MLPR Lab 5
Questions and Answers
Maan Kumawat, U20240010, DSEB

Question 1. What are some common distance based metrics used in distance-based classification algorithms?

Answer 1.

1. Euclidean Distance: Represents shortest distance between two vectors but does not account for distribution of the data.

$$d(x,\ y)\ =\sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

2. Mahalanobis Distance: Represents distance between a point P and a distribution D. We measure how many standard deviations away is P from the mean of D.

$$D_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1} (\vec{x}-\vec{\mu})}.$$

3. Manhattan Distance: Represent distance between two points measured along the axes at a right angle.

$$L^1 = |x_2 - x_1| + |y_2 - y_1|$$

4. Chebyshev Distance: Represents distance between two vectors is the greatest of their differences along any coordinate dimension.

$$\max(\ |x_A\ -\ x_B|\ ,\ |y_A - y_B|\ )$$

5. Minkowski Distance: Represents a generalised distance metric that can be modified by substituting the value of ‘p’ to calculate the distance between two points.

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

6. Cosine Distance: Represents the degree of angle between two vectors. It is used when orientation matters not the magnitude.

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2 \cdot \sum\limits_{i=1}^{n} B_i^2}}$$

$$\text{cosine distance} = D_C(A,B) := 1 - S_C(A,B)$$

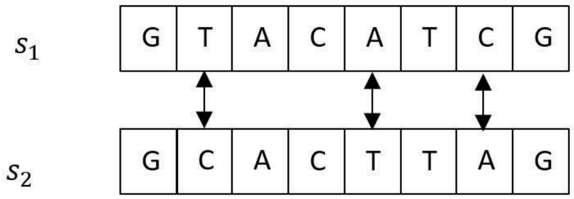
7. Hamming Distance: Represents the distance that measures minimum number of substitutions required to jump from one string (or bits) to another.

$$D_H = \sum_{i=1}^{k} \left| x_i - y_i \right|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

For example, In the following example Hamming Distance = 3,



s_1 G T A C A T C G

s_2 G C A C T T A G

Question 2. What are some real-world applications of distance-based classification algorithms?

Answer 2.

1. In clustering algorithms (K-Means and DBScan Algorithms)
2. Detect Outliers
3. Calculate distance between houses, length of wire connections
4. Warehouse Logistics
5. Computer Aided Manufacturing Processes
6. Find similarities between documents
7. Recommendation Systems
8. Human Pose Matching
9. Digital Communication
10. Error Computation

Question 3. Explain various distance metrics.
Answer 3. I’ve answered this in question 1.

Question 4. What is the role of cross validation in model performance?

Answer 4. It essentially evaluates model performance on an unseen set of data. It Prevents Overfitting, Ensures Model Generalization, Gives reliable estimate of Model Accuracy and helps us pick Optimal Parameters/Hyperparameters.

In an n-fold cross validation, we go through n iterations with n different set of validation datasets and average performance is calculated.

Question 5. Explain variance and bias in terms of KNN.

Answer 5. In KNN,

Bias is essentially how far the model’s predictions are from the true values on average. It’s caused by extremely simple assumptions made by the model.

Variance is the error caused by the model being too sensitive to small fluctuations in the training data. It measures how the model’s predictions change with changing training datasets.

When K is small, bias is low and fits the training data well. However, variance is high because small data changes lead to higher changes in predictions. The vice versa is also true.