Lec9

Machine learning = learn from data , different from giving “if”

Types of machine learning

1. Unsupervised learning
2. Supervised learning
3. Reinforcement learning

**UNsupervied learning**

**Cluestering**

1. **K-means clustering** 
   1. Breaks obevations into k pre-defined mnumber of clusters
   2. You define k the number of clusters
   3. K = number of types
      1. Select k
      2. Select k=3 distant data points at random
      3. Measure the distance between the 1 st point and the k=3 initial clusters
      4. Assign the 1st point to the nearest cluster
      5. Update cluster centers – calculate the mean value for the first cluster including the new point
      6. Assign next point to closest cluster
      7. Update cluster centers and move to next pont
      8. Repear
         1. ->
      9. Pick three new initial point

**It will choose the shortest cluster combinations as winner !**

In the two dimensional -> distance measure by the pethargorithm theom

In the three dimensional -> calculate the distance by d= sqrt(x^2 +y^2 +z^2)

* Function that can do is **K-means** in R

Model selection

Recall K-means has a random component

Nstart = # -> number of running the cluster

Pick the smallest within cluster SS as winner

* **Kmeans will always give you the renumber of clusters you request**

Scree plot : trial and error is not the best approach

Systemically try a range of different K VALUES AND PLOT A SCREE PLOT

Rev = reverse

Cbind = column bind

Rnorm = normal distribution

K= 2 means “center=2”

1. **Hierarchical clustering**

Bottom-up

Top-down

Bottom-up

Set all individual as individual cluster

Narrow down to one cluster

HCLUST( DIST(X))

CUTREE(**HC, h=6**)

Linking clusters in hierarchica clustering

How is distance between clusters deterimined?

There are four main methods to determine which cluster should be linked:

Complete : uses largest of all pair-wise similarities

Single: uses smallest of all pair wise similarities

Average: uses average of all pair-wise similarities

Dimensionality reduction, visualization and ‘structure’ analysis

• Principal Component Analysis (PCA)

Different cell type (hard to visually differentiated) -> measure the mRNA for specific gene

See the correlation between two different cell type

But if there is more than 4 cell types -> hard to see coreelaton in chart

Color them in the clusters as cell types

<https://bioboot.github.io/bimm143_W20/class-material/pca/>