Lesson 7

K-NN Regression Analysis

K-Nearest Neighbours (KNN) analysis

- Consider the saying, ‘a person is known by the company he/she keeps.’
- K-NN predicts using neighbouring data points?
- An machine learning based technique using the concept of using up to k-neighbours to the data points; k = 1, 2, …,

K-NN

- K = 1 means the nearest neighbour data points
- K = 2 means the next nearest neighbour data points (xi, yi)
- K = 3 means the next to next nearest neighbour data points (xi, yi), and so on.

KNN

- An algorithm, which is usually used for classifiers
- Useful for regression also
- Predictions can use all k examples (global examples) or just K examples (K-neighbours with K = 1, 2 or 3)

KNN

- Predicts the unknown value y_p using predictor variable x_p using the available values at the neighbours
- Training dataset consists of available values of y_{ni} at x_{ni} with ni = 1 to K, where ni is the K-th neighbour, means just the local examples

K-NN Method

- First find all available neighbouring target (xi, yi) cases,
- Then predict the numerical value to be predicted based on a similarity measure

Prediction Methods

(i) Simple interpolation, when predictor variable is inside the training subset

(ii) Extrapolation, when predictor variable is outside the training subset

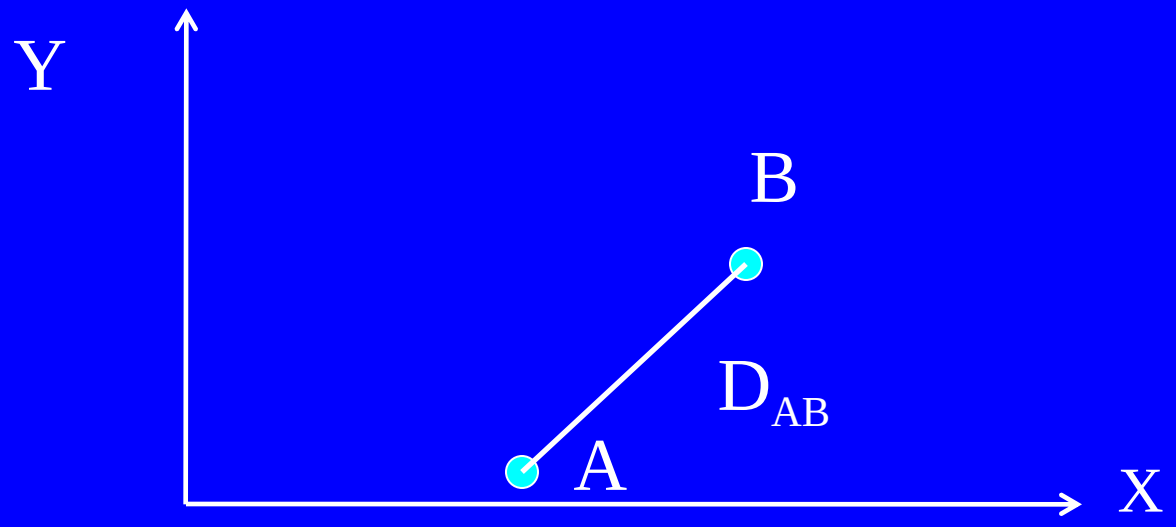
(iii) Averaging, local linear regression or local-weighted regression.

KNN Analysis

- Assumes that weight is inversely proportional to the square of distance (w is proportional to D^{-2}),when using Squared Euclidean D_{Eu}^{2} for interpolation or extrapolation for predictor variables
- Refer Equations (6.20a and 6.20b)

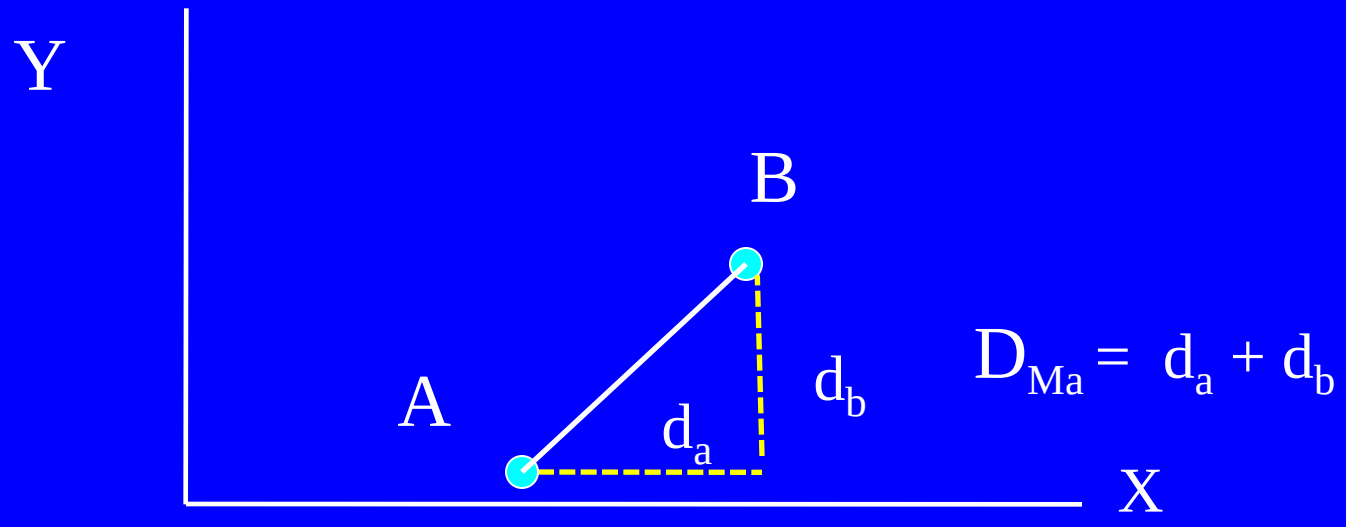
Euclidean D_{Eu}

- In terms of distance between two data-points A and B (Equations 6.20a and 20b)



Manhattan Distance D_{Ma}

- In terms of sum of axial distances between two data-points A and B (Equation 6.20c)



Euclidean and Manhattan Distances

- When v = 2, Euclidean distance is the diagonal distance between the points on an x-y graph
- Manhattan distances are faster to calculate as compared to Euclidean distances. Manhattan distances are proportional to Euclidean distances in case of linear regression

KNN Analysis

- Assumes that weight is inversely proportional to the of distance ($w \alpha D^{-1}$),when using Manhattan distance D_{Ma} for interpolation or extrapolation for predictor variables
- Refer Equation (6.20c)

KNN Analysis

- Assumes that weight is inversely proportional to qth power of the distance D^{-q}, called Minkowski D_{Mi} distance
- Refer Equation (6.20d)

Coefficients (Weights) Assignments

- When predicting, a weight assignment may require computations using a kernel function, like a Gaussian or tri-cube function in cases where the dependent variable varies according to the kernel function.

Hamming Distance

- Used when predictions are on the basis of categorical variables
- A measure of the number of instances in which corresponding values are found

Summary

We learnt:

- K-NN Regression analysis
- Interpolation and extrapolation using distance computations
- Euclidean, Manhattan, Minkowski and Hamming distances

End of Lesson 7 on K-NN Regression Analysis