Different Distance Based Classification techniques on IRIS Data set

Data Set Description

No of Classes	No of Features	No of observation of each class
Setosa Versicolour Virginica	sepal length sepal width petal length petal width	C -1: 50 C -2: 50 C-3: 50

Training Set: 60% of Each class instances
Testing Set: 40% of each class Instances

Distance Metrics

- Euclidean Distance (Squared ED, Normalized Square ED)
- City Block Distance (=Manhattan Distance)
- Chess Board Distance
- Mahalanobis Distance
- Minkowski Distance
- Chebyshev Distance
- Correlation Distance
- Cosine Distance
- Bray-Curtis Distance
- Canberra Distance

Vector Representation

$A = \mathbb{R}^N \rightarrow N\ dimensional\ Real\ Space$

$= \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \dots \dots \dots \dots \times \mathbb{R}$

$\mathbb{R} = (-\infty, \infty) \quad (set)$

$\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$

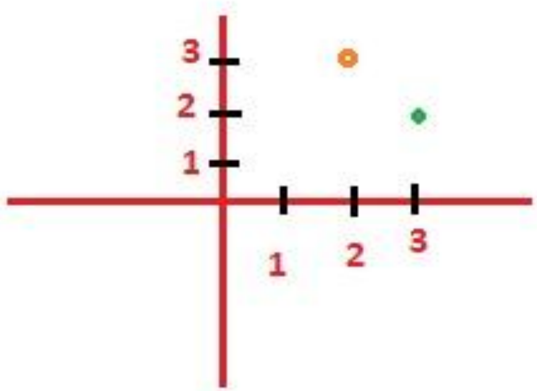
$$\underline{a}_{n\times 1} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ . \\ . \\ a_n \end{pmatrix} \quad \underline{b}_{n\times 1} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ . \\ . \\ b_n \end{pmatrix}$$

$n\ dimensional\ column\ vector$

Distance between them

$\sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$

$\sum_{i=1}^{n} |a_i - b_i|$



2D Euclidean Space

For Example: $(2,3) \neq (3,2)$

$\mathbb{R}^2 = (2,3) \times (3,2).$

Properties of Metric

Def:

Let $A \neq \emptyset$

Let $d : A \times A \rightarrow [\,0, \infty)$

$d(x, y) \rightarrow ordered\, pair$

$d(x, y) \neq d(y, x)$ [$may\, not\, be$]

$d(x, y)$ $takes\ value\ in\ the\ interval$ $[0, \infty)$.

1) $d(x, y) = d(y, x) \ \ \forall\, x, y \in A$

2) $d(x, y) = 0 \quad \leftrightarrow \quad x = y$

3) $d(x, y) + d(y, z) \geq d(x, z) \ \ \forall\, x, y, z \in A$ → **Triangular Inequality**

1). **Distance is not negative number.**

2) . **Distance can be zero or greater than zero.**

Dissimilarity Measures

Metrics

$$\underline{a}' = \left(a_1, a_{2,\ldots\ldots\ldots\ldots} a_M\right)$$

$$\underline{b}' = \left(b_1, b_{2,\ldots\ldots\ldots\ldots} b_M\right)$$

$$d_p(\underline{a}, \underline{b}) = \left(\sum_{i=1}^{M} |a_i - b_i|^p\right)^{\frac{1}{p}} ; p \geq 1$$

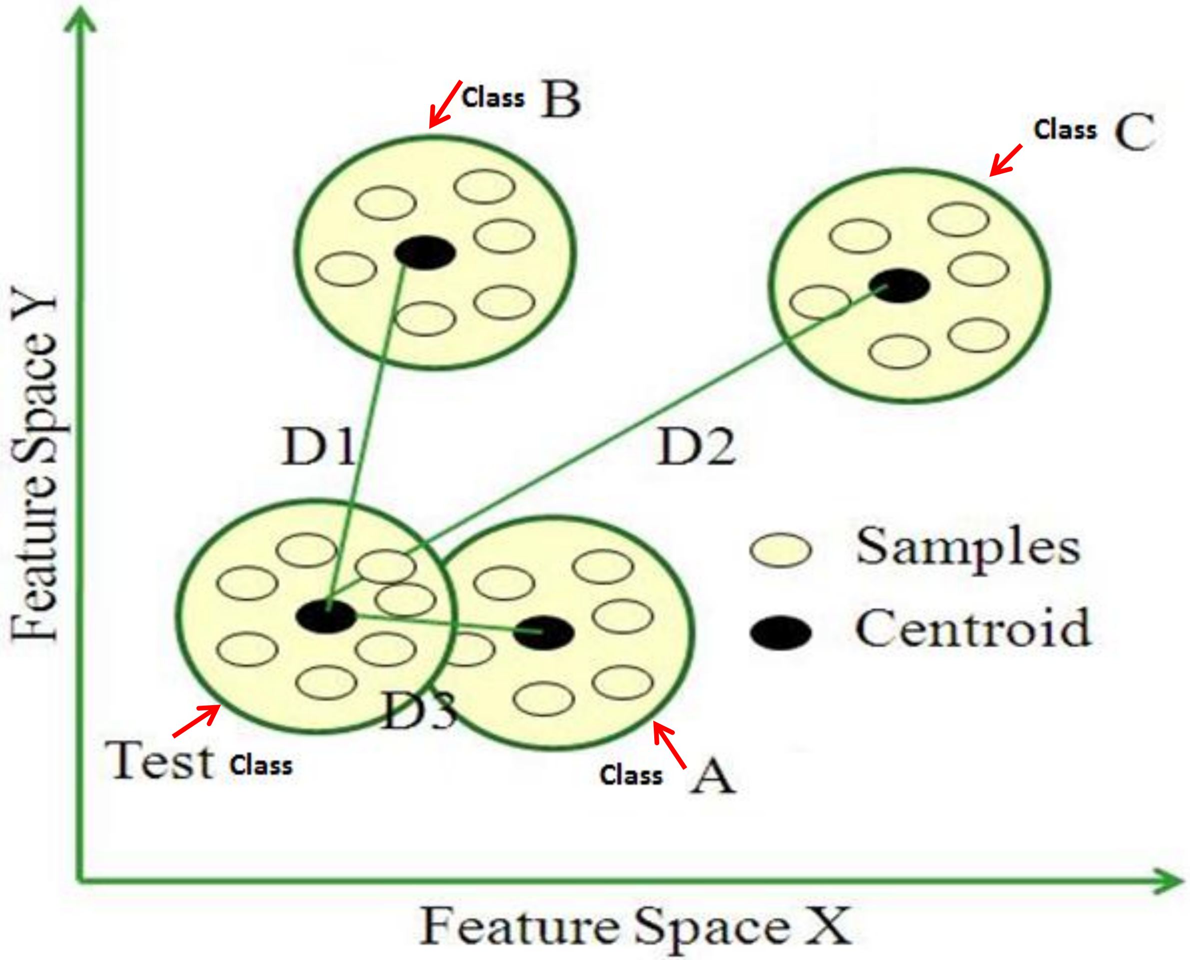
$$p = 2 \rightarrow Euclidean\ Distance$$

Classification Approaches

Generalized Distance Metric

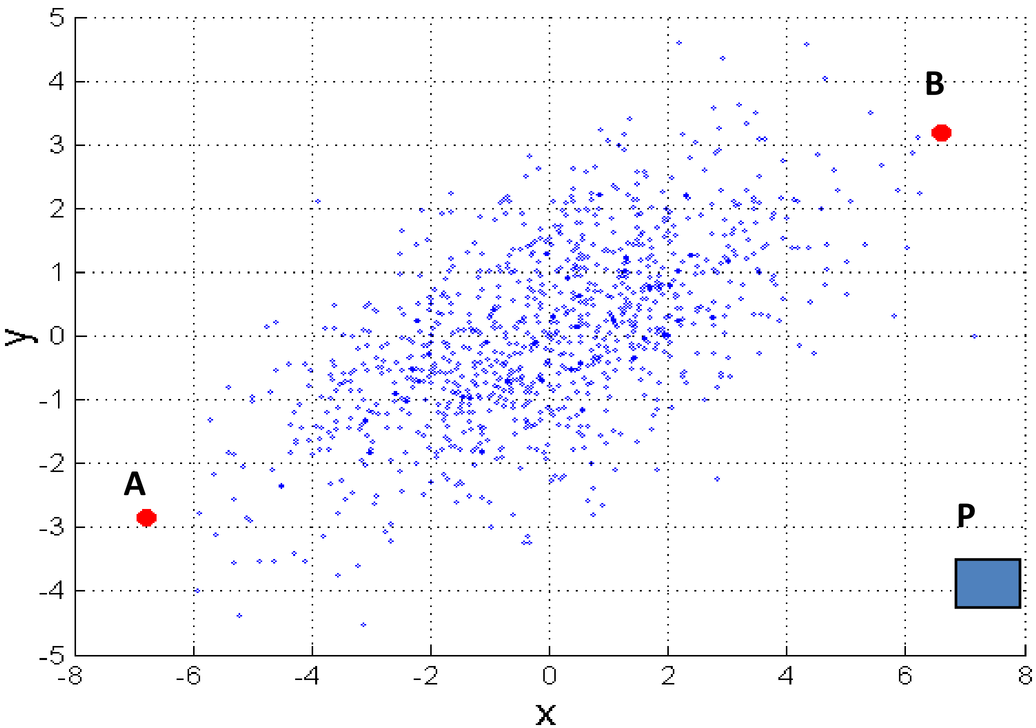
- **Step 1:** Find the average between all the points in training class C_k .
- **Step 2:** Repeat this process for all the class k
- **Step 3:** Find the **Euclidean distance/City Block/ Chess Board** between Centroid of each training classes and all the samples of the test class using $d_p(\underline{a}, \underline{b}) = \left(\sum_{i=1}^{M} |a_i - b_i|^p\right)^{\frac{1}{p}}$; $p \geq 1$
- **Step 4:** Find the class with minimum distance.

Euclidean Metric Measurement



Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$



Σ is the covariance matrix of the input data *X*

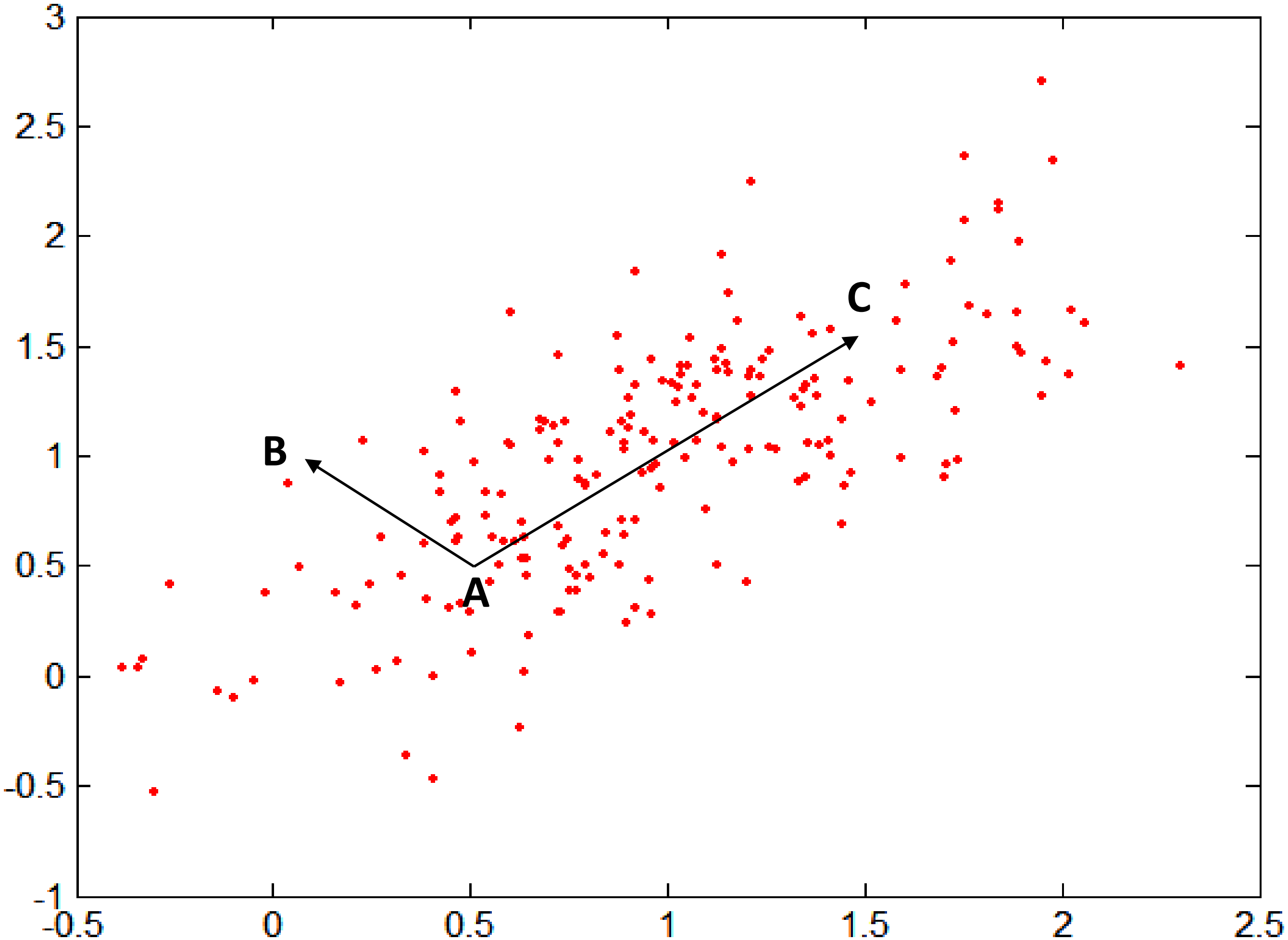
$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

When the covariance matrix is identity Matrix, the mahalanobis distance is the same as the Euclidean distance.

Useful for detecting outliers.

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

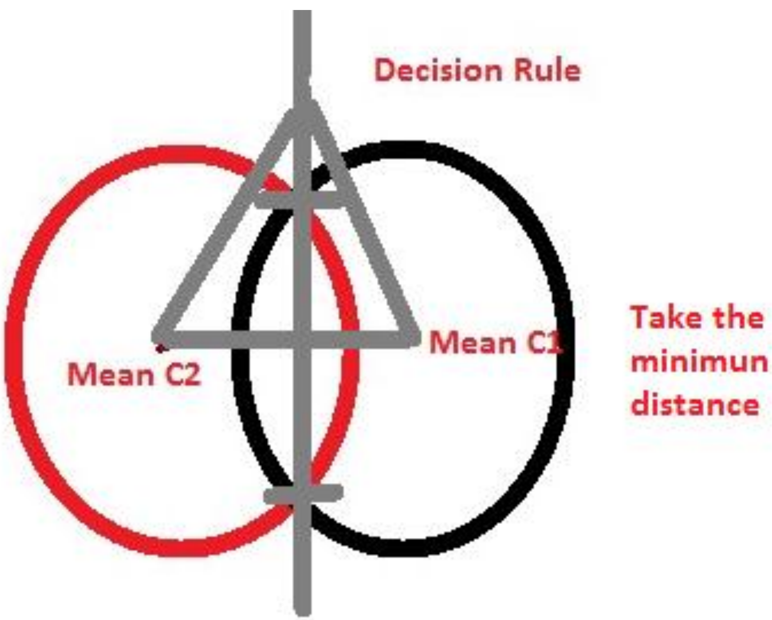
C: (1.5, 1.5)

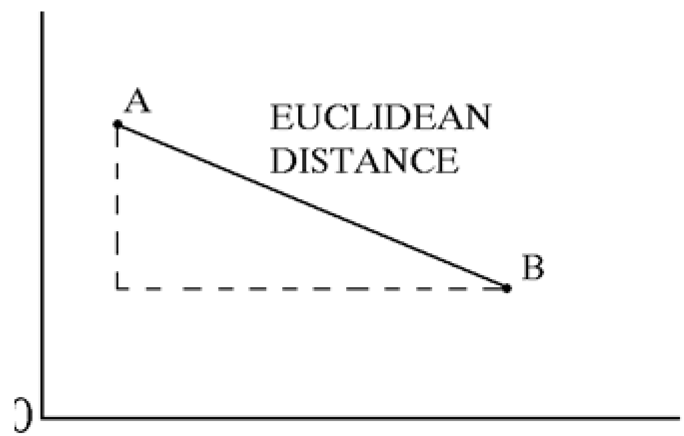
Mahal(A,B) = 5

Mahal(A,C) = 4

Geometric Representations of Euclidean Distance

$$ED_{i,h} = \sqrt{\sum_{j=1}^{p}(a_{i,j} - a_{h,j})^2}$$





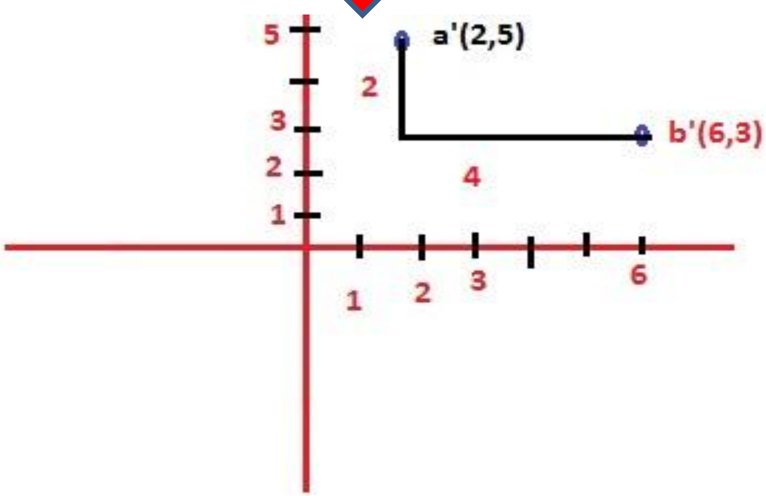
City Block Distance

$\underline{a} = \binom{2}{5}$

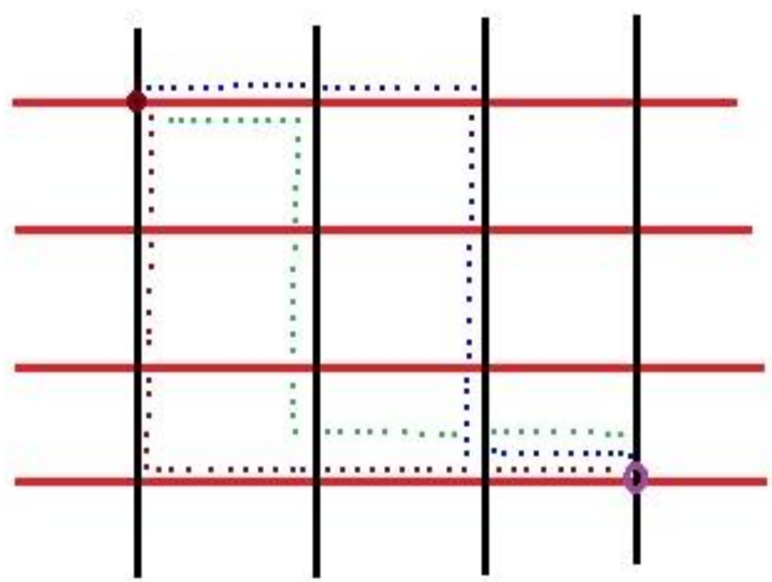
$\underline{b} = \binom{6}{3}$

(Two dimensional vector)

$d_1(\underline{a}, \underline{b}) = 4 + 2 = 6$



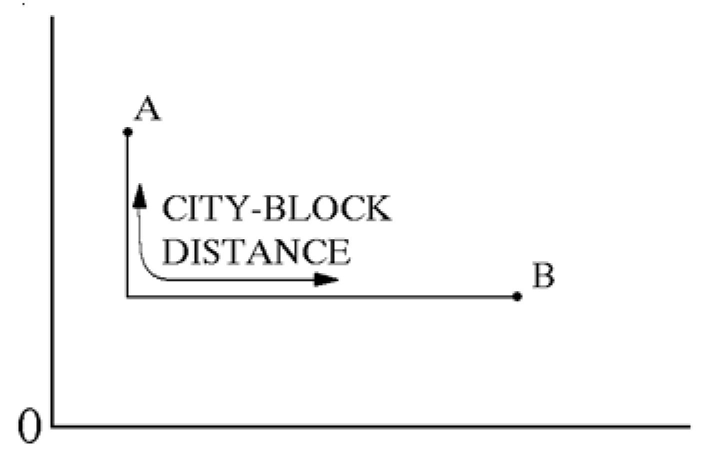
City Block Distance $d_1(\underline{a}, \underline{b}) = \sum_{i=1}^{n} |a_i - b_i|$

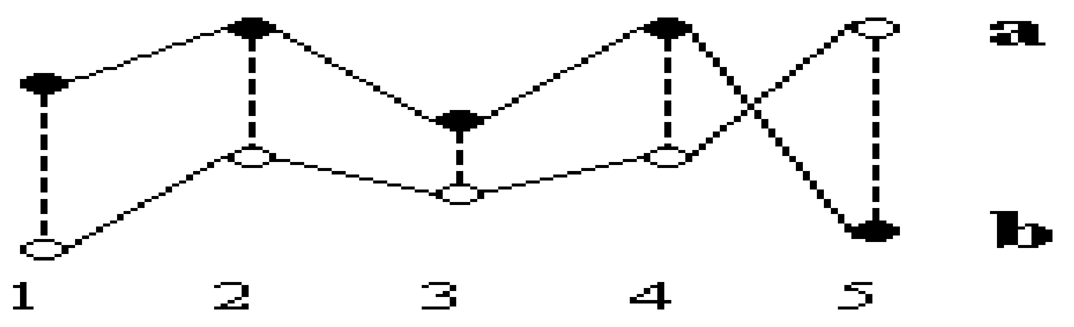


Geometric Representations of City Block Distance

City-block distance (= Manhattan distance)

$$CB_{i,h} = \sum_{j=1}^{p} |a_{i,j} - a_{h,j}|$$





The dotted lines in the figure are the distances (a_1-b_1), (a_2-b_2), (a_3-b_3), (a_4-b_4) and (a_5-b_5)

Chess Board Distance

Evaluate:

$$\lim_{p \to \infty} d_p\left(\underline{a}\,,\underline{b}\right) = ?$$

$$\underset{i=1,2,3\,\ldots.n}{\operatorname{argmax}} \;\; |a_i - b_i|$$

$\underline{a} = \binom{1}{3}$

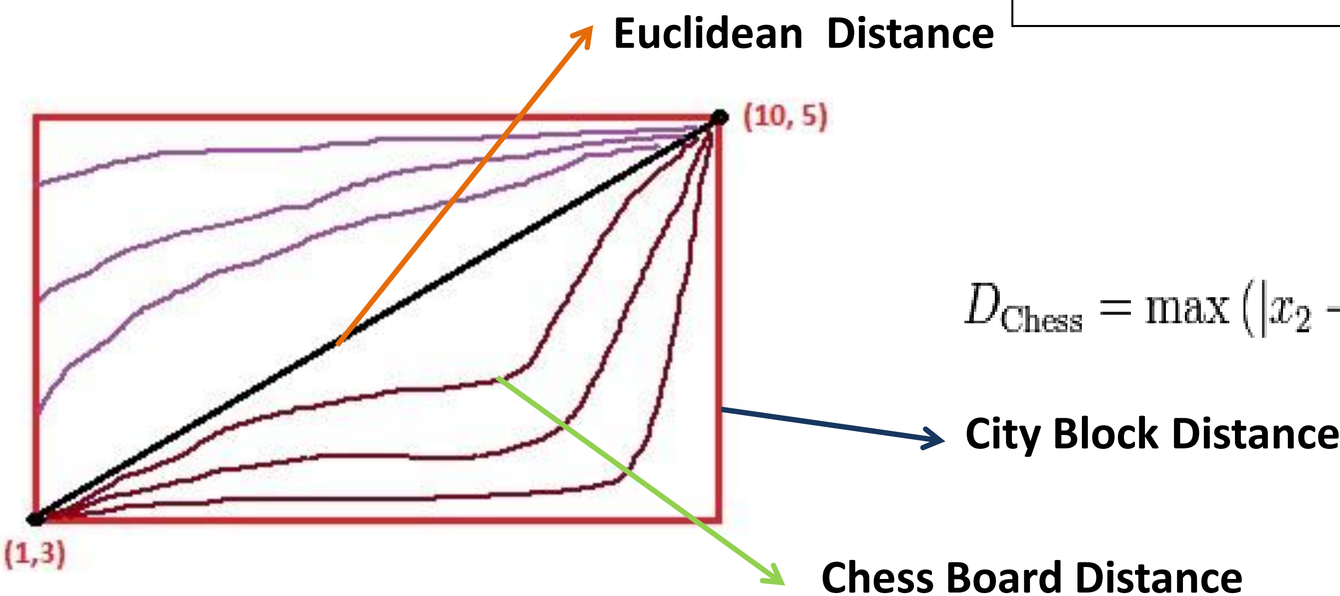
$\underline{b} = \binom{10}{5}$

$p = 1;\; d_1 \;= 11$

$p = 2;\; d_2 \;= \sqrt{81+4} = \;\sqrt{85}$

$p = \infty;\; d_p \;= \max\left(|10-1|, |5-3|\right)$

Result of Chess board Distance = 9.



$$D_{\text{Chess}} = \max\left(|x_2 - x_1|\,,|y_2 - y_1|\right).$$

Correlation Distance

- Correlation Distance [u, v]. Gives the correlation coefficient distance between vectors u and v.

Correlation Distance [{a, b, c}, {x, y, z}]; u = {a, b, c};
v = {x, y, z};

CD = 1 - (u – Mean [u]).(v – Mean [v]) / (Norm[u - Mean[u]] Norm[v - Mean[v]])

Cosine Distance

Cosine distance [u, v]; Gives the angular cosine distance between vectors u and v.

- Cosine distance between two vectors:

Cosine Distance [{a, b, c}, {x, y, z}]

CoD = 1 - {a, b, c}.{x, y, z}/(Norm[{a, b, c}] Norm[{x, y, z}])

Bray Curtis Distance

- Bray Curtis Distance [u, v];
  Gives the Bray-Curtis distance between vectors u and v.

- Bray-Curtis distance between two vectors:

Bray-Curtis Distance[{a, b, c}, {x, y, z}]

BCD: Total[Abs[{a, b, c} - {x, y, z}]]/Total[Abs[{a, b, c} + {x, y, z}]]

Canberra Distance

- Canberra Distance[u, v]
  Gives the Canberra distance between vectors u and v.
- Canberra distance between two vectors:

Canberra Distance[{a, b, c}, {x, y, z}]

CAD: Total[Abs[{a, b, c} - {x, y, z}]/(Abs[{a, b, c}] + Abs[{x, y, z}])]

Minkowski distance

- The **Minkowski distance** can be considered as a generalization of both the Euclidean distance and the Manhattan Distance.

Output to be shown

- Error Plot (Classifier Vs Misclassification error rates)

- MER = 1 – (no of samples correctly classified)/(Total no of test samples)

- Compute mean error, mean squared error (mse), mean absolute error