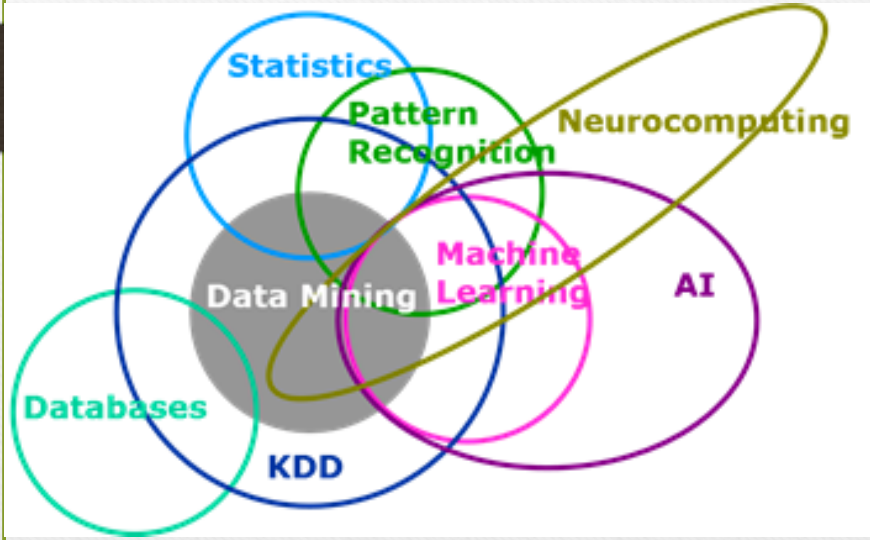
Statistical Machine Learning (Pattern Recognition)

CSE 342/542

Statistical Machine Learning (formerly Pattern Recognition)

- The course will introduce salient topics in machine learning and pattern recognition with emphasis on statistics.
- Fundamentals and advanced theoretical and mathematical concepts related to classification techniques and learning paradigms will be discussed.

The field of Data Science



Statistics	Machine Learning
Estimation	Learning
Classifier	Hypothesis
Data Point	Example/ Instance
Regression	Supervised Learning
Classification	Supervised Learning
Covariate	Feature
Response	Label

Relation between AI, ML, PR

- Artificial Intelligence: Started first
- Pattern Recognition: Started in 1970's, focused on learning interpretable patterns in data
- Machine Learning: Started in late 1980', with emphasis on reducing the error rate
- Data mining, deep learning, information retrieval are related areas

Statistical Machine Learning (formerly Pattern Recognition)

- **Pre-requisites**
  - Programming
  - Probability, statistics and linear algebra
- **Post Condition**
  - Understand various key paradigms for pattern classifications and statistical machine learning, and approaches in each
  - Ability to apply suitable feature extraction and classification technique to solve a given classification problem

Topics to be Covered

- **Introduction:** Review of probability, performance evaluation, generative and discrimination classification
- **Bayesian decision theory**: Minimum error rate classification, Discriminant function and decision surfaces, Error Bounds: Chernoff and Bhattacharya, Missing and Noisy Features
- **Parameter Estimation**: Parametric (MLE, Bayesian)
- **Discriminant Analysis**: Principal Component Analysis, Linear Discriminant Analysis, and Subclass Discriminant Analysis

Topics to be Covered

- **Hidden Markov Models**
- **Unsupervised Learning:** Unsupervised Bayesian learning, Hierarchical clustering, Online clustering
- **Algorithm Independent Machine Learning:** Bias and variance tradeoff, bootstrapping, No free lunch, Ugly Duckling, Bagging, Boosting, and Combining classifiers
- **Non-parametric Regression**
- **Ensemble Learning**

Reading Material

- Textbook
  - Pattern Classification by Duda, Hart and Stork, Wiley Interscience, 2000
- Reference books
  - Pattern Recognition by S. Theodoridis, K. Koutroumbas, Elsevier/Academic Press
  - Pattern Recognition and Machine Learning by C. M. Bishop, Springer
  - Introduction to Statistical Machine Learning by Masashi Sugiyama, Elsevier

Evaluation

- Assignments: 25%
- Midsem Exam: 15%
- Endsem: 20%
- Project: 30% (Continuous evaluation)
- Quizzes: 10%

Cheating and Collaboration

- No collaboration is allowed during quizzes and exams
- In the assignments, you are permitted to discuss the questions with others. However, you must write up your own solutions to these questions.
- Any indication to the contrary will be considered an act of academic dishonesty.
- Institute policy for academic dishonesty

Grading Policy: Absolute Grading

Marks	Grade
>= 95 %	A+
>= 87 %	A
>= 80 %	A-
>= 72 %	B
>= 65 %	B-
>= 57 %	C
>= 50 %	C-
>= 35 %	D
< 33 %	F

Along with scoring more than 33% in total, you have to score at least 35% in both exams.

Statistical Machine Learning

- Course Website: https://www.usebackpack.com/iiitd/ w2018/cse542
- Course Mailing List: cse542@iiitd.ac.in

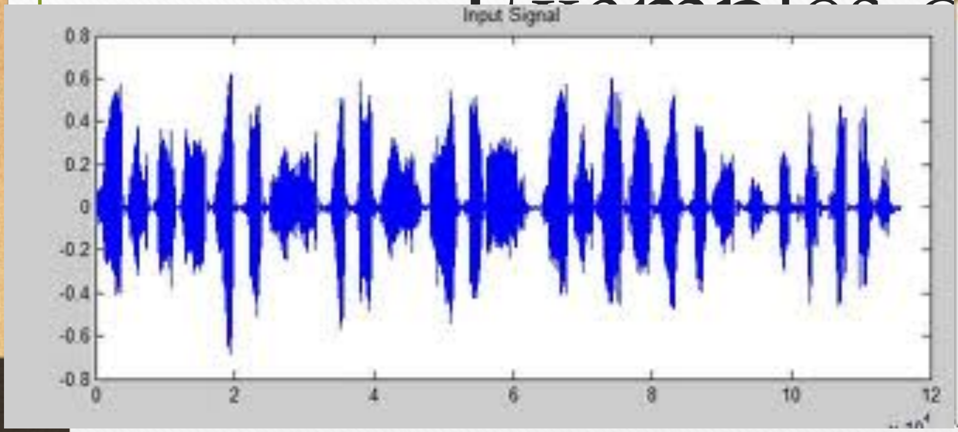
- Course slides are compiled from several resources on the internet

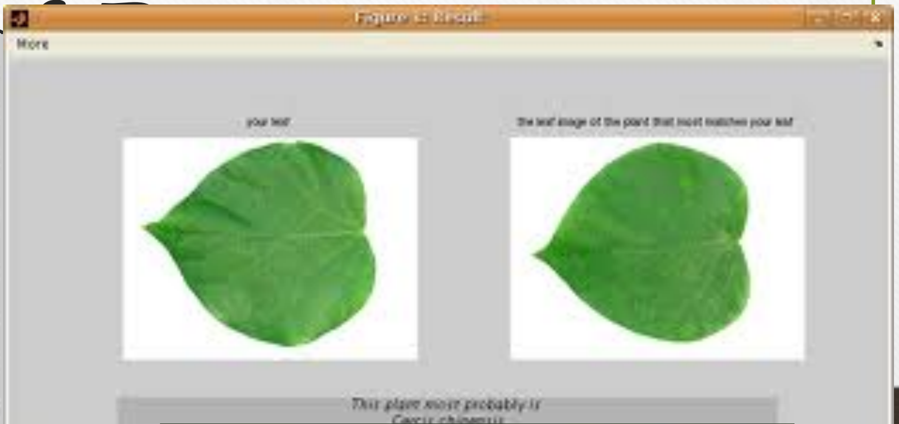
What is Pattern Recognition

- “The assignment of a physical object or event to one of several pre-specified categories” -- Duda & Hart

- “A pattern is the *opposite of a chaos; it is an entity vaguely defined, that could be* given a name.” (Watanabe)

Examples of P

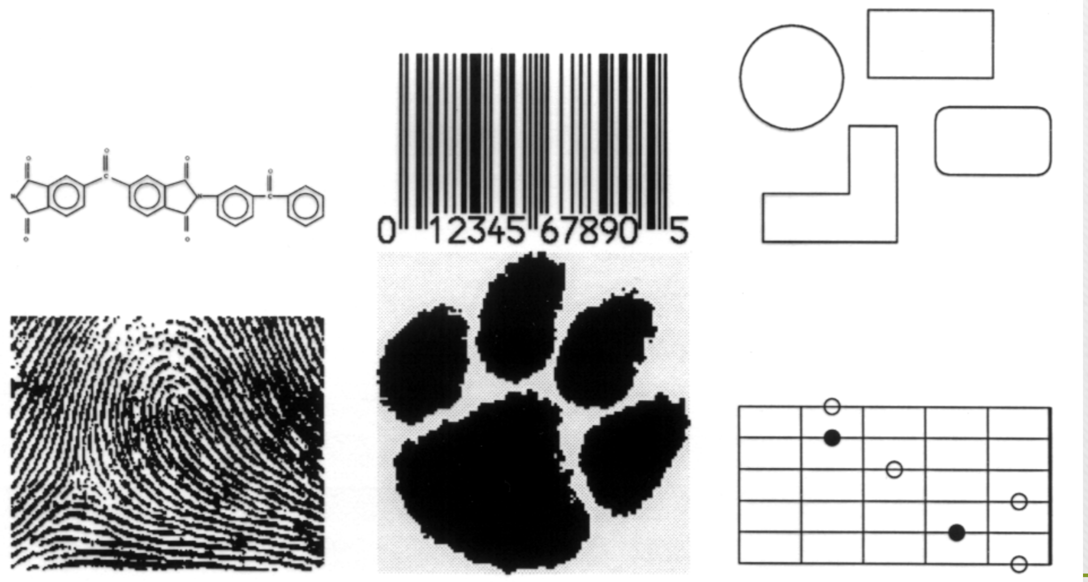




your leaf

the leaf image of the plant that most matches your leaf

This plant most probably is Cercis chinensis



Examples of Patterns

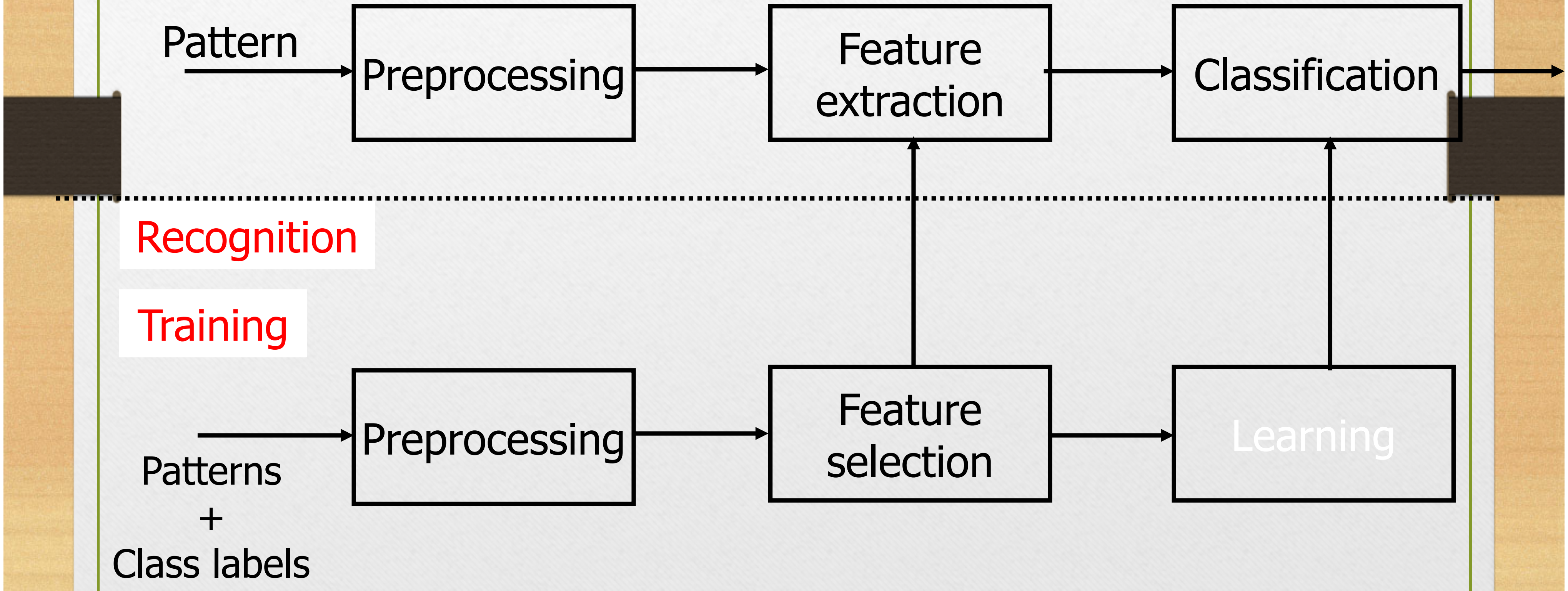
- Insurance, credit cards, loans
  - Income, number of dependents, credit worthiness, loan amount
- Web documents
  - Keywords, content, organization
- Medical data
  - Symptoms, test reports, previous history
- Emotions
  - Audio – pitch, spoken text, frequency
  - Images – facial features

Pattern Class

- A **pattern class** (or category) is a set of patterns sharing common attributes and usually originating from the same source

- Emotions: happy, sad, angry, surprised
- Web documents: sports, medicine, technology, politics
- Fruits: apple, mango, guava

Statistical Pattern Classification



Important Issues

- Noise / Segmentation
- Data Collection / Feature Extraction
- Pattern Representation / Invariance/Missing Features
- Model Selection / Overfitting
- Prior Knowledge / Context
- Classifier Combination
- Costs and Risks
- Computational Complexity

Issue: Noise

- Various types of noise (e.g., shadows, conveyor belt might shake, etc.)
- Noise can reduce the reliability of the feature values measured.
- Knowledge of the noise process can help improve performance.

Issue: Segmentation

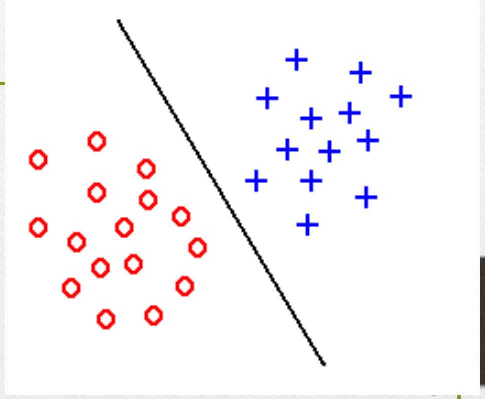
- Individual patterns have to be segmented
  - How can we segment without having categorized them first ?
  - How can we categorize them without having segmented them first ?
- How do we "group" together the proper number of elements ?

Issue: Data Collection

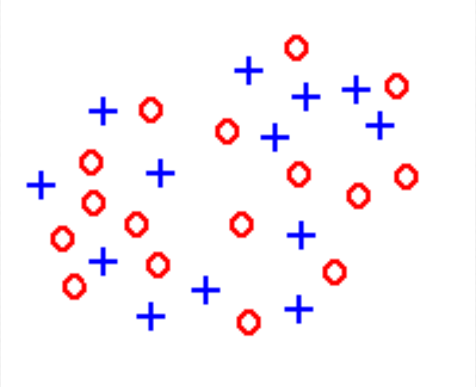
- How do we know that we have collected an adequately large and representative set of examples for training/testing the system?

Feature Extraction

- Extract features which are good for classification
- Good features
  - Objects from the same class have similar feature values.
  - Objects from different classes have different values.



“Good” features



“Bad” features

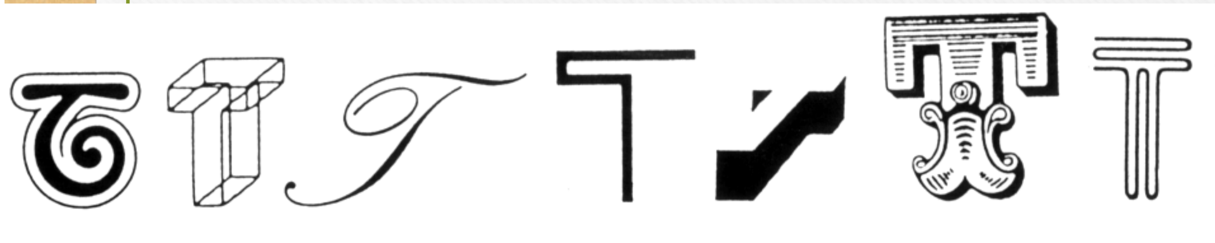
Issue: Feature Extraction

- It is a domain-specific problem which influences classifier's performance.
- Which features are most promising ?
- Are there ways to automatically learn which features are best ?
- How many should we use ?
- Choose features that are robust to noise.
- Favor features that lead to simpler decision regions.

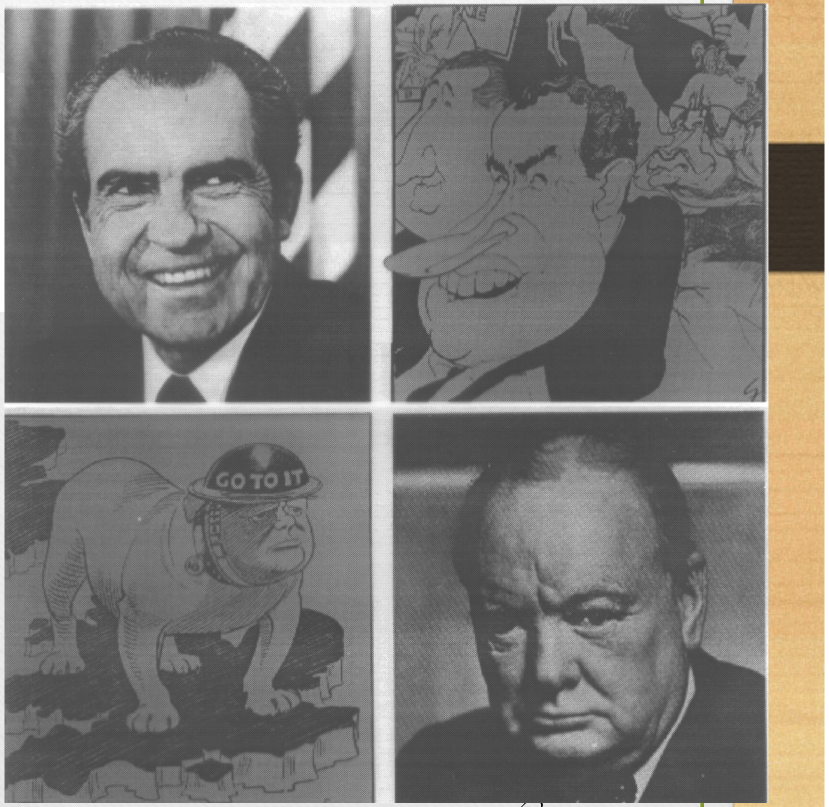
Issue: Pattern Representation

- Similar patterns should have similar representations
- Patterns from different classes should have dissimilar representations
- Pattern representations should be invariant to transformations such as:
  - translations, rotations, size, reflections, non-rigid deformations
- Small intra-class variation, large inter-class variation

Intra-class Variability



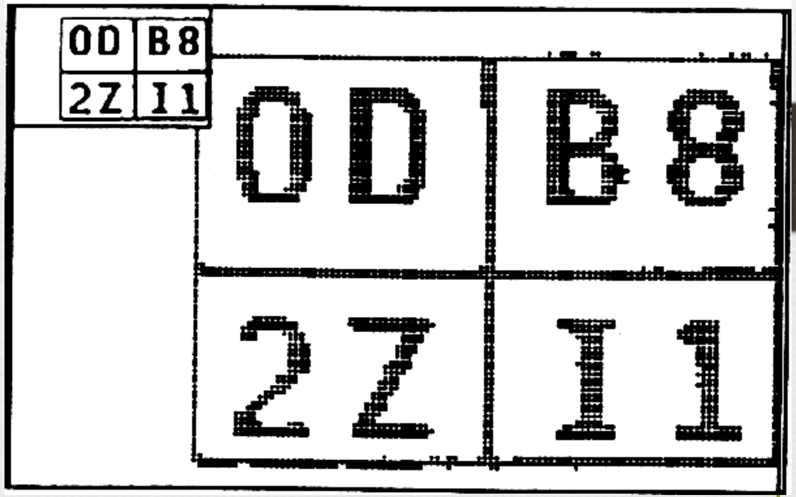
The letter "T" in different typefaces



John P. Frisby, *Seeing. Illusion*, *Brian and Mind*, Oxford University Press, 1980

Inter-class Similarity

Identical twins



Characters that look similar

Issue: Missing Features

- Certain features might be missing (e.g., due to occlusion).
- How should the classifier make the best decision with missing features ?
- How should we train the classifier with missing features ?

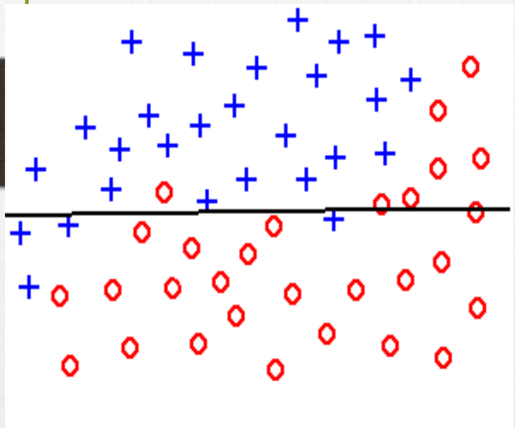
Issue: Model Selection

- How do we know when to reject a class of models and try another one ?
- Is the model selection process just a trial and error process ?
- Can we automate this process ?

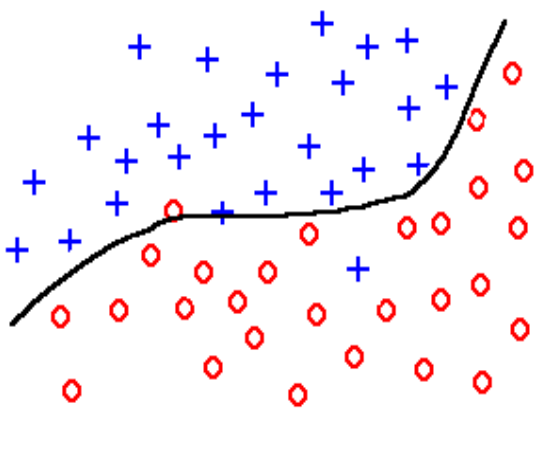
Issue: Overfitting

- Models complex than necessary lead to overfitting (i.e., good performance on the training data but poor performance on novel data).
- How can we adjust the complexity of the model ? (not very complex or simple).
- Are there principled methods for finding the best complexity ?

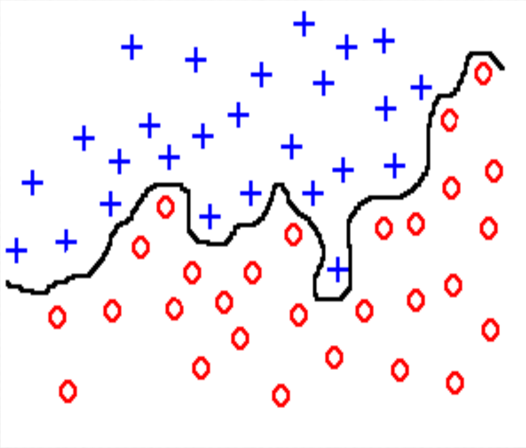
Overfitting and Underfitting



Underfitting



Good fit



Overfitting

Issue: Domain Knowledge

- When there is not sufficient training data, incorporate domain knowledge:
  - Model how each pattern is generated (analysis by synthesis) - this is difficult !! (e.g., recognize all types of chairs).
  - Incorporate some knowledge about the pattern generation method. (e.g., optical character recognition (OCR) assuming characters are sequences of strokes)

Issue: Context

How m ch info mation are y u mi sing

Issue: Classifier Combination

- Performance can be improved using a "pool" of classifiers
- How should we combine multiple classifiers ?

Issue: Costs and Risks

- Each classification is associated with a cost or risk (e.g., classification error)
- How can we incorporate knowledge about such risks ?
- Can we estimate the lowest possible risk of any classifier ?

Issue: Computational Complexity

- How does an algorithm scale with
  - the number of feature dimensions
  - number of patterns
  - number of categories
- Brute-force approaches might lead to perfect classifications results but usually have impractical time and memory requirements.
- What is the tradeoff between computational ease and performance ?

General Purpose PR/ML Systems?

- Humans have the ability to switch rapidly and seamlessly between different pattern recognition tasks
- It is very difficult to design a device that is capable of performing a variety of classification tasks
    - Different decision tasks may require different features.
    - Different features might yield different solutions.
    - Different tradeoffs (e.g., classification error) exist for different tasks.

Thanks.