19Z601- Machine Learning

Presented by

Ms.Anisha.C.D

Assistant Professor

CSE

Syllabus

INTRODUCTION : Types of Learning - Designing a learning system - concept learning - **Find-s Algorithm - Candidate Elimination** - Data Preprocessing - Cleaning - Data Scales - Transformation - **Dimensionality Reduction.** (9)

LINEAR MODELS : Linear Regression Models ,Maximum Likelihood Estimation - Least Squares - Bias-Variance Decomposition - Bayesian Linear Regression - **Linear Models for Classification, Probabilistic Generative Models - Probabilistic Discriminative Models - Linear Discriminant Analysis** (9)

Syllabus

NEURAL NETWORKS AND DECISION TREES : **Feed-forward Networks - Network Training - Delta Rule- Gradient Descent – Error Backpropagation - Regularization in Neural Networks - Generalisation - Decision Tree Learning- Representation - Inductive Bias- Issues** (9)

KERNEL AND GRAPHICAL METHODS : Constructing Kernels - Radial Basis Function Networks – Gaussian Processes - Maximum Margin Classifiers - SVM - Bayes Theorem - Naive Bayes - Bayesian Networks (9)

Syllabus

UNSUPERVISED AND REINFORCEMENT LEARNING : Measures of Similarity and Dissimilarity - Clustering - Partitioning methods - KMeans - Hierarchical Methods - Outliers - Reinforcement Learning - Reinforcement Learning Tasks - **Q-learning** (9)

Text Books and Reference Books

TEXT BOOKS:

1. **Tom Mitchell , "Machine Learning", McGraw Hill, 2017.**
2. **Christopher M Bishop , "Pattern Recognition and Machine Learning Learning", Springer, 2011.**

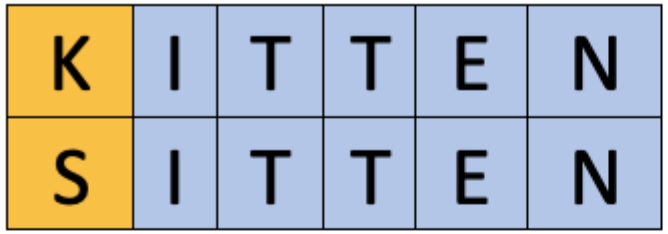
REFERENCES:

1. Ethem Alpaydin , "Introduction to Machine Learning", 3rd Edition, PHI Learning, 2015.
2. Trevor Hastie, Robert Tibshirani, Jerome friedman , "The Elements of Statistical learning", 2nd Edition, Springer, 2017.
3. Kevin Murphy , "Machine Learning - A Probabilistic Perspective", MIT Press, 2012.
4. Yaser S. Abu-Mostafa , "Learning from Data", AML, 2017.

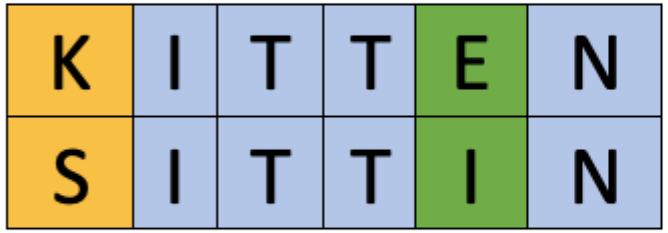
Measures of Similarity and Dissimilarity

Levenshtein Distance

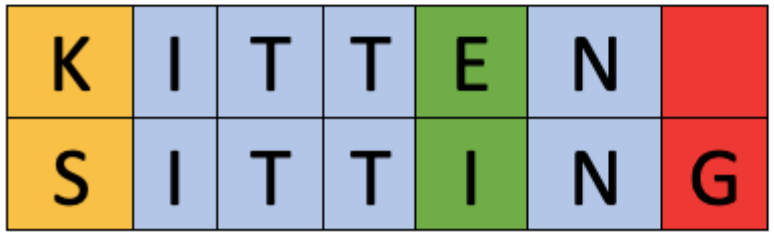
- is a **string metric for measuring the difference between two sequences**.
- Informally, the **Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other**



substitution
K with S



substitution
E with I



insertion G

https://medium.com/analytics-vidhya/levenshtein-distance-for-dummies-dd9eb83d3e09

Levenshtein distance algorithm – Uses

- Spell checking
- Speech recognition
- DNA analysis
- Plagiarism detection

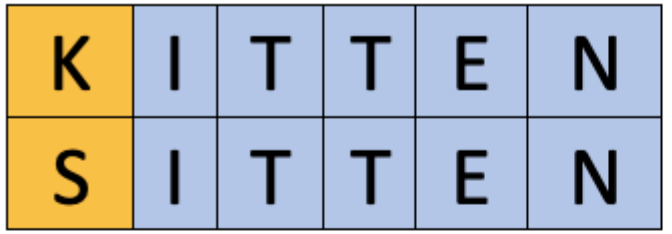
Levenshtein distance algorithm - Example

- Levenshtein Distance between **FORM and FORK is 1**. There is one substitution from M to K.

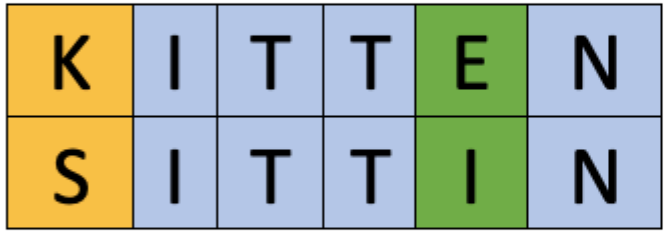
F	O	R	M	substitution
F	O	R	K	M to K

Levenshtein distance algorithm - Example

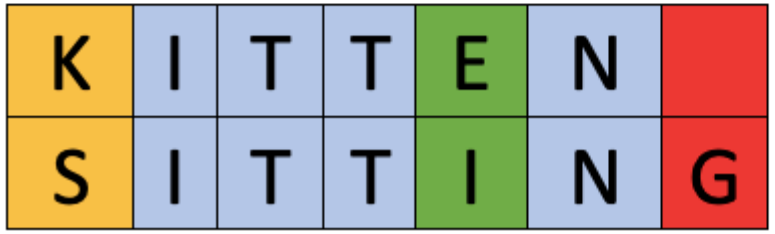
- Levenstein Distance between **KITTEN and SITTING is 3** because there are 3 characters edits.



substitution
K with S



substitution
E with I



insertion G

The Levenshtein distance Formula

$$D(i,j) = \begin{cases} 0, & i = 0,\ j = 0 \\ i, & j = 0,\ i > 0 \\ j, & i = 0,\ j > 0 \\ \min\{ & \\ \quad D(i, j-1) + 1, & \\ \quad D(i-1, j) + 1, & j > 0,\ i > 0 \\ \quad D(i-1, j-1) + \mathrm{m}(S_1[i], S_2[j]) & \\ \} & \end{cases},$$

https://medium.com/@art3330/levenshtein-distance-fundamentals-817b6f7f1718

Illustration : The Levenshtein distance between the strings APPLE and APRLE.

- Matrix Initialization

	-	A	P	P	L	E
-						
A						
P						
R						
L						
E						

When comparing an empty string with another empty string, it's evident that the number of operations to turn an empty string into another empty string is zero. Thus, the first value in the matrix will be 0.

	-	A	P	P	L	E
-	0					
A						
P						
R						
L						
E						

The only difference is that in the case of the column, we are considering the number of insertion operations. how many insertion operations are needed to turn an empty string into the substring APR.

	-	A	P	P	L	E
-	0	1	2	3	4	5
A	1					
P	2					
R	3					
L	4					
E	5					

For the cell [1,1], which represents the intersection of the characters A and A,

Need to check the cost of deletion, the cost of insertion, and the cost of substitution.

The given cell has the following values: i = 1, j = 1.

- **The cost of deletion is [i, j-1] + 1 = [1, 0] + 1 = 1 + 1 = 2.**
- **The cost of insertion is [i-1, j] + 1 = [0, 1] + 1 = 1 + 1 = 2.**
- **The cost of substitution is [i-1, j-1] + m(a, b) = [0, 0] + m(a, b) = 0 + 0 = 0.**

Here, m(a, b) is a function that compares two characters, returning 0 if the characters are identical and 1 if they are different. Therefore, min(2, 2, 0) is equal to 0, and hence the value of the cell D(1, 1) is 0.

	-	A	P	P	L	E
-	0	1	2	3	4	5
A	1	0				
P	2					
R	3					
L	4					
E	5					

Final Matrix

	-	A	P	P	L	E
-	0	1	2	3	4	5
A	1	0	1	2	3	4
P	2	1	0	1	2	3
R	3	2	1	1	2	3
L	4	3	2	2	1	2
E	5	4	3	3	2	1