mathbf{S}_{p\ell}^{-1}(\mathbf{x}-\mathbf{x}_i)$$

  - k_i of the k nearest neighbors are from group G_i, so $\sum_{i=1}^{g} k_i = k$
  - Rule:

    Assign $\mathbf{x}$ to $\hat{G}_i$ if $k_i = \max\limits_j k_j$ (If there is a tie for largest k_i, do not classify $\mathbf{x}$ into any $\hat{G}_i$)

    - Rule assumes $p_i = \frac{n_i}{n},\ i = 1, \ldots, g$. If $p_i \neq \frac{n_i}{n}$, consider the revised rule:

      Assign $\mathbf{x}$ to $\hat{G}_i$ if $\frac{k_i p_i}{n_i} = \max\limits_j \frac{k_j p_j}{n_j}$
  - Choice of k
    - Use k with best error rate
    - $k \approx \sqrt{n_i}$ (Loftsgaarden and Quesenberry, 1965)

- Classification Based on Density Estimates
  - Use kernal density estimators to obtain $\hat{f}_1(\mathbf{x}), \ldots, \hat{f}_g(\mathbf{x})$
  - Rule:

    Assign $\mathbf{x}$ to group $\hat{G}_{i'}$ if $p_i \hat{f}_i(\mathbf{x})$ is maximized when $i = i'$.

Evaluating Classification Functions

AER = Actual error rate (aka "test error rate")

= Probability that classification functions based on current sample will misclassify a future observation.

→ Want an estimate of AER

- Resubstitution

  Use classification rules based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to predict group membership for each observation $\mathbf{x}_1, \ldots, \mathbf{x}_n$

  – Among the n_i observations from group G_i, n_{ij} are classified into $\hat{G}_j$ so that $n_i = \sum_{j=1}^{g} n_{ij}$

  – APCCR = Apparent correct classification rate

  $$= \frac{\sum_{i=1}^{g} n_{ii}}{n} = \frac{\text{total \# correctly classified}}{\text{total \# of observations}}$$

  – APER = Apparent error rate

  $$= 1 - \frac{\sum_{i=1}^{g} n_{ii}}{n} = 1 - \text{ APCCR}$$

  APER underestimates AER since observations used to create rules also used to evaluate rules

- Holdout Method (aka "Crossvalidation" or "n-fold crossvalidation")
  - All but one observation ($\mathbf{x}_i$) is used to create the classification rule
  - $\mathbf{x}_i$ is classified into one of the g groups using the rule just calculated
  - Repeat process for $i = 1, \ldots, n$
  - Among the n_i observations from group G_i, n_{ij} are classified into $\hat{G}_j$
  - "Expected AER" $= 1 - \frac{\sum_{i=1}^{g} n_{ii}}{n}$ is a better estimate of AER.

- k-fold crossvalidation (e.g., 10-fold CV)
  - Process:
    - $*$ Break training data into k random "folds"
    - $*$ Treat each holdout fold as test data, fitting model to the remaining $k-1$ folds
    - $*$ Using fitted model, classify observations in the holdout fold and calculate test error ("Expected AER")
    - $*$ Repeat process for $i = 1, \ldots, k$ and use mean test error across all k folds as k-fold CV estimate of the test error
  - Disadvantage of k-fold CV ($k < n$) vs. hold-one-out CV
    - $*$ Estimated test error rate biased high due to only $\frac{k-1}{k}n$ observations used to fit model
  - Advantages of k-fold CV ($k < n$) vs. hold-one-out CV
    - $*$ Lower computational cost — k model fits instead of n
    - $*$ Some claim (??) potential lower variance in estimate of test error rate due to less correlation in the k estimates

$\boxed{\text{ex}}$ Football helmet data

- Evaluate each of the following rules using resubstitution and crossvalidation
  - ∗ Linear classification rule
  - ∗ Quadratic classification rule
  - ∗ 5 nearest neighbors rule

$\boxed{\text{ex}}$ Olive Data (8 areas)

- Linear classification rule
  (priors proportional to sample size)

V.A.iii. “Modern” Classification Tools: Trees, Random Forest, Boosting, SVM, Naive Bayes

V.A.iii.a. Trees & Random Forest

Basic Approach for Forming Classification Trees

- Beginning with all n observations in one group or *root node*
- Find a predictor variable x and an associated cutoff criterion such that splitting the n subjects into the two groups will minimize the *impurity* (diversity) within each of the two *child nodes*
- Continue splitting child nodes into new generations
- Stopping/pruning rules based on cross-validation

Criteria for choosing optimal splits

1. Prior probabilities $p_1, \ldots, p_k$ associated with each group
2. Misclassification cost for classifying an observation from group G_i into group G_j (for $i, j = 1, \ldots, k$)
   - messy for $k > 2$
3. Measure of impurity that is appropriate to scenario
   - Gini index is common

Gini index

Impurity for node A:

$$I_A = \sum_{i=1}^{k} p_{i|A}(1 - p_{i|A}) \tag{7}$$

where $p_{i|A}$ is the probability that an observation is in group G_i given that it is classified into node A

- $I_A \approx 0$ when each of $p_{1|A}, p_{2|A}, \ldots, p_{k|A}$ is either near zero or near 1 (i.e., observations in a node are predominantly from one group)
- $p_{i|A} = \frac{p_i(n_{iA}/n_i)}{\sum_{i=1}^{k} p_i(n_{iA}/n_i)}$
  - p_i: prior probability for G_i
  - n_i: number of observations in G_i
  - n_{iA}: number of observations from G_i that are in node A
  - Note: if prior probabilities are proportional to the size of the group G_i or "PPS" (e.g., training data is random sample from population of interest) then $p_{i|A} = \frac{n_{iA}}{n_A}$, where n_A = number of observations in node A.

Using Impurity to Choose Optimal Split

Probability of an observation being in node A:

$$p_A = \sum_{i=1}^{k} p_i(n_{iA}/n_i) \overset{\text{PPS}}{=} n_A/n$$

Optimal split of parent node A into two child nodes A_L and A_R maximizes change in impurity:

$$\Delta I = p_A I_A - (p_{A_L} I_{A_L} + p_{A_R} I_{A_R}),$$

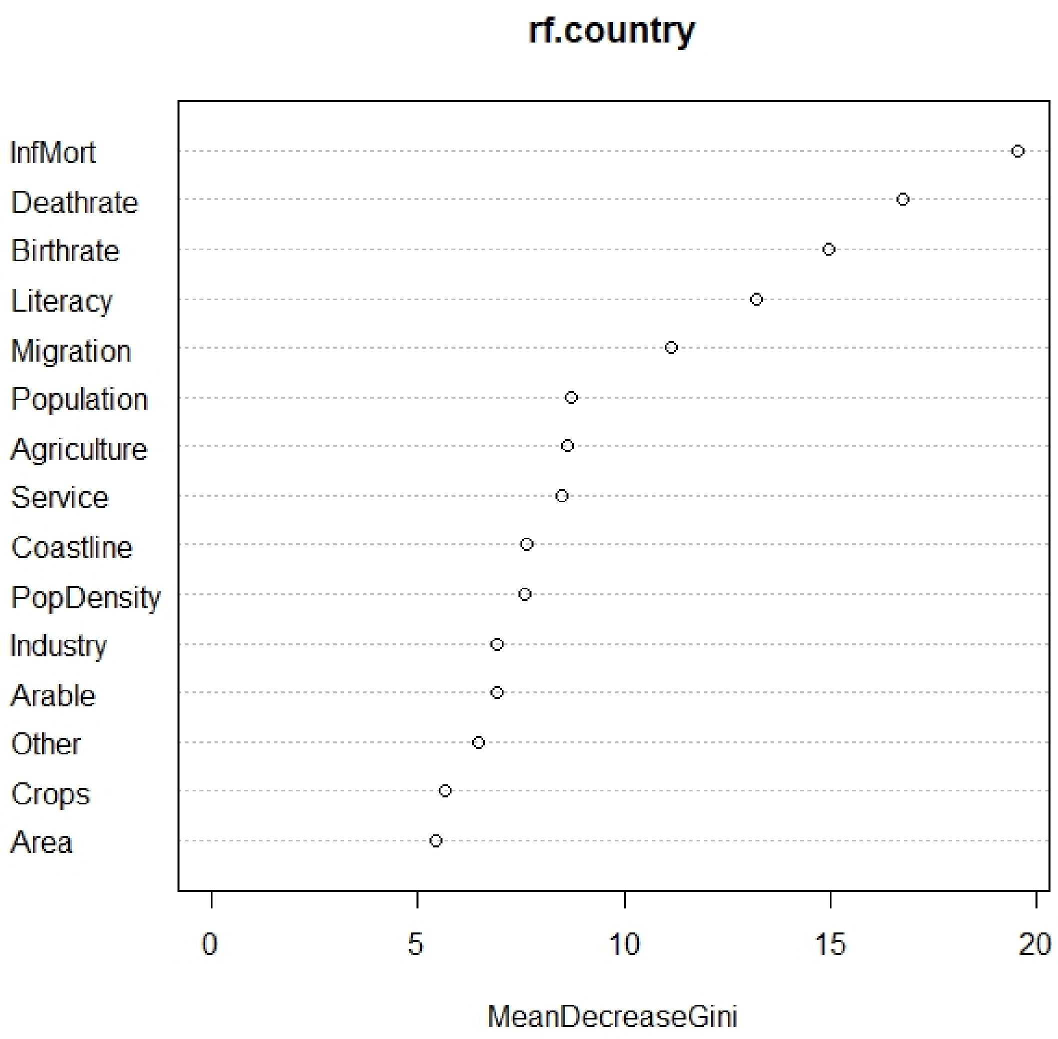
- ΔI guaranteed to be ≥ 0

$\boxed{\text{ex}}$ Predict a country's global region from its demographics

Random Forest

- Bagging: build ensemble of trees based on bootstrapped samples of training data (let each tree "vote" on prediction)
- Random Forest: similar to bagging, but at each split, only m of the total p predictors are considered as splitting variable
  - $m = p$: RF = Bagging
  - $m < p$: Leads to an ensemble that is less correlated across trees $\Rightarrow$ (generally) more stable predictions with lower test error
  - Choosing m:
    - Classification: $m \approx \sqrt{p}$
    - Regression (quantitative response): $m \approx p/3$

– Checking variable importance: calculate total decrease in node impurities from splitting on the variable, averaged over all trees



ex Predict a country's global region from its demographics

V.A.iii.b. Boosting

Rough idea (details omitted)

- Like bagging in that we use multiple trees, but trees are formed sequentially instead of in parallel
  - "Weak learner" trees are generally a single split or "stump"
  - The speed of learning can be adapted—slow learning generally yields better results
- Begin with all observations in the training set having equal weights
- Observations that were misclassified in the previous iteration are upweighted

  $\boxed{\text{ex}}$ Predict a country's global region from its demographics

- xgboost: Variation of boosting that fits trees to the residuals from the previous model (instead of updating weights) to adapt each successive tree
  - When data are traditional data tables, xgboost has been the most successful tool for classification problems (e.g., Kaggle competitions)
    - For “unstructured” data like text, images, video, speech: neural networks are generally the best performers
  - Implementing xgboost in R can be complicated

V.A.iii.d. Support Vector Machines (SVM)

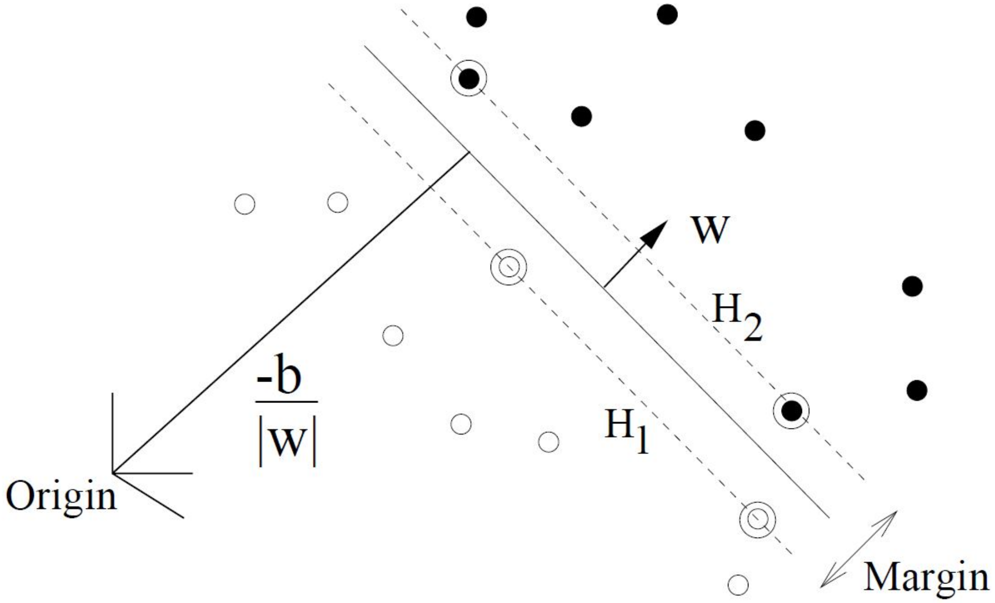
Maximal Margin Classifiers (aka Separable Linear SVM)

Consider p-dimensional observations from two groups (labeled $y = -1$ and $y = 1$...multi-class SVM considered later). Suppose that the groups are neatly separated by a $(p-1)$-dimensional hyperplane defined by:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

- Problem: if there exists one such hyperplane, there exist an infinite number
- Maximal margin hyperplane is the hyperplane with the largest distance to the data
  - separating hyperplane depends only on a small number of points of data (as few as 3)
  - maximal margin classifier may be overfitting the data

Example:

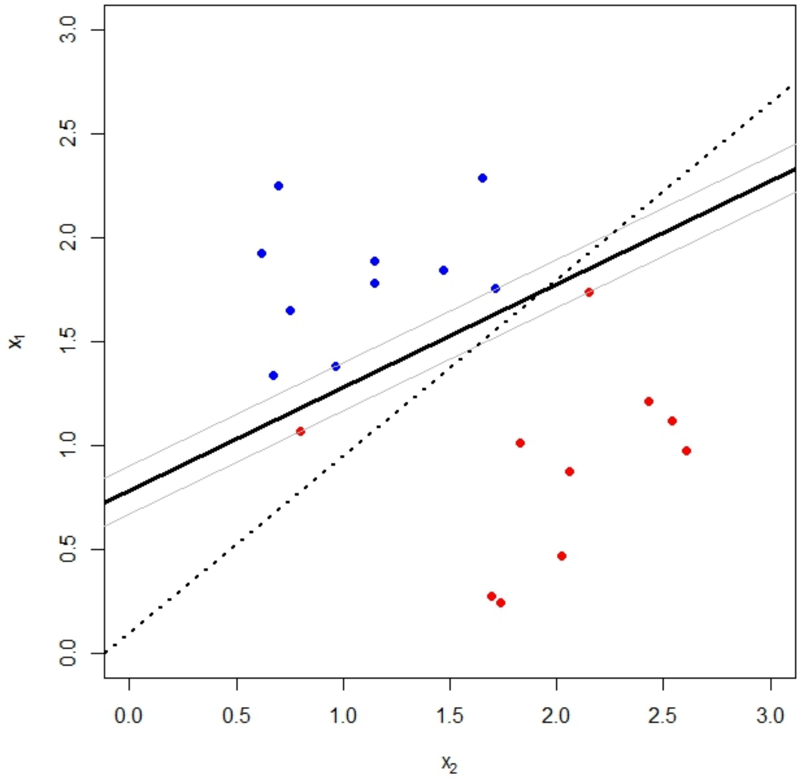


Source: Burges, 1998, *Data Mining and Knowledge Discovery*. b=intercept, $\mathbf{w}$ =vector of slopes.

- Finding the maximal margin classifier:
  - Maximize margin M subject to
    - $\sum_{j=1}^{p} \beta_j^2 = 1$
    - $y_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) > M$ for all $i = 1, ..., n$
- Points defining the hyperplane (≥ 3) are called *support vectors*

Support Vector Classifiers (aka Non-Separable Linear SVM)

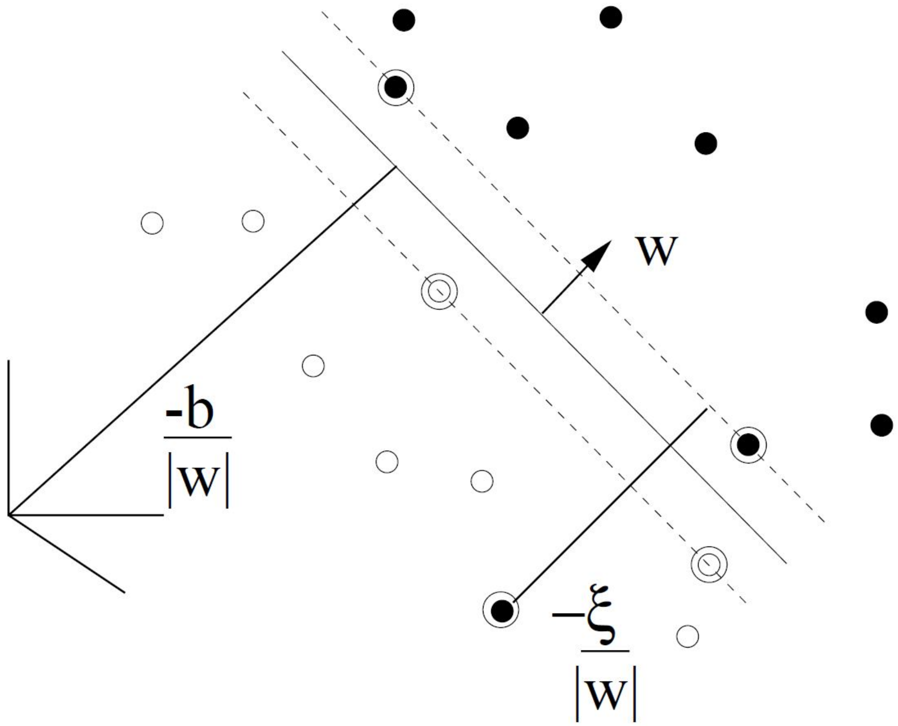
Usually, groups are not perfectly separable, and even if they are, the separating hyperplane (solid line below) may be overly influenced by only one or two points.



- Can we do better out of sample prediction with another classifier (like the dotted line above)?

Finding the support vector classifier:

- Maximize margin M subject to
  - $\sum_{j=1}^{p} \beta_j^2 = 1$
  - $y_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) \geq M(1 - \epsilon_i)$
  - $\epsilon_i \geq 0$
  - $\sum_{i=1}^{n} \epsilon_i \leq C$
- $\epsilon_1, \ldots, \epsilon_n$ are *slack variables* that allow observations to cross the margin boundaries (or even the hyperplane) subject to a total slack *budget* C
  - $\epsilon_i > 0 \Rightarrow$ observation on wrong side of the margin
  - $\epsilon_i > 1 \Rightarrow$ observation on wrong side of the hyperplane



Source: Burges, 1998, *Data Mining and Knowledge Discovery*

- Budget $C = 0$:
  - No violations of margin allowed
  - Yields maximal margin classifier
- As budget C increases:
  - More flexibility with outliers (observations can violate the margin or hyperplane)
  - Margin gets wider
  - More observations (those on margins and those violating margin) affect the classifier—more support vectors...but still unaffected by all other observations
  - More bias / less variance
- `svm` function in `R` uses `cost` parameter instead of budget (C)
  - `cost` inversely related to budget: higher costs ⇒ tighter budget ⇒ fewer violations allowed

Support Vector Machines (aka Non-Separable Nonlinear SVM)

What about cases where nonlinear boundaries are needed for classification?

- It can be shown that the classifier

  $$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

  can be rewritten as

  $$\begin{aligned} f(\mathbf{x}) &= \beta_0 + \sum_{i=1}^{n} \alpha_i \mathbf{x}'\mathbf{x}_i \\ &= \beta_0 + \sum_{i=1}^{n} \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle \\ &= \beta_0 + \sum_{i \in S} \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle \end{aligned}$$

  where the final equality holds because $\alpha_i = 0$ unless $\mathbf{x}_i$ is in the set S of support vectors

- The inner product $\langle \mathbf{x}, \mathbf{x}_i \rangle = \mathbf{x}'\mathbf{x}_i$ can be generalized to $K(\mathbf{x}, \mathbf{x}_i)$
  - Linear kernel: $K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}'\mathbf{x}_i$
  - Polynomial kernel of degree d: $K(\mathbf{x}, \mathbf{x}_i) = (k_0 + \gamma \mathbf{x}'\mathbf{x}_i)^d$
  - Radial kernel: $K(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma(\mathbf{x} - \mathbf{x}_i)'(\mathbf{x} - \mathbf{x}_i))$

Support Vector Machines for Multi-class Problems

Two approaches when $K > 2$:

- One vs. One (this is used by `svm` in `R`)
  - Run the $\binom{K}{2}$ one-versus-one SVM classifications
  - Assign test observation to the class to which it was most frequently assigned
- One vs. All
  - For $k = 1, \ldots, K$, run the SVM classification comparing class k (coded as $y = +1$) to the collection of observations not in class k (coded as $y = -1$)
  - Assign test observation to the class which maximizes

$$f_k(\mathbf{x}) = \beta_{0k} + \beta_{1k}x_1 + \beta_{2k}x_2 + \ldots + \beta_{pk}x_p$$

    * $f_k(\mathbf{x})$ large $\Rightarrow$ strong evidence for membership in class k

$\boxed{\text{ex}}$ Predict a country's global region from its demographics

V.A.iii.d. Naive Bayes

Simple approach based on a strong (often unrealistic) assumption: conditional on class, the predictor variables are independent. Although the assumption is rarely valid, the classifier has been shown to do well in large data scenarios such as text classification.

Suppose we have a categorical response variable y and categorical predictor variables $X_1, \ldots, X_p$, with y coming from one of the classes $C_1, \ldots, C_k$.

$$
\begin{aligned}
P(C_k|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = & \\
= \quad & \frac{P(x_1, x_2, \ldots, x_p, C_k)}{P(x_1, x_2, \ldots, x_p)} \\
\propto \quad & P(x_1, x_2, \ldots, x_p, C_k) \\
= \quad & P(x_1|x_2, \ldots, x_p, C_k)P(x_2, \ldots, x_p, C_k) \\
\vdots \quad & \\
= \quad & P(x_1|x_2, \ldots, x_p, C_k)P(x_2|x_3, \ldots, x_p, C_k) \ldots P(x_p|C_k)P(C_k) \\
\overset{\text{Cond.Ind.}}{=} \quad & P(x_1|C_k)P(x_2|C_k) \ldots P(x_p|C_k)P(C_k) \\
= \quad & P(C_k) \prod_{i=1}^{p} P(x_i|C_k)
\end{aligned}
$$

$\boxed{\text{ex}}$ Predict a mushroom's class (edible vs. poisonous) using its facets

V.C. Cluster Analysis (A <u>Brief</u> Intro)

Goal: Group objects that are similar into "clusters" using p-variate observation $\mathbf{x}$.

- Usually little is known about structure in the groups (or even if groups exist)
- An "exploratory" method — little in the way of formal inference
- Most clustering approaches based on a measure of similarity/dissimilarity among objects.

Distance/Dissimilarity for Interval-Scaled Data

Desired properties for a distance metric:

- $d(\mathbf{x}, \mathbf{x}) = 0$
- $d(\mathbf{x}, \mathbf{y}) \geq 0$
- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

1. Euclidean distance
   $d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})}$
   Generally preferred to $\sqrt{(\mathbf{x} - \mathbf{y})\mathbf{S}^{-1}(\mathbf{x} - \mathbf{y})}$ since $\mathbf{S}$ cannot be computed without knowledge about group membership

2. Minkowski metric
   $d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$
   - $m = 2$ → Euclidian distance
   - $m = 1$ → "Manhattan" or "city block" distance (sum of component-wise distances)

→ Regardless of dissimilarity metric, scaling of variables can dramatically affect nature of clusters (variables with large variances have disproportionate influence on cluster formation)

- ***We usually standardize variables if they are not commensurate***

Distance/Dissimilarity for Binary Data

e.g.

$$\mathbf{x}' = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

1. Euclidian Distance

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{(0-0)^2 + (1-1)^2 + \cdots + (0-1)^2} \\ &= \sqrt{0 + 0 + \cdots + 1} \\ &= \sqrt{\text{\# of mismatches}} \end{aligned}$$

2. Other Ideas

Create table for p attributes. E.g.,

		$\underset{\sim}{y}$		
		0	1	
$\underset{\sim}{x}$	0	a=2	b=2	4
	1	c=1	d=1	2
		3	3	p=6

$$d_1 = \frac{b+c}{a+b+c+d}$$ = % mismatches

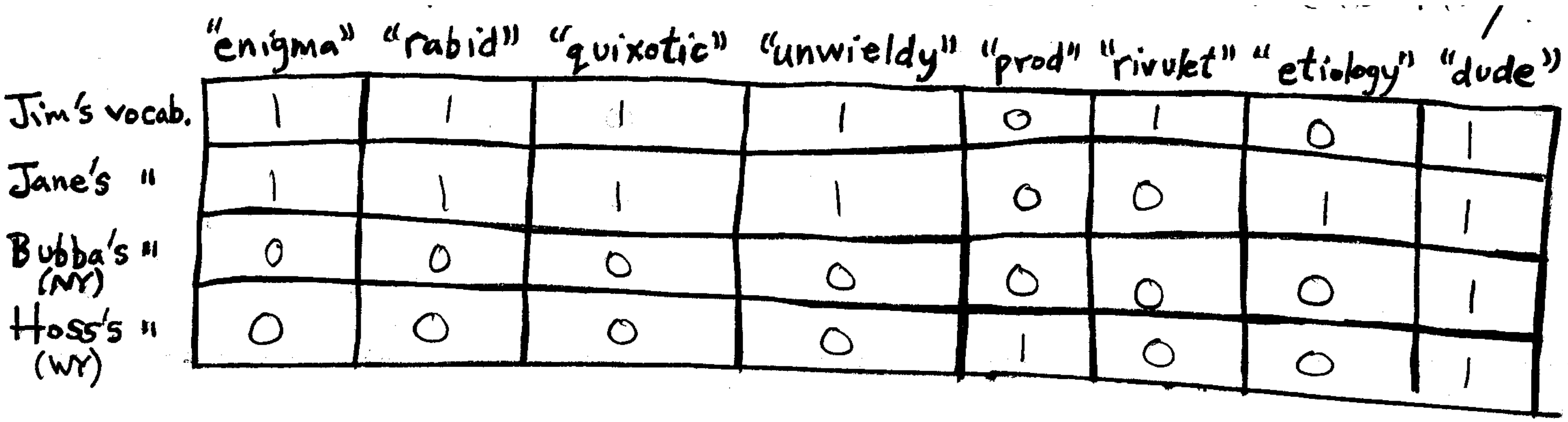
(same weight for 1-1 matches and 0-0 matches)

$$d_2 = \frac{b+c}{b+c+d}$$ = % mismatches among attributes for which at least one object has attribute present

(0-0 matches are treated as irrelevant)

? Which pair has more in common: two people who <u>DO</u> speak Haitian Creole or two people who don't?

$\boxed{\text{ex}}$ Which 2 persons have most similar vocabulary?



	"enigma"	"rabid"	"quixotic"	"unwieldy"	"prod"	"rivulet"	"etiology"	"dude"
Jim's vocab.	1	1	1	1	0	1	0	1
Jane's "	1	1	1	1	0	0	1	1
Bubba's " (NY)	0	0	0	0	0	0	0	1
Hoss's " (WY)	0	0	0	0	1	0	0	1

$d_1(\text{"Jim"}, \text{"Jane"}) = \frac{2}{8} = .25$

$d_1(\text{"Bubba"}, \text{"Hoss"}) = \frac{1}{8} = .125$

$d_2(\text{"Jim"}, \text{"Jane"}) = \frac{2}{7} = .29$

$d_2(\text{"Bubba"}, \text{"Hoss"}) = \frac{1}{2} = .5$

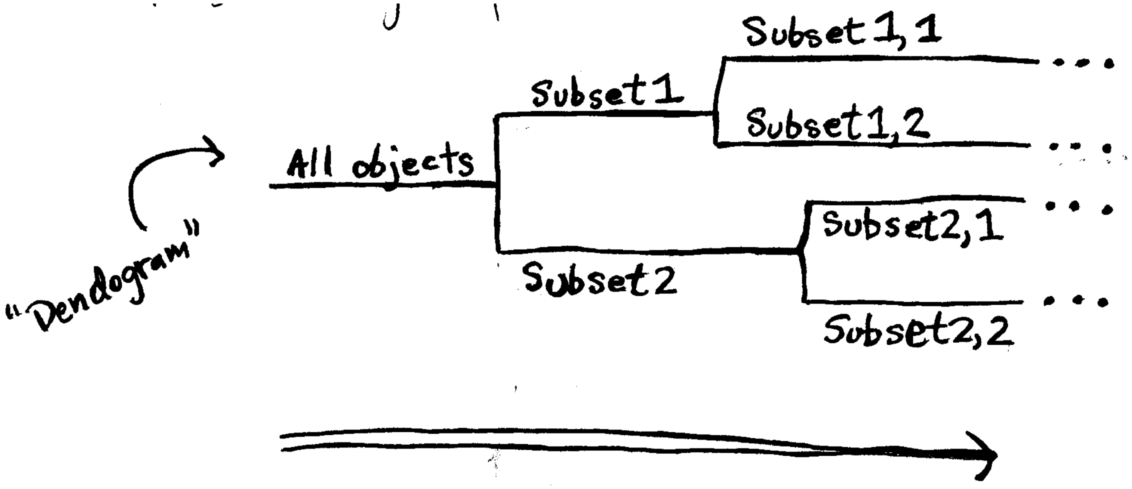
d_1 says that Bubba and Hoss have most similar vocabulary

d_2 says that Jim and Jane have most similar vocabulary

Hierarchical Clustering Methods

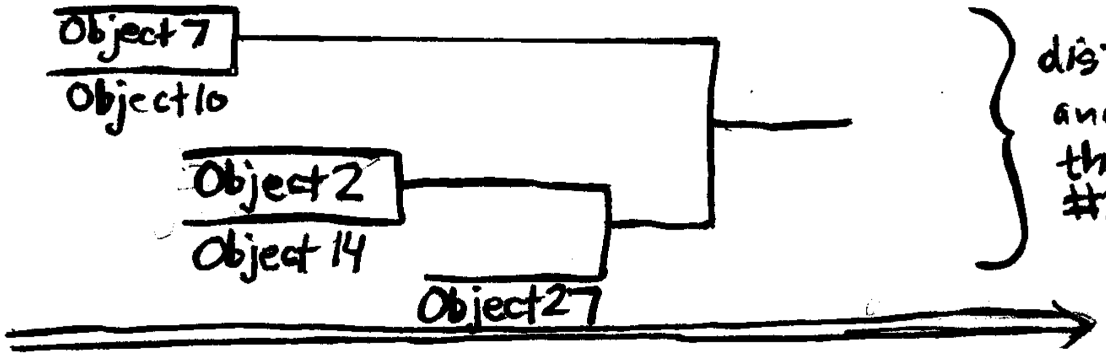
Divisive Hierarchical Methods

- Start with one large group and then successively break groups into dissimilar subgroups
- Process continues until each object is its own group



Agglomerative Hierarchical Methods

- More commonly used
- Start with each object in its own group and then successively combine groups that are most similar (least dissimilar).
- Process continues until all objects in one group



distance between #7 and #10 is smaller than distance between #2 and # 14

- Note that “distance/dissimilarity” between clusters must be defined (more complicated than distance between objects)
  - 3 Linkage Methods:

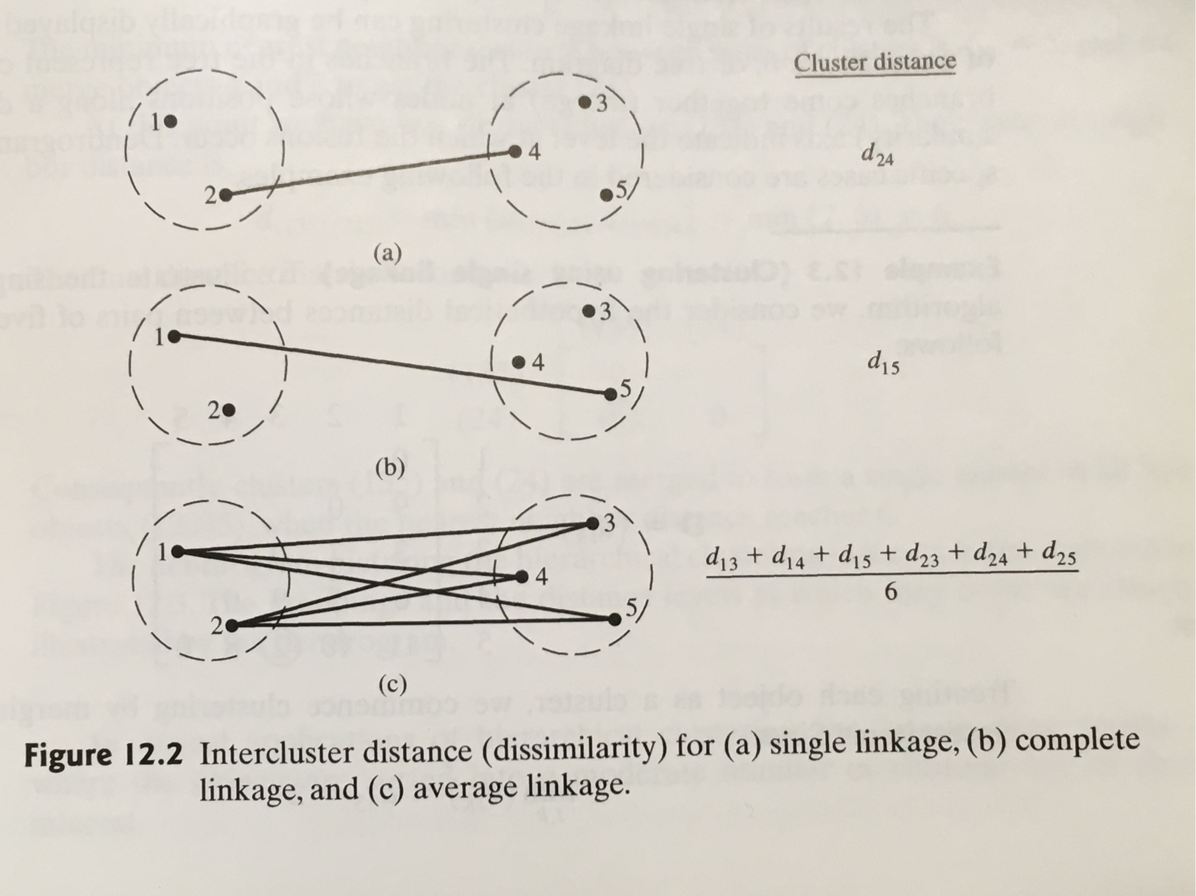
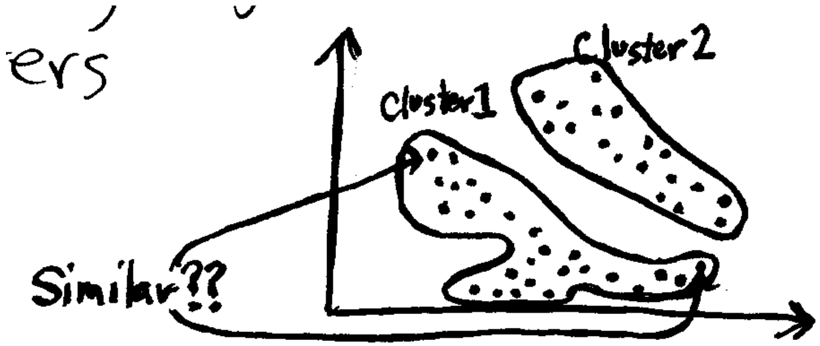

Cluster distance
1
2
3
4
5
(a)
d_{24}
1
2
3
4
5
(b)
d_{15}
1
2
3
4
5
(c)
$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$

Figure 12.2 Intercluster distance (dissimilarity) for (a) single linkage, (b) complete linkage, and (c) average linkage.

Source: JW

- Single linkage cannot discern poorly separated clusters
- Single linkage can pick out long stringlike cluster (linking objects into such non-elliptical clusters is known as "chaining"—generally considered a disadvantage)



Procedure:

1. Start with n clusters (the n observations) and an $n \times n$ symmetric matrix of distances denoted $\mathbf{D} = (d_{ik})$
2. Search for most similar pair of clusters U and V (d_{UV} is smallest)
3. Merge clusters U and V to create cluster "UV."
   - Delete row and column of $\mathbf{D}$ corresponding to clusters U and V.
   - Add a row and column giving distances between cluster UV and remaining clusters.
4. Repeat Steps 2 and 3 a total of $n - 1$ times (until all objects in a single cluster). Record levels (distances) at which mergers take place for dendrogram.

k-means Method (Non-hierarchical)

Object: Find an optimal clustering for a given number of clusters.

- Need to start with either (i) an initial partitioning of the objects or (ii) a set of k initial cluster means ("centroids") or "seed points." Using one of the linkage (hierarchical) methods is a good way to obtain an initial partition.
- Procedure:
  1. Partition the objects into k initial clusters.
  2. Assign each object to closest cluster centroid (using Euclidean distance).
     – Recalculate the centroid for the cluster receiving new object and the cluster losing the object
  3. Repeat Step 2 until no more reassignments occur.
- Hartigan (1975) gives rule-of-thumb for choosing k based on decrease in within-group sum of squares when increasing k by 1.

$\boxed{\text{ex}}$ Texas Cities

- Dendrograms for single, complete, and average linkage
- Use 4 clusters
- k-means (using complete linkage to obtain original partitioning)
- Look at groupings on PC1 vs. PC2 plot
- Do groupings make sense geographically using map?

$\boxed{\text{ex}}$ U.S. Cities (with population between 1 and 5 million)

- Same as above . . . using 5 clusters

⋆ *Often useful to follow-up cluster analysis with a discriminant analysis to describe the nature of the clusters.*