

Lecture 5: Clustering, Linear Regression

Reading: Chapter 10, Sections 3.1-2

STATS 202: Data mining and analysis

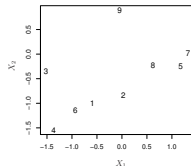
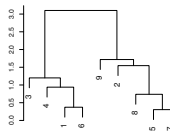
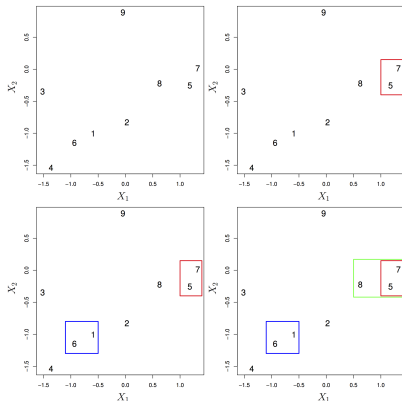
Sergio Bacallado
September 19, 2018

Announcements

- ▶ Starting next week, Julia Fukuyama will be having her office hours after business hours for SCPD students (through Skype). The times will be updated on the website as usual.
- ▶ Probability problems.
- ▶ Homework 2 will go out this afternoon. It's a long one.

Hierarchical clustering

Most algorithms for hierarchical clustering are *agglomerative*.

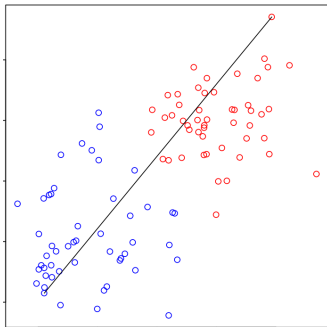


The output of the algorithm is a *dendrogram*. We must be careful about how we interpret the dendrogram.

Notion of distance between clusters

At each step, we link the 2 clusters that are “closest” to each other.

Hierarchical clustering algorithms are classified according to the notion of distance between clusters.



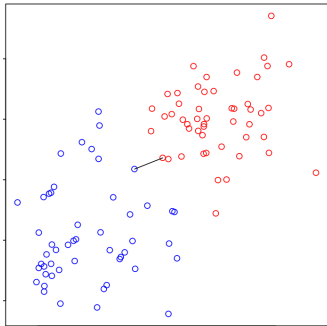
Complete linkage:

The distance between 2 clusters is the *maximum* distance between any pair of samples, one in each cluster.

Notion of distance between clusters

At each step, we link the 2 clusters that are “closest” to each other.

Hierarchical clustering algorithms are classified according to the notion of distance between clusters.



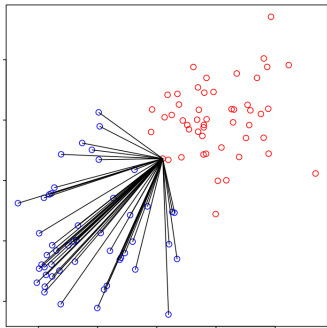
Single linkage:

The distance between 2 clusters is the *minimum* distance between any pair of samples, one in each cluster.

Notion of distance between clusters

At each step, we link the 2 clusters that are “closest” to each other.

Hierarchical clustering algorithms are classified according to the notion of distance between clusters.



Average linkage:

The distance between 2 clusters is the average of all pairwise distances.

Example

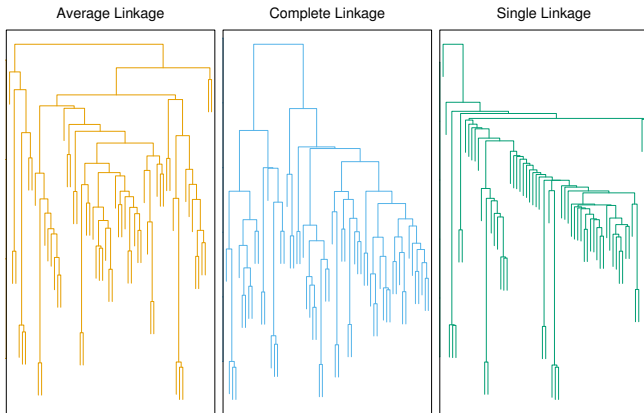


Figure 10.12

Clustering is riddled with questions and choices

- ▶ Is clustering appropriate? i.e. Could a sample belong to more than one cluster?
 - ▶ Mixture models, soft clustering, topic models.
- ▶ How many clusters are appropriate?
 - ▶ Choose subjectively — depends on the inference sought.
 - ▶ There are formal methods based on gap statistics, mixture models, etc.
- ▶ Are the clusters robust?
 - ▶ Run the clustering on different random subsets of the data. Is the structure preserved?
 - ▶ Try different clustering algorithms. Are the conclusions consistent?
 - ▶ Most important: temper your conclusions.

Clustering is riddled with questions and choices

- ▶ Should we scale the variables before doing the clustering.
 - ▶ Variables with larger variance have a larger effect on the Euclidean distance between two samples.

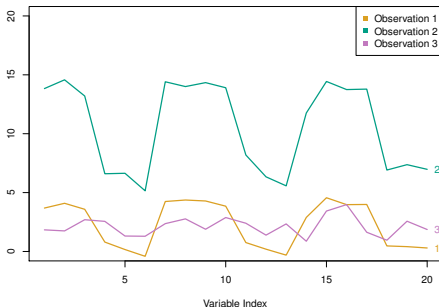
	(Area in acres,	Price in US\$,	Number of houses)
Property 1	(10,	450,000,	4)
Property 2	(5,	300,000,	1)

- ▶ Does Euclidean distance capture dissimilarity between samples?

Correlation distance

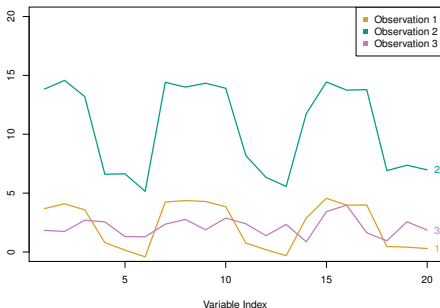
Example: Suppose that we want to cluster customers at a store for market segmentation.

- ▶ Samples are customers
- ▶ Each variable corresponds to a specific product and measures the number of items bought by the customer during a year.



Correlation distance

- ▶ Euclidean distance would cluster all customers who purchase few things (orange and purple).
- ▶ Perhaps we want to cluster customers who purchase *similar* things (orange and teal).
- ▶ Then, the **correlation distance** may be a more appropriate measure of dissimilarity between samples.



Simple linear regression

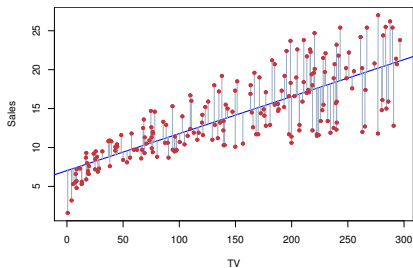


Figure 3.1

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize the residual sum of squares (RSS):

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \end{aligned}$$

Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

A little calculus shows that the minimizers of the RSS are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Assesing the accuracy of $\hat{\beta}_0$ and $\hat{\beta}_1$

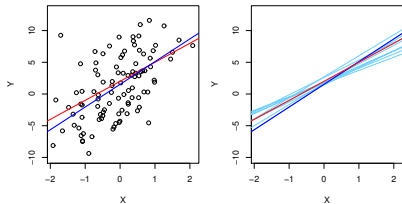


Figure 3.3

The Standard Errors for the parameters are:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The 95% confidence intervals:

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

Hypothesis test

H_0 : There is no relationship between X and Y .

H_a : There is some relationship between X and Y . H_0 :
 $\beta_1 = 0$.

H_a : $\beta_1 \neq 0$.

Test statistic: $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$.

Under the null hypothesis, this has a t -distribution
with $n - 2$ degrees of freedom.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. For the **Advertising** data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of \$1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the **sales** variable is in thousands of units, and the **TV** variable is in thousands of dollars).

Interpreting the hypothesis test

- ▶ If we reject the null hypothesis, can we assume there is a linear relationship?
 - ▶ No. A quadratic relationship may be a better fit, for example.
- ▶ If we don't reject the null hypothesis, can we assume there is no relationship between X and Y ?
 - ▶ No. This test is only powerful against certain monotone alternatives. There could be more complex non-linear relationships.

Multiple linear regression

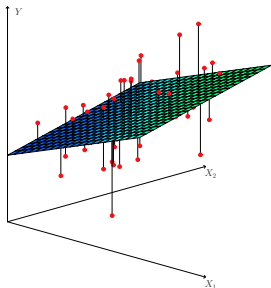


Figure 3.4

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

or, in matrix notation:

$$\mathbf{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$,
 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ and \mathbf{X} is our
usual data matrix with an extra
column of zeroes on the left to
account for the intercept.

Multiple linear regression answers several questions

- ▶ Is at least one of the variables X_i useful for predicting the outcome Y ?
- ▶ Which subset of the predictors is most important?
- ▶ How good is a linear model for these data?
- ▶ Given a set of predictor values, what is a likely value for Y , and how accurate is this prediction?

The estimates $\hat{\beta}$

Our goal again is to minimize the RSS:

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2.\end{aligned}$$

One can show that this is minimized by the vector $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Which variables are important?

Consider the hypothesis:

H_0 : The last q predictors have no relation with Y .

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F -statistic is defined by:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}.$$

Under the null hypothesis, this has an F -distribution.

Example: If $q = p$, we test whether any of the variables is important.

$$RSS_0 = \sum_{i=1}^n (y_i - \bar{y})^2$$

Which variables are important?

Consider the hypothesis:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F -statistic is defined by:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}.$$

Under the null hypothesis, this has an F -distribution.

Example: If $q = p$, we test whether any of the variables is important.

$$RSS_0 = \sum_{i=1}^n (y_i - \bar{y})^2$$

Which variables are important?

A multiple linear regression in R has the following output:

```
Residuals:
    Min       1Q   Median       3Q      Max
-15.594  -2.730  -0.518   1.777   26.199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim         -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox          -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116 < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis          -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad           3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax          -1.233e-02  3.761e-03  -3.280 0.001112 **
ptratio      -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black         9.312e-03  2.686e-03   3.467 0.000573 ***
lstat        -5.248e-01  5.072e-02  -10.347 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-Squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Which variables are important?

The t -statistic associated to the i th predictor is the square root of the F -statistic for the null hypothesis which sets only $\beta_i = 0$.

A low p -value indicates that the predictor is important.

Warning: If there are many predictors, even under the null hypothesis, some of the t -tests will have low p -values.

How many variables are important?

When we select a subset of the predictors, we have 2^p choices.

A way to simplify the choice is to define a range of models with an increasing number of variables, then select the best.

- ▶ **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- ▶ **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest p-value at each step.
- ▶ **Mixed selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate that variable.

Choosing one model in the range produced is a form of *tuning*.

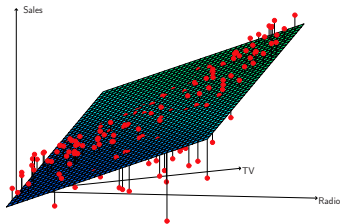
How good is the fit?

To assess the fit, we focus on the residuals.

- ▶ The RSS always decreases as we add more variables.
- ▶ The residual standard error (RSE) corrects this:

$$\text{RSE} = \sqrt{\frac{1}{n - p - 1} \text{RSS}}.$$

- ▶ Visualizing the residuals can reveal phenomena that are not accounted for by the model; eg. synergies or interactions:



How good are the predictions?

The function `predict` in R output predictions from a linear model:

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="confidence")  
      fit   lwr   upr  
1 29.80 29.01 30.60  
2 25.05 24.47 25.63  
3 20.30 19.73 20.87
```

Confidence intervals reflect the uncertainty on $\hat{\beta}$.

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="prediction")  
      fit   lwr   upr  
1 29.80 17.566 42.04  
2 25.05 12.828 37.28  
3 20.30  8.078 32.53
```

Prediction intervals reflect uncertainty on $\hat{\beta}$ and the irreducible error ε as well.