Classification I

UW ECE 657A - Core Topic

Mark Crowley

Lecture Outline

1. Supervised vs. Unsupervised Machine Learning
   - Clustering vs. Classification
   - Classification
2. Some Definitions
3. Similarity Based Classifiers
   - k-Nearest Neighbor Classifier
4. Density Based Classifiers
   - Parzen Window Density Estimation

Lecture Outline

1. Supervised vs. Unsupervised Machine Learning
   - Clustering vs. Classification
   - Classification
2. Some Definitions
3. Similarity Based Classifiers
   - k-Nearest Neighbor Classifier
4. Density Based Classifiers
   - Parzen Window Density Estimation

Outline

1. Supervised vs. Unsupervised Machine Learning
   - Clustering vs. Classification
   - Classification
2. Some Definitions
3. Similarity Based Classifiers
   - k-Nearest Neighbor Classifier
4. Density Based Classifiers
   - Parzen Window Density Estimation

Descriptive versus Inferential Analysis

- We have data (samples). This data is a sample of a population (more than just the measured or observed sample).
- **Descriptive Analysis** is the type of analysis and measures that seek to describe and summarize the data, the available samples. We can not in general use it for interpretation of unobserved data.
- **Inferential Analysis (predictive)** is the type of analysis that can describe measures over the population of data. That is observed and unobserved.

Another way to look at it: Machine Learning [Murphy, 2012]

Supervised Learning

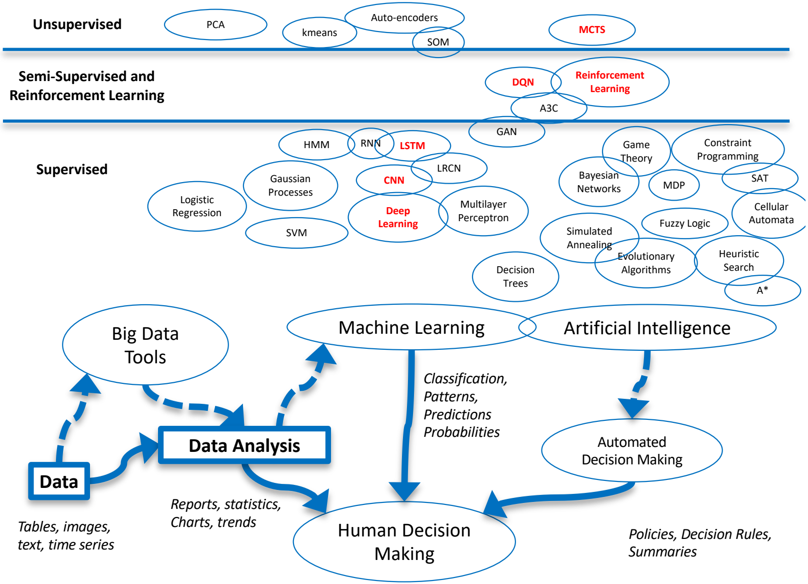
- Also called: predictive, inferential
- Given inputs and outputs
- Model based: model or distribution is known
  - Parameter Estimation: distribution known, fitting parameters
  - Linear Regresion, least squares estimation
- output is linear: regression, prediction
- output is categorical (label) : **classification**

Unsupervised Learning

- Also called: descriptive, knowledge discovery
- Given inputs only
- output is categorical : clustering

Reinforcement Learning

- Input and output given only when action (prediction, choice, activity) provided.
- Given feedback about utility of choice made.
- No given model or labels ahead of time.
- Learning is interactive.



Unsupervised

PCA

kmeans

Auto-encoders

SOM

MCTS

Semi-Supervised and Reinforcement Learning

DQN

Reinforcement Learning

A3C

GAN

Supervised

HMM

RNN

LSTM

LRCN

CNN

Gaussian Processes

Logistic Regression

SVM

Deep Learning

Multilayer Perceptron

Decision Trees

Game Theory

Constraint Programming

Bayesian Networks

MDP

SAT

Cellular Automata

Fuzzy Logic

Simulated Annealing

Evolutionary Algorithms

Heuristic Search

A*

Big Data Tools

Machine Learning

Artificial Intelligence

Data Analysis

Data

Tables, images, text, time series

Reports, statistics, Charts, trends

Classification, Patterns, Predictions Probabilities

Automated Decision Making

Human Decision Making

Policies, Decision Rules, Summaries

Clustering vs. Classification

Clustering

- Unsupervised
- Uses unlabeled data
- Organize patterns w.r.t. an optimization criteria
- Notion of similarity
- Hard to evaluate
- Example: K- means, Fuzzy C- means, Hierarchical

Classification

- Supervised
- Uses labeled data
- Requires training phase
- Domain sensitive
- Easy to evaluate
- Examples: Bayesian, KNN,SVM, Decision Trees

Classification Definition

Definition:

- Classification is a learning method that uses training samples (with known class labels) to learn how to assign the samples into their proper classes. The task of the clasifier is to use the feature vector provide by the feature extractor to assign the object (datapoint) to a category.
- This learning can then be used to label test samples (with unknown labels).
- Classification can be facilitated if the features (or a subset of them) of samples in the same class share characteristics that discriminate them from other classes.
- It can be seen as the discrete form of Prediction, can you map the input values to a discrete set of output values rather than to a continuous number.

Classification

- How do we design a classifier that can deal with the variability in feature values?
- Designing or selecting a classifier for the data or application is not an obvious task.

Lecture Outline

1. Supervised vs. Unsupervised Machine Learning
   - Clustering vs. Classification
   - Classification
2. Some Definitions
3. Similarity Based Classifiers
   - k-Nearest Neighbor Classifier
4. Density Based Classifiers
   - Parzen Window Density Estimation

Notation:

$\mathbb{R}$: the set of real numbers

$\mathrm{x}, y, \mathcal{D}$: test data point, true output label, training data $\mathcal{D} = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$

K: number of nearest neighbors to include in kNN algorithm

$N_k(\mathrm{x}, \mathcal{D})$: set of K nearest neighbors of point x

c: class label

$\mathbb{I}(\texttt{COND})$: **Indicator function** which equals 1 if and only if `COND` is true.

$\kappa(\mathrm{x})$: kernel mapping function for Kernel density estimation. Takes an input count of density as input and maps it directly to a new space.

$\kappa(\mathrm{x}, \mathrm{x}')$: binary kernel mapping function returning similarity score between two points in the remapped space.

h : Parzen window bandwith, determines size of neighborhood included in score.

Classification Definition

- Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ where each datapoint $x_i \in X$ contains F features denoted $x_{i,f}$,
- and given a set of classes $C = \{C_1, \ldots, C_K\}$,
- the **classification problem** is to define a mapping $\delta(x_i) : X \rightarrow C$ where each x_i is assigned to one class.
- A **class**, C_j, contains precisely those points mapped to it, no more and no less.
- Note: the classes are predefined, nonoverlapping and partition the entire set of datapoints.

Considerations for Designing a Classification Approach

1. Data, labelled, one class, multiclass, overlap, missing values
2. Training and Error Estimation procedures
3. Classification methods
4. Performance measures

Classifier Methods

Similarity Based Approach

Template matching

Nearest Mean classifier

Neural Network classifiers

Deep Learning

Other

Decision Trees

Random Forests

Probabilistic Approach

Bayesian Classification

Logistic regression

Decision Boundary Approach

Geometrical

Neural Networks

SVM

Lecture Outline

1. Supervised vs. Unsupervised Machine Learning
   - Clustering vs. Classification
   - Classification
2. Some Definitions
3. Similarity Based Classifiers
   - k-Nearest Neighbor Classifier
4. Density Based Classifiers
   - Parzen Window Density Estimation

Parametric vs Non-parametric Models

Analysing some data, produce a probability model

Supervised: $p(y|\mathbf{x}, \theta)$

Unsupervised: $p(\mathbf{x}, \theta)$

Number of parameters θ of the model

Parametric: fixed in size, regardless of the amount of data $\mathcal{N}(\mu, \sigma)$

- Usually fast.
- Use strong assumptions, could be wrong.
- Can use MLE or MAP.

Examples: linear regression, logistic regression, Naive Bayes, Simple Neural Networks, k-means, HMMs

Parametric vs Non-parametric Models

Non-parametric: number of parameters grows with the amount of training data.

- More flexible.
- Slower than parametric or even completely intractable for large datasets

Examples: k-nearest neighbors, Support Vector Machines (SVMs), Parzen Window/KDE, Decision Trees, Gaussian processes

k-Nearest Neighbor (KNN) Classifier

A simple example of a supervised, non-parametric, classification model. General Algorithm

- For some test datapoint x define a set of K points in training set *nearest* to x
- Count how many members of each class are in the nearest neighbors set
- Return empirical fraction for each class as a probability
- Optionally take highest probability as class label for x

$$p(y = c|\mathrm{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i} \in N_K(\mathrm{x}, \mathcal{D}) \mathbb{I}(y_i = c)$$

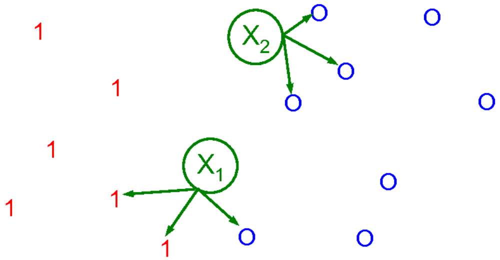
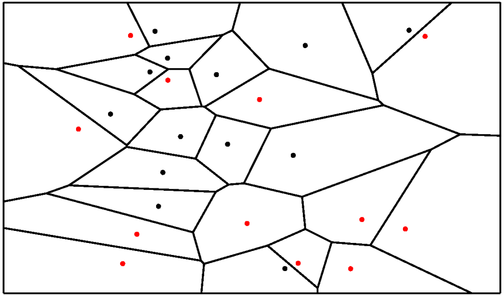


Figure: KNN in 2D for K=3.

$$
\begin{aligned}
x_1 &: p(y=1|\mathrm{x}_1,\mathcal{D},K=3) = 2/3 \\
x_1 &: p(y=0|\mathrm{x}_1,\mathcal{D},K=3) = 1/3 \\
x_2 &: p(y=1|\mathrm{x}_2,\mathcal{D},K=3) = 0/3 = 0 \\
x_2 &: p(y=0|\mathrm{x}_2,\mathcal{D},K=3) = 3/3 = 1
\end{aligned}
$$

Special case K=1

If $K = 1$ then we have **1-Nearest Neighbour**.
Induces a Voronoi Diagram. All points in polygons have point as their nearest point.



KNN Example

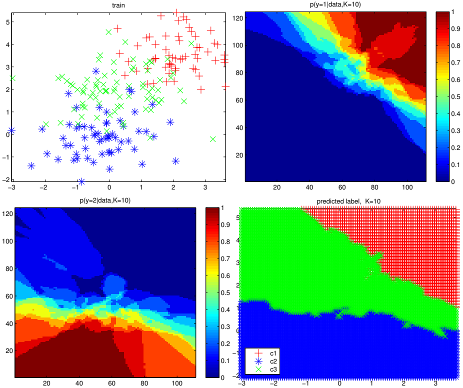


Figure: (a) 3-class dataset. (b) Prob of class 1 for $K = 10$. (c) Class 2. (d) MAP estimate of class label.

Comments on KNN

- Non-parametric because the number of neighboring points used depends on the data.
- Biased towards classes with more samples.
- Sensitive to noise in the data (eg. random neighbors).
- Computationally expensive in large dimensions or large K
  - Could compute distance using only some dimensions (which ones?)
  - Remove redundant points from training set (eg. ones surrounded by same class)
  - For low dimensions could use search trees to organize training points into independent subsets.

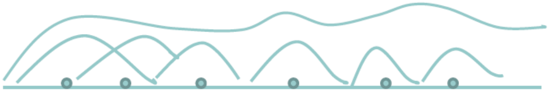
Lecture Outline

1. Supervised vs. Unsupervised Machine Learning
   - Clustering vs. Classification
   - Classification
2. Some Definitions
3. Similarity Based Classifiers
   - k-Nearest Neighbor Classifier
4. Density Based Classifiers
   - Parzen Window Density Estimation

Density Estimation

- Estimate the density of points in a part of the dataset.
- **parametric approach:** Could set the location of one or multiple distributions to model clusters and fit parameters to the data.
- **non-parametric approach:** Allocate a cluster for each training data point and combines the distributions.
  - Parzen window density estimation
  - Kernel density estimation (KDE)

Parzen Window



$$\hat{p}_h(\mathrm{x}|\mathcal{D}) = \frac{1}{Nh}\sum_{i=1}^{N}\kappa\left(\frac{\mathrm{x}-\mathrm{x}_i}{h}\right)$$

where *h* is called the **bandwidth** (usually chosen by cross-validation on the risk).

Kernel Function κ

$$\begin{aligned}\kappa(\mathrm{x}) &= \mathcal{N}(\mathrm{x}|\mathrm{x}_i, \sigma^2\mathrm{I}) \\ &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \quad \text{Gaussian kernel} \\ \text{or } &= \mathcal{I}(|x| \leq 1) \quad \text{Boxcar kernel}\end{aligned}$$

Parzen Window Example

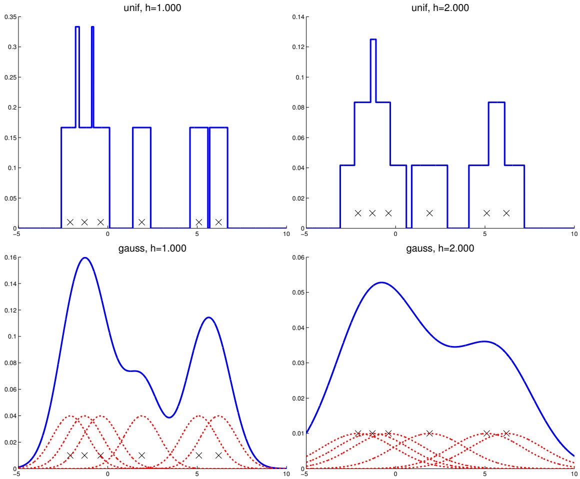


Figure: Boxcar kernel vs Gaussian kernel on $h = 1$ and $h = 2$

From KDE to KNN

- We can define K-Nearest Neighbors using the Parzen window or KDE approach
- Instead of a single bandwidth h, define h_i for each datapoint x_i such that the K points nearest to x_i are in the box.