Classifiers/Icons

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 12, 2016

What did we think about last time?

- Partitioning our Space to Separate Classes
- Classification Trees
  - Can prune trees to reduce complexity
  - Pay attention to what happens to your smaller classes
  - Can stabilize with ensembles like random forests
- Can use tree structure for all sorts of decision rules; need idea of split criteria
- General Discriminant Analysis
- Choosing Decision Boundaries based on Posterior Probabilities

Now we'll try

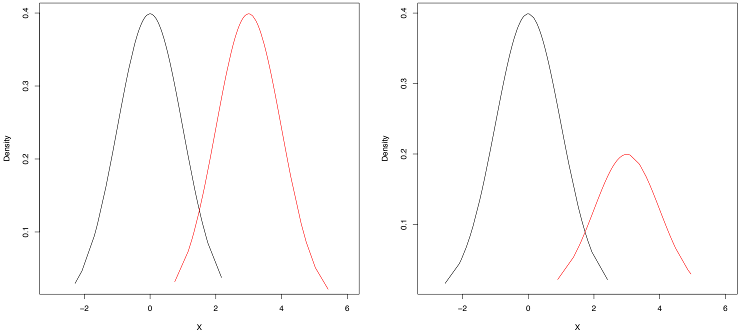
- modeling these posterior probabilities with linear/quadratic discriminant analysis
- Start looking for structure without labels

Basing Decision Boundary on Posterior Probability

$$P(Class\ j|x) = \frac{\pi_j \cdot p_j(x)}{\sum_{l=1}^{L} \pi_l p_l(x)}$$

Dependent on group size (π_j) and shape of density ($p_j(x)$)

Can find post. prob for any class at any location in feature space



Choose most likely class with post probs: $\text{argmax}_k\ P(Class\ k|x)$

Linear Discriminant Analysis

Assume densities are Gaussian and the covariances are equal
What happens if we compare the posterior probs of two classes?

$$p_j(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} e^{-1/2(x-\mu_j)^t\Sigma_j^{-1}(x-\mu_j)}$$

$$\frac{P(Class\ j|x)}{P(Class\ k|x)} = \frac{\frac{\pi_j p_j(x)}{\sum \pi_l p_l(x)}}{\frac{\pi_k p_k(x)}{\sum_l \pi_l p_l(x)}} = \frac{\pi_j p_j(x)}{\pi_k p_k(x)}$$

Taking log:

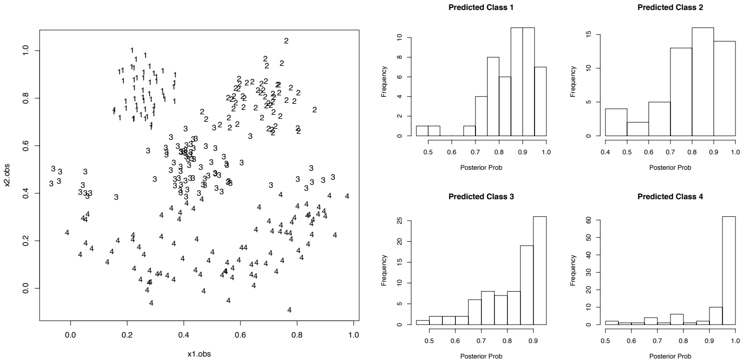
$$\log(\frac{\pi_j}{\pi_k}) + \log\frac{(2\pi)^{d/2}|\Sigma_j|^{1/2}}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} - \frac{1}{2}(x-\mu_j)^t\Sigma_j^{-1}(x-\mu_j) + \frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)$$

Covariances equal:

$$\log(\frac{\pi_j}{\pi_k}) - \frac{1}{2}(\mu_j - \mu_k)^t\Sigma^{-1}(\mu_j - \mu_k) + x^t\Sigma^{-1}(\mu_j - \mu_k)$$

Linear in x; what do we need to estimate?

Why So Serious?



Quadratic Discriminant Analysis

Back to comparing posterior probabilities of two groups:

$$\log(\frac{\pi_j}{\pi_k}) + \log\frac{(2\pi)^{d/2}|\Sigma_j|^{1/2}}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} - \frac{1}{2}(x-\mu_j)^t\Sigma_j^{-1}(x-\mu_j) + \frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)$$

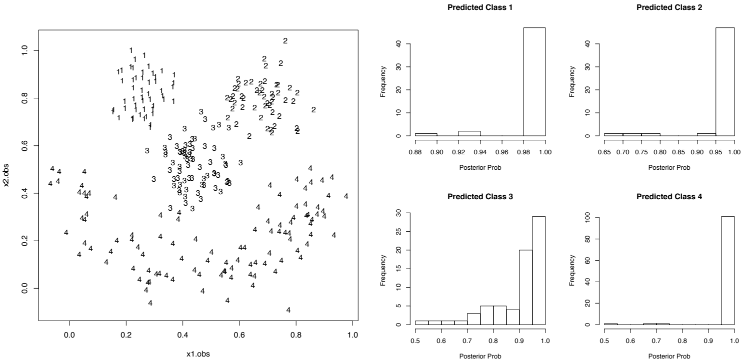
Allow covariances to be unequal; boundary stays quadratic in x.

- Far more flexible boundaries
- Comes at the cost of far more parameter estimation
- Can have fitting problems in high dimensions

Can also use discriminant analysis for dimension reduction

- LDA/QDA essentially project separation information given the classes into "discrimination" space. Dimensions are in decreasing order of "information"
- Can choose smaller number of "discrimination variables"

Seriously, Why So Serious?



Finding Structure without Labels

Often referred to as *unsupervised learning*: determining and extracting structure in data without the use of a response variable.

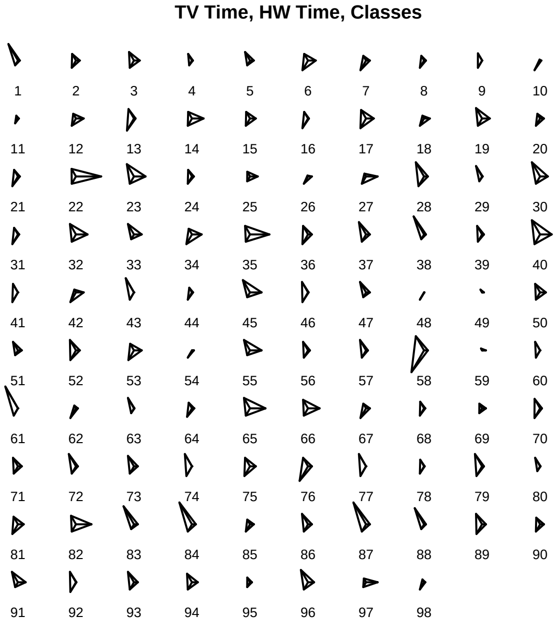
We'll start with trying to visualize groups of similar observations.

- Turns out that humans are pretty bad at seeing similarities in rows and columns of data
- Problem gets much harder in high dimensions
- Much better at seeing similarities in objects or pictures
- Often use icons or glyphs to represent high-dimensional multivariate data
- Easier to compare attributes of pictures than visually compare high-dim data
- Look for groups, patterns, outliers, etc in the pictures

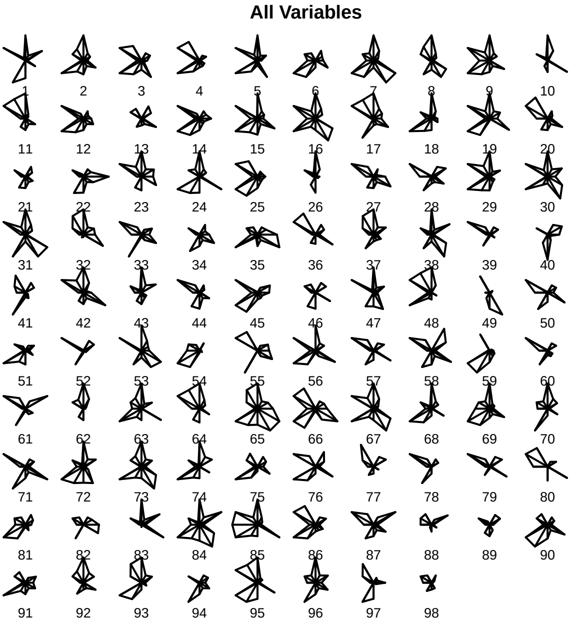
Common Glyphs

- *Stars:* each variable represented by length of vector; vectors are connected; variables counterclockwise from x-axis
- *Spider/Radar:* can put all stars on top of each other to assess similarities; each variable is a "spoke" of the spider web
- *Segment Diagrams:* each variable has a piece of a circle; length of radius corresponds to variable value
- *Thermometers:* start with two variables being the x,y coordinates; then add more variables as features of thermometer (width, height, proportion filled, etc)

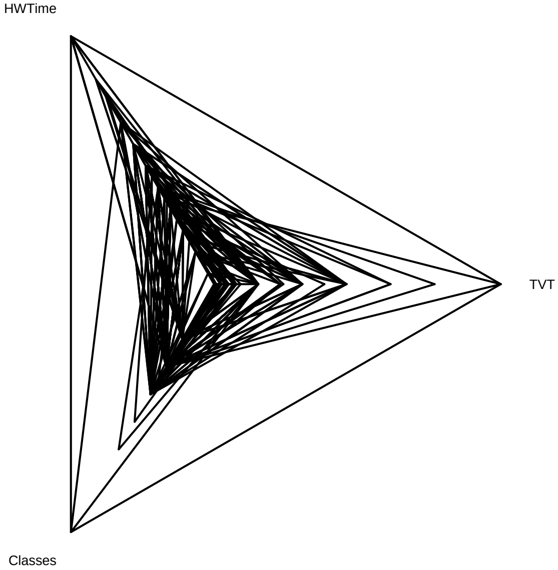
CMU Undergrads: Stars



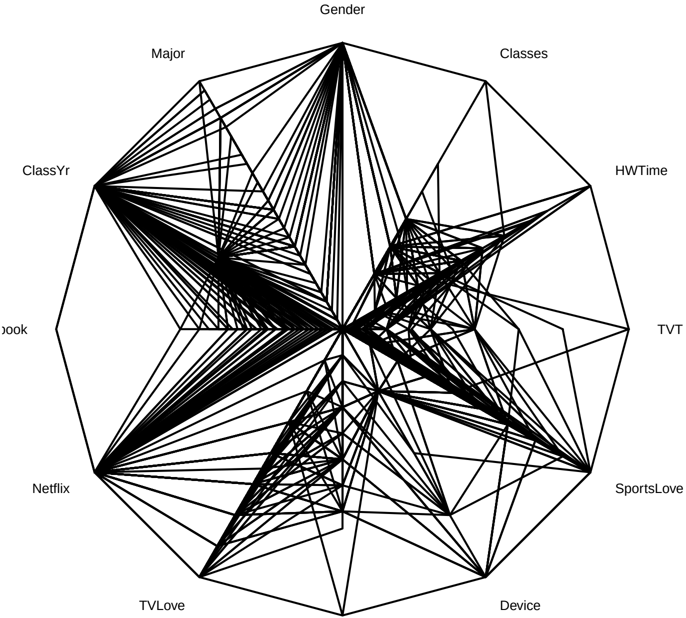
CMU Undergrads: Stars



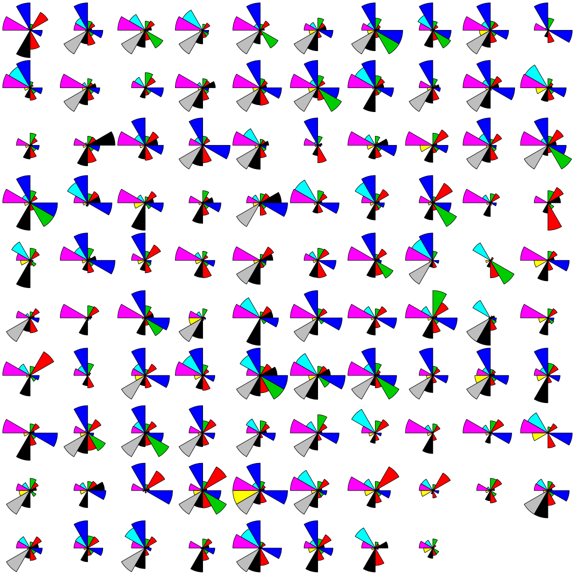
CMU Undergrads: Spiders



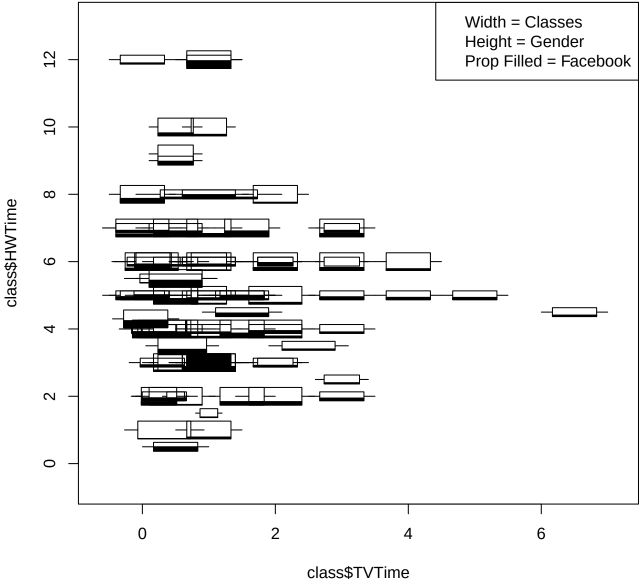
CMU Undergrads: Spiders



CMU Undergrads: Segments



CMU Undergrads: Thermometers



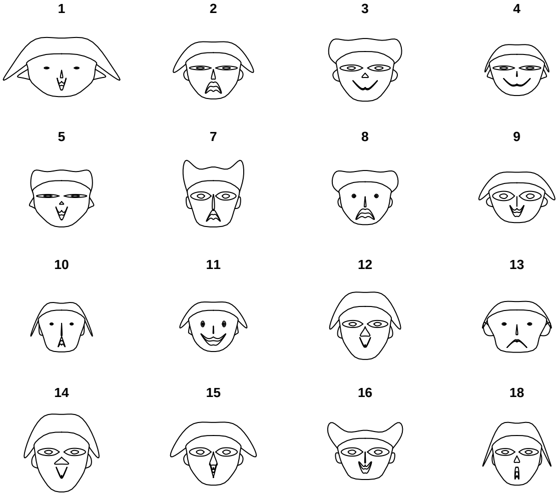
Chernoff Faces

Probably the most famous statistical icon/glyph; based on human's ability to distinguish differences in people's faces

One face per observation; represent variables by facial features

- TV Time = Height of Face
- HW Time = Width of Face
- Classes = Shape of Face
- Gender = Height of Mouth
- Major = Width of Mouth
- Class Yr = Curve of Smile
- Facebook = Height of Eyes
- Netflix = Width of Eyes
- TV Love = Height of Hair
- Avg Sleep = Width of Hair
- Device = Hairstyle
- Sports Love = Height of Nose
- could also have width of nose, width of ears, height of ears

CMU Undergrads: Chernoff Faces



CMU Undergrads: Chernoff Faces