ADVANCED DATA ANALYTICS

UE23AM343AB1

Unit 1 Case Study

Dr. Bhaskarjyoti Das
Department of Computer Science and Engineering
in AI & ML

A case for Statistical Tests in Machine Learning

- You have a **dataset of 2 populations (say USA and India) in 2 CSV**. There is a common **class label assigned**. The CSVs has **some categorical (nominal) features** and some **numerical features**.
- You are trying to **design a classifier that can predict class labels** (output is categorical)
- You are thinking of **combining the two datasets**
- Since the USA dataset is larger, you want to **build the ML model on that and predict for India dataset!**

Open Questions

- You need to find **the more significant features, eliminating correlated variables as they carry the same information**
  - the nominal features
  - The numerical features
- You are trying to **analyze the numerical features**
  - You need a test to determine if they are correlated
  - Will you use parametric or nonparametric? How to test ?
  - You also want to see if they are significant
- You are trying to **analyze the numerical features**
  - How would you test the significance of the numerical feature with the categorical output ?
- Since you are short of dataset**, you wan to combine the two CSVs or use one to develop your ML model and predict on the other.**
  - **how would you test the equivalence of two distributions** (USA and India). The population sizes are different. Which test will work ?
- Which statistical library will be used ?

Your To Do

- Come up with an approach
- Provide an answer to each of the questions in the previous slide

THANK YOU

Dr. Bhaskarjyoti Das
Professor, Department of Computer Science and Engineering in AI & ML, PES University, Bengaluru
Email: bhaskarjyotidas@pes.edu