CSE 475: Statistical Methods in AI **Monsoon 2018**

Lec 1: Introduction

Lecturer: Anoop M. Namboodiri *Date: Jul. 30*

Class Logistics

- Lectures and Tutorials
- Grading Policy
  - Mini Projects (20%)
  - Homeworks (30%)
  - Mid Terms (30%)
  - Final Exam (20%)
- Homeworks and Shiksha Portal
- Assignments / Mini Projects
- See moodle page for details/resources

Topics Covered

1. *Category or Class*: A collection of objects with similar properties, which can together be given a name.
2. *Cognition and Recognition*: Cognition is the process of formation of new concepts from a set of unlabelled examples. Often referred to as unsupervised learning. Recognition is the process of associating a new (unlabelled) sample to a class that we already know of.
3. *Inter-Class and Intra Class Variability*: Similarity between objects of different classes and Differences between objects of the same class. These make the problem difficult.
4. *Features and Feature Vector*: A set of informative measurements taken from an object, which forms the representation of the object. Often features that humans think of are not meaningful or are difficult to extract automatically.
5. *Recognition*: The process of mapping an input feature vector to a class label. Depending on applications, the input and output changes, but the process outline remains the same.
6. *Recognition Applications*: Think of the input and output (class labels) for each of the following problems.
   - Speech Recognition
   - Speaker Identification
   - Non-destructive Testing
   - Natural Resource Identification
   - Character Recognition (OCR)
   - Web Search
   - Fingerprint Identification
   - Identification and Counting of Cells
   - Disease detection/diagnosis from EEG/EKG
   - Aerial Reconnaissance
7. *Recognition Pipeline*: Works in two stages.
   - *Training*: **Labelled Data** → Feature Extraction → Learning Classifier → **Model**.
   - *Testing*: **Unlabelled Data** → Feature Extraction → Model → **Class Label**.
8. *Approaches to Classification*:
   - *Generative*: Model the distribution of features among samples of a class. Assign class labels to new samples based on the distribution function value for the feature vector.
   - *Discriminative*: Model the boundary between classes in the feature space. Assign class labels to new samples based on which side of the boundary, the sample falls in.
9. *Complexity of the boundary*:
   - Given two boundaries that separate training samples, do we choose a simple one or complex one? (Occam's razor)
   - Are complex boundaries that separate the training samples better than simple boundaries that make some errors?