10/25/2021

Finishing up visualization of multivariate data.

Heirarchical clustering.

Snapshot: dendrograms: a tree diagram to show groups within data.

Need a notion of distance b/w data points

ex. x_i, x_j
$d_{ij} = |x_i - x_j|$ (euclidean distance)

e.g. say $x_i = (a, b)$
$x_j = (c, d)$

$d_{ij} = \sqrt{(a-c)^2 + (b-d)^2}$

$d_{ij} = |x_i - x_j|^2$

You also need a notion of distance b/w groups

Agglomerative or bottoms up.

In the beginning, we will have $n \times n$ distance matrix where we have pairwise distances

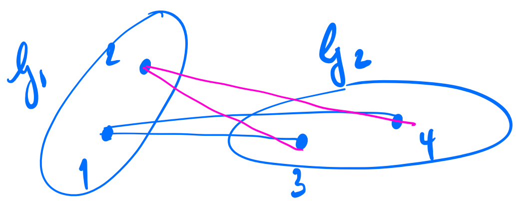
Now join groups
Suppose x_i, x_j are in a group (or cluster)
and x_k is outside.
we could update by joining x_k to the existing group with x_i, x_j
or make more groups.

Generally, the default notion of distance b/w groups is called Complete Linkage

$$\underbrace{d(g_1, g_2)}_{\text{2 groups}} = \text{maximum distance b/w points in } g_1, g_2$$

$$= \max_{\substack{i \in g_1 \\ j \in g_2}} d_{ij}$$

d_{ij} = dist b/w points i & j



d_{14}: max

$d(g_1, g_2) = d_{14}$

Say $g_1 = \{x_k\}$

Average Linkage: to measure distance b/w g_1, g_2.
Look at all pairwise distances b/w points in g_1 with pts in g_2 & take average.

$$d_{avg}(g_1, g_2) = \underset{\substack{i \in g_1 \\ j \in g_2}}{\text{mean}}(d_{ij})$$

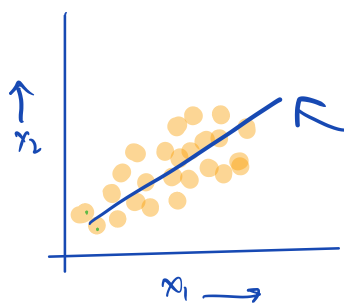
<u>Single Linkage</u>

$$d_{SL}(g_1, g_2) = \min_{\substack{i \in g_1 \\ j \in g_2}} d_{ij}.$$

Using complete linkage (max) does not allow very large clusters

PRINCIPAL COMPONENT ANALYSIS (PCA)

- For dimension reduction
- PCA refers to the process by which we compute "principal components" and how we interpret them
- We often have redundancy in variables (different variables give us similar information) (Even 10 variables give 45 pairwise plots)
- PCs are NEW variables that are linear combinations of the old ones. (not subsets of old variables)
- Goal : dimension reduction : look for underlying structure in data set to simplify original data set.



most of variation is in this direction
both x_1, x_2 give us similar info.

Try to extract one or more dimensions which have most of the variation, creating new variables which are linear combinations of the original variables.

2 equivalent methods to think about this:

① Capture direction of max. variability
② Look for line that is closest to the points.

(dist. from the line to the points is measured along orthogonal projections)

[google data camp tutorial pca mtcars]

Our original variables are $X_1, X_2, \dots X_p$.

$\vec{x}_1, \vec{x}_2 \dots x_n \quad \vec{x}_1 : (x_1^{(1)}, x_1^{(2)}, \dots x_1^{(p)}) \in \mathbb{R}^p$

$$\begin{array}{c|c}
 & \begin{matrix} X_1 & X_2 & \dots & X_p \end{matrix} \\
\begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} & \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & \dots & & x_2^{(p)} \\ & & & \\ & & & \end{bmatrix}
\end{array}$$

No response variable, want a new variable z_i that incorporates info from existing x_i

$z_i = a_1 x_i^{(1)} + a_2 x_i^{(2)} + \dots + a_p x_i^{(p)}$

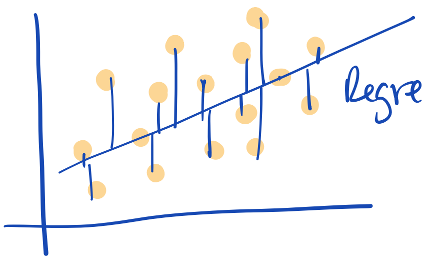
Need to find a_i, or the vector of coefficients that is "best".

Maybe find a_i that maximizes sample variance of z_i.

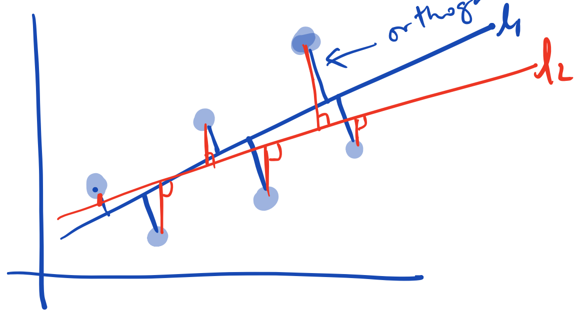
Idea 1
Find direction of max variance

$$\begin{aligned} Var(z_i) &= \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (a_1 x_i^{(1)} + a_2 x_i^{(2)} + \dots + a_p x_i^{(p)} - \bar{z})^2 \end{aligned}$$

Idea 2: Find the "best" line that fits the data cloud. (line "ℓ" that is closest to all the points)



Now look for line that minimizes distance from all the points, where distance is orthogonal projection from point to line



Idea 1. looking for direction of max. variability

Look for a new coordinate system with fewer variables

Look for a_i as described.

Compute a_i by maximizing sample variance of z_i.

Usually constrain vector $\vec{a} = (a_1, \dots, a_p)$ to have magnitude 1. $\sum_{j=1}^{p} a_j^2 = 1$

$$\text{Max} \quad \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 \quad \text{subject to} \quad \sum_{j=1}^{p} a_j^2 = 1$$

First k-principal components give us a coordinate system ("span a subspace") that gives us a k-dimensional view of the data.

Scree plot or elbow plot