Supervised and unsupervised examples

Regression

k-means clustering

Introduction to Regression

- Regression analysis is the part of statistics that investigates the relationship between two or more variables related in a nondeterministic fashion.
- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.
- One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.
- For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.
- Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest.

Introduction to Linear Regression

- This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables

- A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.

- If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model.

Introduction to Linear Regression

- The simplest deterministic mathematical relationship between two variables x and y is a linear relationship.

$$y = b_0 + b_1 x$$

- The set of pairs (x, y) for which determines a straight line with slope b_1 and y-intercept b_0.

- More generally, the denoted by x and will be called the independent, predictor, or explanatory variable.

- For fixed x, the second variable will be random; we denote this random variable and its observed value by Y and y, respectively, and refer to it as the dependent or response variable.

Introduction to Linear Regression

- Usually observations will be made for a number of settings of the independent variable.
- Let $x_1, x_2, \ldots, x_n$ denote values of the independent variable for which observations are made, and let Y_i and y_i, respectively, denote the random variable and observed value associated with.
- The available bivariate data then consists of the n pairs

  $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$.
- A picture of this data called a **scatter plot** gives preliminary impressions about the nature of any relationship.
- In such a plot, each (x_i, y_i) is represented as a point plotted on a two-dimensional coordinate system.

Scatter Plot Example: Linear Regression

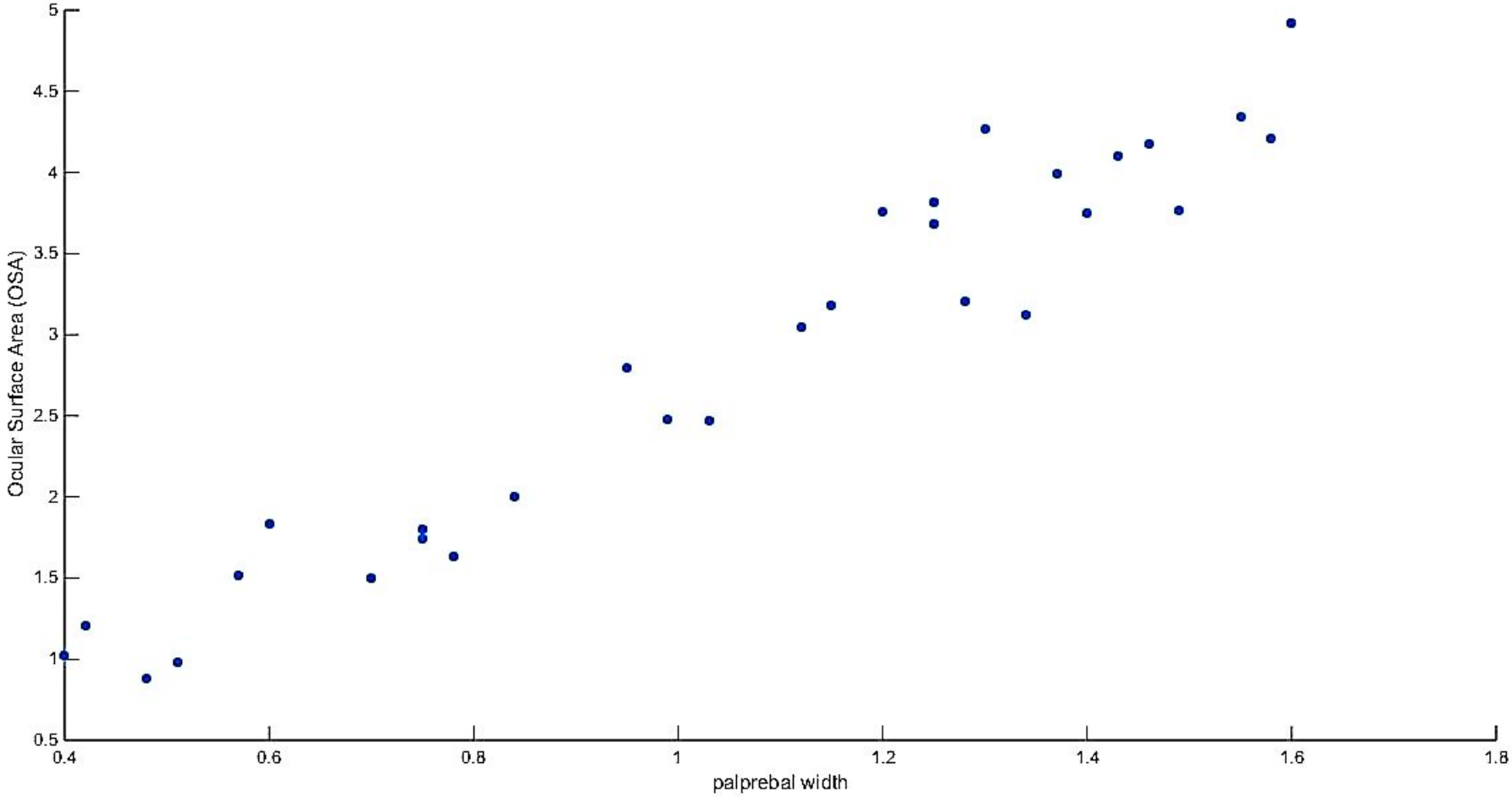
Visual and musculoskeletal problems associated with the use of visual display terminals (VDTs) have become rather common in recent years. Some researchers have focused on vertical gaze direction as a source of eye strain and irritation. This direction is known to be closely related to ocular surface area (OSA), so a method of measuring OSA is needed. The accompanying representative data on y = OSA (cm^2) and x = width of the palprebal fissure (i.e., the horizontal width of the eye opening, in cm) is from the article "Analysis of Ocular Surface Area for Comfortable VDT Workstation Layout" (*Ergonomics*, 1996: 877–884). The order in which observations were obtained was not given, so for convenience they are listed in increasing order of x values.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	.40	.42	.48	.51	.57	.60	.70	.75	.75	.78	.84	.95	.99	1.03	1.12
y_i	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

Credits: Probability and Statistics for Engineering and the Science by Jan Devore

Scatter Plot Example



Scatter Plot Example: Linear Regression

- Several observations have identical x values yet different y values.
  - *e.g., x_8=x_9 =0 .75, but y_8 =1.80 and y_9 = 1.74). Thus the value of y is not determined solely by x but also by various other factors.*
- There is a strong tendency for y to increase as x increases. That is, larger values of OSA tend to be associated with larger values of fissure width—a positive relationship between the variables.

- It appears that the value of y could be predicted from x by *finding a line that is reasonably close to the points in the plot* (the authors of the cited article superimposed such a line on their plot).

- In other words, there is evidence of a substantial (though not perfect) linear relationship between the two variables.

Simple Linear Regression Model

- For the deterministic model , the actual observed value of y is a linear function of x.
- The appropriate generalization of this to a probabilistic model assumes that the expected value of Y is a linear function of x, but that for fixed x the variable Y differs from its expected value by a random amount.
- In a simple regression problem (a single x and a single y), the form of the model would be:

$$y = b_0 + b_1 * x$$

- In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. b_0 and b_1 in the above example).

Simple Linear Regression Model

- When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0).

- The values b_0 and b_1 must be chosen so that they minimize the error. If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

Simple Linear Regression Model

- For model with one predictor,

$$b_0 = \bar{y} - b_1 \bar{x}$$

and

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Exploring ‘b1’:
  - If b1 > 0, then x(predictor) and y (target) have a positive relationship. That is increase in x will increase y.
  - If b1 < 0, then x(predictor) and y (target) have a negative relationship. That is increase in x will decrease y.

Credits: **Linear Regression — Detailed View by** Saishruthi Swaminathan

Simple Linear Regression Model

- Exploring ‘b0’
  - If the model does not include x=0, then the prediction will become meaningless with only b0.
  - For example, we have a dataset that relates height(x) and weight(y). Taking x=0(that is height as 0), will make equation have only b0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.
  - If the model includes value 0, then ‘b0’ will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.
  - The value of b0 guarantee that residual have mean zero. If there is no ‘b0’ term, then regression will be forced to pass over the origin. Both the regression co-efficient and prediction will be biased.

Credits: **Linear Regression — Detailed View by** Saishruthi Swaminathan

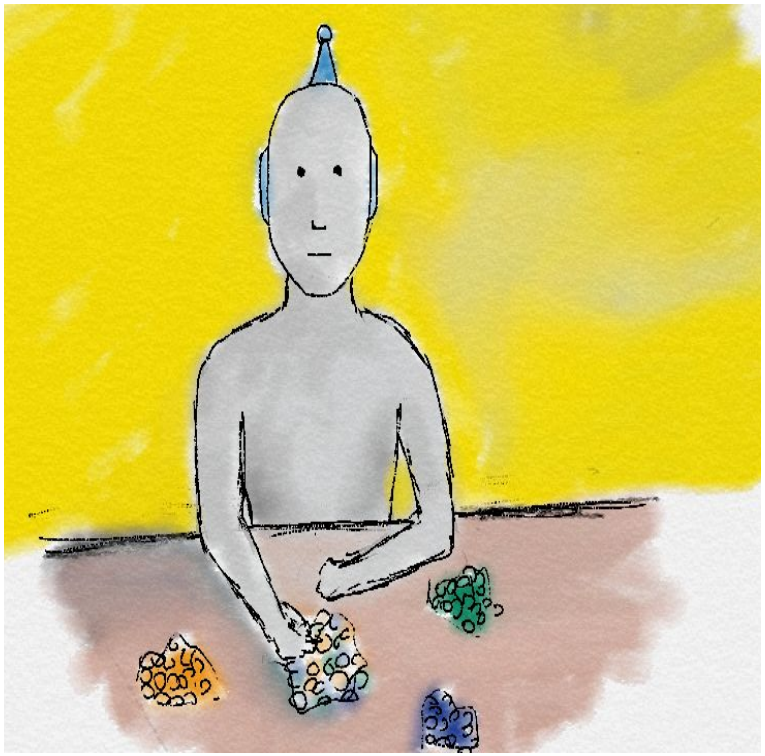
Additional Resources:

- https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86
- http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

Week	Covid Cases in thousands	(Xi-mean(x))^2	(yi-mean(y))^2	mul(col 3 and col 4)	
0	0.4	6.25	1.388469	7.638469	
1	0.56	16	6.4009	22.4009	
2	0.96	14.44	6.948496	21.3885	110.5536
3	1.61	14.0625	3.674122	17.73662	91.752
4	2.59	16	0.388711	16.38871	
5	3.35	25	0.000374	25.00037	

What is clustering?

- Clustering: the process of grouping a set of objects into classes of similar objects.
  - Objects within a cluster should be similar.
  - Objects from different clusters should be dissimilar.



k-means Clustering

- k-means is a partitional clustering algorithm
- Let the set of data points (or instances) D be $\{x_1, x_2, \ldots, x_n\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $x \subseteq R^r$, and r is the number of attributes (dimensions) in the data.
- The k-means algorithm partitions the given data into k clusters.
- Each cluster has a cluster centre, called centroid.
- k is specified by the user

k-means algorithm

1) Randomly choose *k* data points (seeds) to be the initial centroids or cluster centers.
2) Assign each data point to the closest centroid
3) Re-compute the centroids using the current cluster memberships.
4) Repeat the assignment (step 2) and update the centroids (step 3) until centroids remain same or no chances in clusters.

Algorithm: k-means clustering algorithm

1: Select *k* points as initial centroids.
2: **repeat**
3: Form *k* clusters by assigning each point (x) to its closest centroid (m_j) using sum of squared error (SSE),

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2$$

C_j is the jth cluster, m_j is the centroid of cluster C_j (the mean vector of all the data points in Cj), and dist(x, m_j) is the distance between data point x and centroid m_j.

4: Recompute the centroid (m_j) for each cluster (c_j).

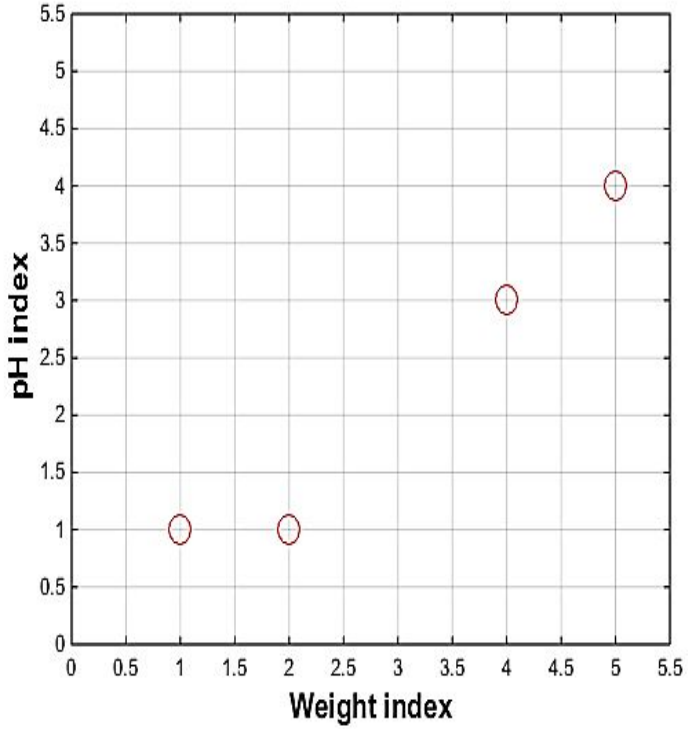
$$m_j = \frac{1}{n} \sum_{k=1, x \in C_j}^{n} x_k$$

5: **until** Centroids do not change.

Example of k means:

- Problem: Suppose we have 4 types of medicines and each has two attributes (pH and weight index). Our goal is to group these objects into k=2 group of medicine.

Medicine	Weight	pH-index
A	1	1
B	2	1
C	4	3
D	5	4



Example of k-means: Iteration 1

Step 1: Use initial seed points for partitioning.

Seed Points: C1= m_1 =(1,1) and C2= m_2 =(2,1)

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2$$

$$Euclidian\ dist(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots . + (x_n - y_n)^2}$$

A=(1,1) m_1 =(1,1) $m_2 = (2,1)$

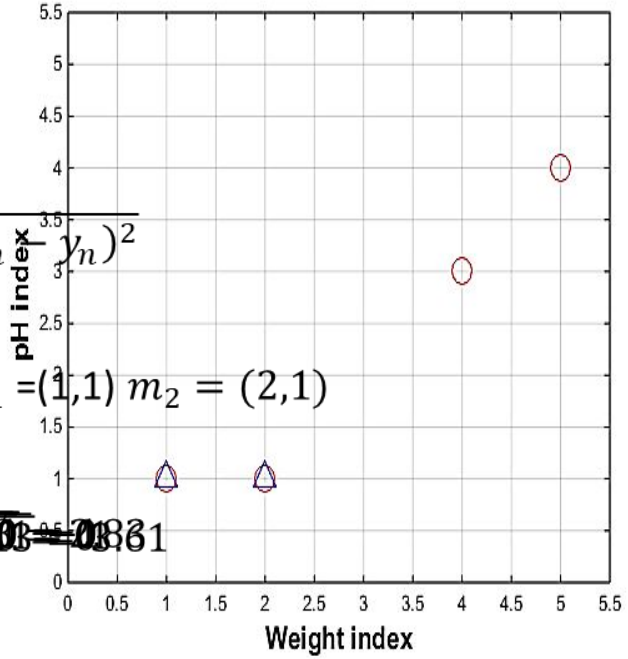
B=(2,1) m_1 =(1,1) $m_2 = (2,1)$

C=(4,3) m_1 =(1,1) $m_2 = (2,1)$

Euclidian dist((4,3),m_1) $= \sqrt{(4-1)^2 + (3-1)^2} = \sqrt{9+4} = \sqrt{13} = 3.61$

Euclidian dist((4,3),m_2) $= \sqrt{(4-2)^2 + (3-1)^2} = \sqrt{4+4} = \sqrt{8} = 2.83$

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 \\ 1 & 0 & 2.83 \end{bmatrix}$$



Example of K-means: iteration 1

Step 1: Use initial seed points for partitioning.

Seed Points:

$$c_1 = A, c_2 = B$$

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

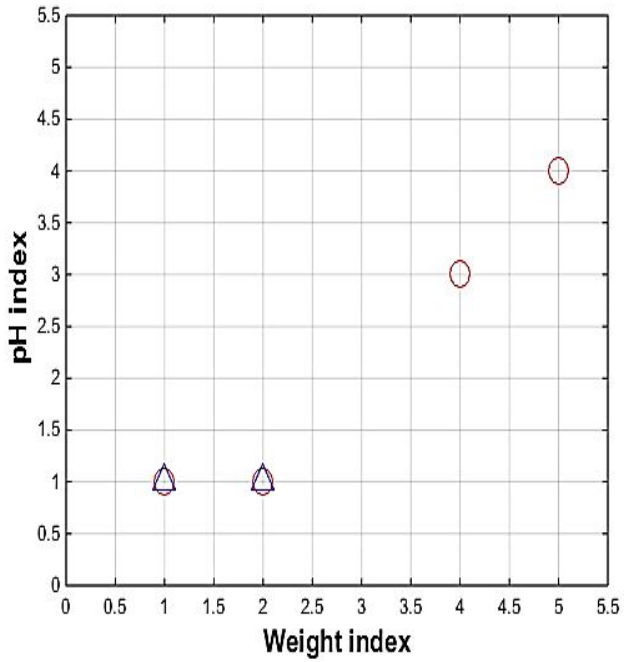
$$\begin{matrix} A & B & C & D \end{matrix} \quad \text{Euclidean distance}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Assign each object to the cluster with the nearest seed point (mean).



Example of K-means: iteration 1

Step 2: Renew membership based on new centroids..

Seed Points: $c_1 = A, c_2 = B$

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

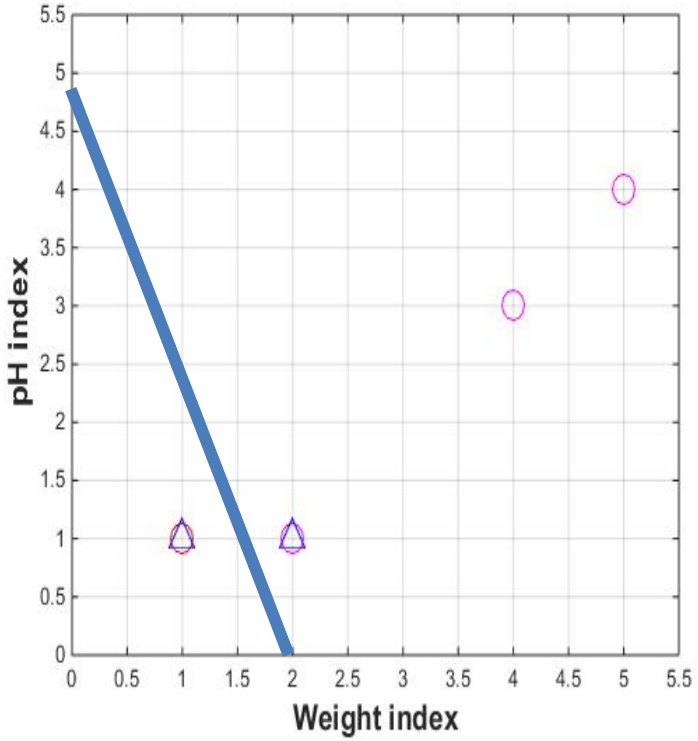
$$\begin{matrix} A & B & C & D \end{matrix} \quad \text{Euclidean distance}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Assign each object to the cluster with the nearest seed point (mean).



iteration 1

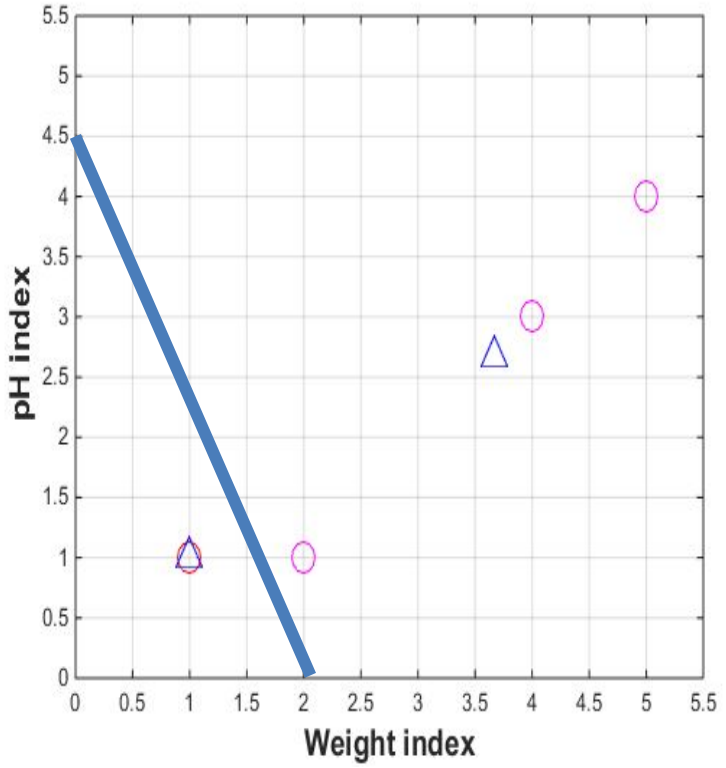
Step 3: Compute new centroids of the current partition.

Knowing the members of each cluster, compute the new centroid for each cluster based on these new membership assignment.

$$c_1 = (1,\ 1)$$

$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right)$$

$$= \left(\frac{11}{3},\ \frac{8}{3}\right) = (3.66,\ 2.66)$$



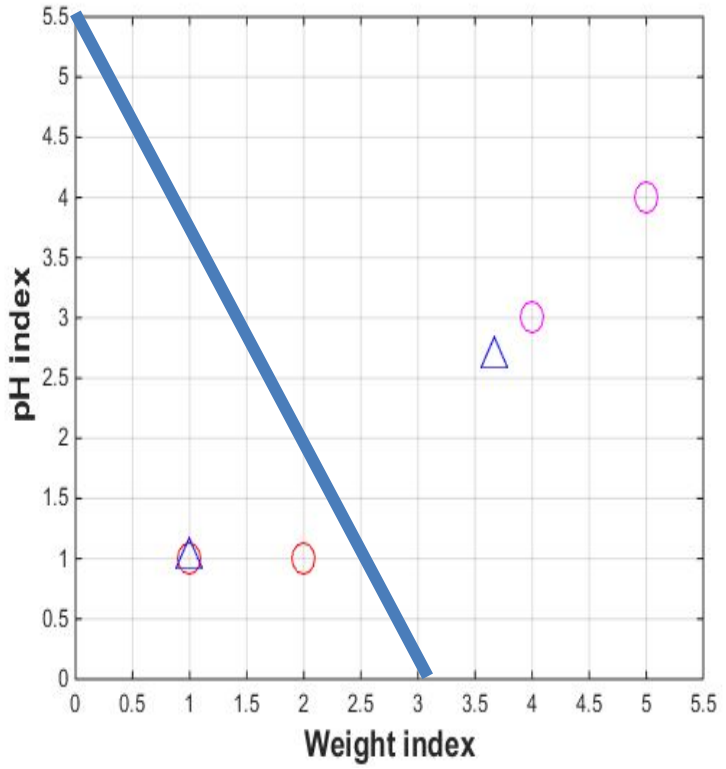
iteration 2

Step 2: Renew membership based on new centroids.

Compute the distance of all objects to the new ce
Assign each object to the cluster with the nearest

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \end{matrix} \quad \begin{matrix} X \\ Y \end{matrix}$$



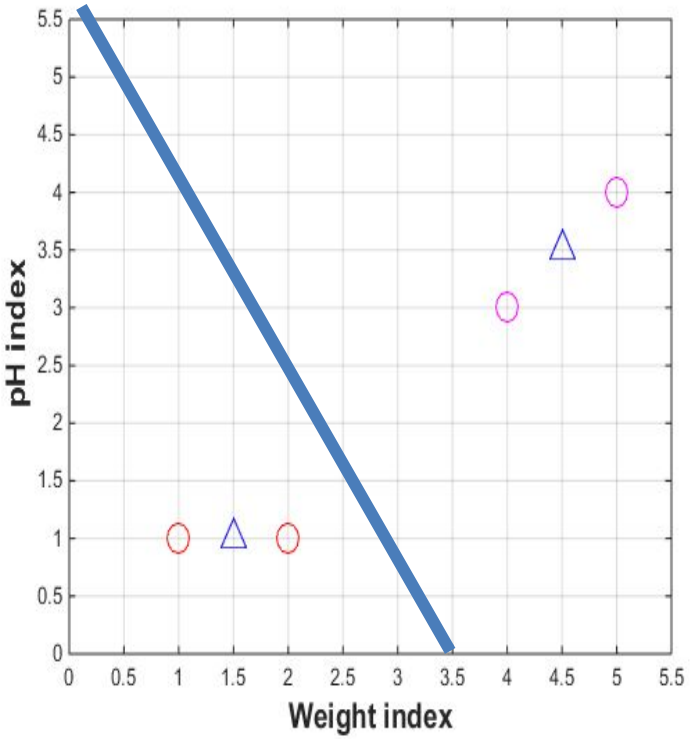
Example of k-means: iteration 2

Step 3: Compute new centroids of the current partition.

Knowing the members of each cluster, compute the new centroid for each cluster based on these new membership assignment.

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\frac{1}{2}, 1)$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4\frac{1}{2}, 3\frac{1}{2})$$



iteration 3

Step 3: Repeat the first two steps until its convergence.

Compute the distance of all objects to the new c
Assign each object to the cluster with the neares

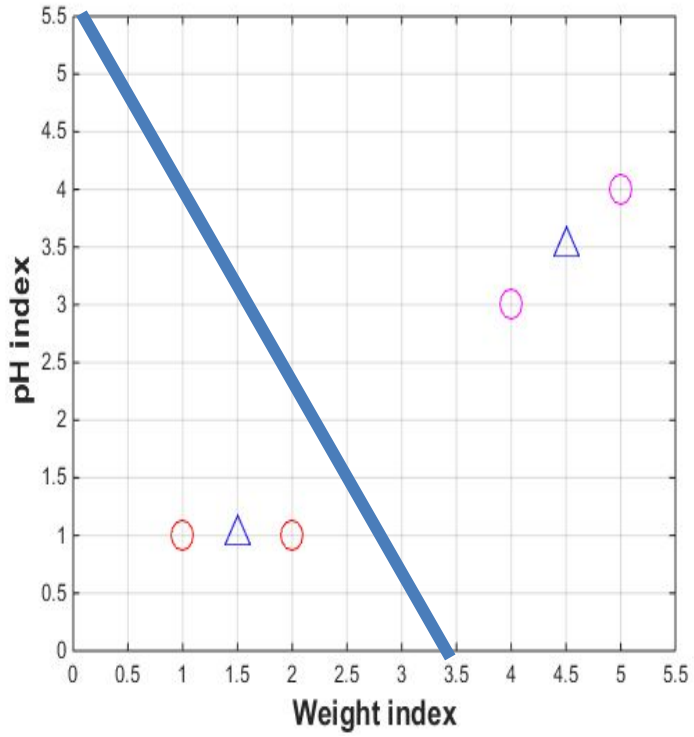
$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} c_1 = (1\frac{1}{2}, 1) & group-1 \\ c_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Knowing the members of each cluster, compute the new centroid for each cluster based on these new membership assignment.

Since there is no change in cluster assignments and centroids the algorithm terminates.



Applications of k-means clustering

- Behavioural segmentation:
- Inventory categorization:
- Sorting sensor measurements:
- Image Segmentation