Decision Making Technologies for Business

BT5212
2018/2019 Semester I

Week 1

Associate Professor
HUANG, Ke-Wei

Today's Class

1. Course Logistics
2. Overview of Machine Learning and Data Mining
3. Different Types of Algorithms
4. Statistics versus Data Mining: Causality versus Prediction
5. Terminologies of Data Mining
6. The Easiest Algorithm: K-Nearest Neighbor Method

A Bit About Myself

- Academia
  - Undergraduate, Electrical Engineering. (National Taiwan University)
  - MBA, Finance. (National Taiwan University)
  - MSc, Information Systems. (New York University)
  - Ph.D., Information Systems. (New York University)
  - Joined NUS in 2007. (IS and Economics departments, 100% IS since 2009.)
- Contact Information
  - COM2, Room 04-18
  - Office No. 6516-2786
  - Email: huangkw@comp.nus.edu.sg (typically I can reply with 48 hours and mostly 24 hours. If I do not reply, you can email again.)
  - Office Hour: by appointment.

Course Overview

Class days, time, place

- **2 Sessions on Tuesday and Wednesday**
- **2 hours of lecture and 1 hour of tutorial in R**
- **11 sessions + Project Presentation**

Class requirements

- **60%: 5 Individual Assignments**
- **40%: Final Project, including Presentation**

Two sessions are identical

Course Overview

- Why 2 hours of lecture and 1 hour tutorial by TA?
- SoC classes are 2 hours only, whereas BIZ classes are 3 hours
- Lectures more about how the algorithms work, applications of algorithms, pros and cons of each algorithm…etc.
- Tutorial will help you (1) learn and practice R (2) understand instructions and solutions of 5 assignments

Part 2: Overview

- There is a surge of interests of data science in recent years.
- Most of the algorithms have been invented in academia for 20-30 years but not widely adopted. What happened in the last 5-10 years?

1. Internet and mobile commerce created a lot of (public) data, much more than 1980s or 1990s.
2. Advances in database technologies make it possible to save, merge, and use Big Data.
3. Several media coverage of the significant success of AI, Machine Learning, and DM.
4. Large companies started to invest in data analytics and they can gain positive ROI => more companies followed the trend.

Successful Marketing Applications

Most new inventions are in marketing

- Demand forecasting
- Segmentation of customers for targeted advertising
- Analyzing the ROI of advertising campaigns or promotion activities
- Analyzing social media or social network
- Product line design (for each market segment or at each retail location)
- Deciding the retail location
- Multi-channel management (physical, e-commerce, m-commerce)

Successful Financial Applications

- Capital market analytics (relatively old and existed before the surge of interests in data science)
- Fraud detection: credit card transactions, life/health/property insurance claims, money laundering
- Consumer banking analytics (similar to marketing)
- Cash management: predicting the cash flow and better managing cash for banks or firms
- Predicting ATM activities
- Internal operational analytics: auditing and internal control

Successful Operational Applications

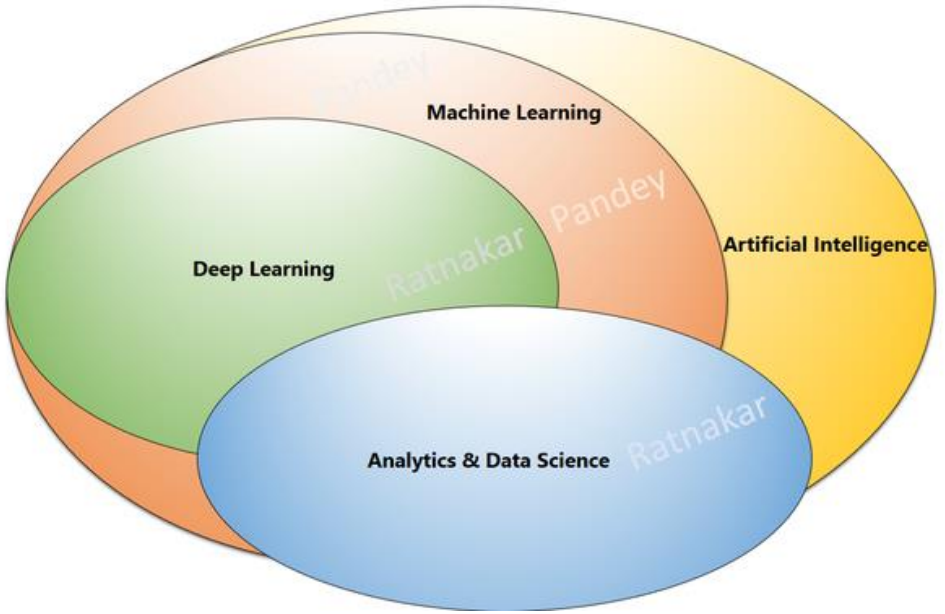
- Inventory management
- Manufacturing analytics: using data mining to reduce the defect production rate
- Production scheduling
- Transportation optimization
- Optimization of energy use in office buildings and especially factories
- Robotics in manufacturing
- Predicting the traffic within a mall, amusement park, or airport
- HR management

Public policy applications

- Prediction of popular election outcomes
- Projection of areas where criminal activity is most likely
- Forecasts of weather behavior and long-term climate changes
- Effects of new policies
- Identification of unwanted spam messages in e-mail
- Actuarial estimates of financial damage of storms and natural disasters
- Discovery of genetic sequences linked to diseases (health care)

Components of Data Mining

- Two reference disciplines: 70-80% of Computer Science and 20-30% of Statistics
- 3 related areas: Artificial Intelligence, Machine Learning, and Data Mining? The boundary is sometimes blur
- Different references may give you difference categorization.



What is Artificial Intelligence?

- Artificial intelligence (AI, also machine intelligence, MI) is intelligence displayed by machines, in contrast with the natural intelligence (NI) displayed by humans and other animals (from Wiki).
- There are three waves of AI success that created hype in AI/Machine Learning interests. (1) IBM Deep Blue beat world champion in western chess in 1996-1997. (2) IBM Watson won in a game show called Jeopardy! In 2011. (3) Google AlphaGo beat world champion in Go Chess game in 2015-2016.
- AI is much broader and this class focused only on the last two about ML or Data Mining.
- AI is still far from human intelligence!!!

Artificial Intelligence Examples

- Robotics from vacuum to auto-driving cars and industrial robotics,
- AI for air-planes
- AI for all Chess games: Chinese go-chess is the most difficult one for AI
- AI for gaming: TV video games and PC games
- Q&A system: Siri, Android, Amazon Echo, IBM Watson
- Algorithm trading
- Music recognition: Shazam app
- AI music composer and AI article writer
- AI customer service call center
- Facial recognition
- Predictive modeling and decision making systems (machine learning and data mining).

What is Machine Learning?

- Machine learning focuses on teaching computers how to use data to solve a problem, while data mining focuses on teaching computers to identify patterns that humans then use to solve a problem. (Lantz)
- Machine learning involves algorithms that teach the machine learn from experience or data.
- **Reinforcement learning** studies how to let machines learn from its own experience. This is beyond the scope of this module.
- We will cover the ML subjects that are more relevant to data mining and predictive modelling.

What is Data Mining?

- Data mining is concerned with the generation of novel insights from large databases. Data mining involves a systematic hunt for nuggets of actionable intelligence. (Lantz)
- Data mining is the extraction of knowledge from data, via technologies that incorporate these principles. (Provost and Fawcett)
- Virtually all data mining involves the use of machine learning, but not all machine learning involves data mining. (Lantz)

3. Categories of Data Mining Algorithms

1. The largest category => Classification algorithms
   - Predicting a binary or categorical variable. For example, predicting a customer will buy or not, a stock up or down, or a patient has cancer or not?
2. Numerical variable prediction
   - Similar to OLS regression, predicting customer spending, predicting stock return, predicting how long a patient can live
3. Clustering algorithms
   - We do not have labels (dependent variables) and we try to group records/examples together.
   - For example, customer segmentation based on demographics.
4. Association algorithms (market basket analysis)
   - E.g., Recommendation Systems in E-Commerce

Classification vs. Numeric Prediction

- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (called class labels) in a classifying attribute and uses it in classifying new data
  - More examples of applications
    - Credit/loan approval:
    - Medical diagnosis: if a tumor is cancerous or benign
    - Fraud detection: if a transaction is fraudulent
    - News article categorization: which category it is
- Numeric Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Most algorithms can handle both.

Supervised vs. Unsupervised Learning

Sometimes you will heard these two terminologies.

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set

- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Different Categories of Algorithms

Traditional Methods

1. Classification algorithms
   - (This module) KNN (the easiest idea)
   - Naïve Bayes (skipped due to limitation of time),
   - (This module) Decision Trees, Rules-based,
   - (This module) Neural Network
   - (This module) Support Vector Machine (SVM)
   - (This module) Gradient Boosting Machines
   - Genetic Algorithms (interesting idea but not cover in the textbook and not covered this semester)
2. Clustering algorithms
   - (Statistics) K-Means, among several other algorithms
   - Outlier detection analysis

Different Categories of Algorithms

Modern Methods for Improving Prediction Performance

1. (This module) Random Forest
2. (This module) Bagging Methods
3. (This module) Boosting Methods
   - AdaBoost
   - XGBoost
   - LightGBM
4. (This module) Stacking of classifiers

"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."

Algorithms for Unstructured Data

1. (This module) Text Mining
   - Dictionary Based Approach
   - Supervised text classification
   - Unsupervised topic modeling

2. (This module) Examples of applications of using unstructured data
   - Image (covered)
   - Audio (not covered)
   - Video (not covered)

4. Statistics and Data Mining

- Data mining is a research area of computer science (CS), or say, most professors working on data mining problems have a PhD in CS.
- Data mining is "engineering" by nature: as long as an algorithm works for its purpose, then it is a good algorithm and we don't need to know why it works.
- Studying why a method or an algorithm works is closer to a **statistical problem** and thus there are some statistics professors working on theoretical problems of data mining or machine learning.
- There also exists a parallel world of statistics/econometrics methods for the same or similar business problems in business school.

Regression vs Data Mining Algorithms

- Regression models are hypothesis-based methods:
  - Formulate a hypothesis of interest => Does X cause Y?
  - Use various regression methods to estimate the beta coefficient of X.
  - Accept or reject hypothesis depending on whether the estimated beta coefficient is different from zero or not.
  - "Causality" and the magnitude of effect are the objectives.
- Data mining is an exploratory-based method:
  - Try to make sense of a bunch of data without an a priori hypothesis!
  - The key assumption is the pattern will repeat itself.
  - The methods usually do not care about causality and powerful methods are black-box methods without interpretable relationship.
  - "Accuracy" of the predictive model is the objective.

Regression versus Data mining Algorithms

- For example,
  - Y is the dependent variable (DV),
  - X1 and X2 are the independent variables (IV).

- In regression approach, YOU assume a function Y=f(X1, X2), where f(X1, X2) is a linear function in linear regression.
  - Y= b0 + b1*X1 + b2*X2 + error
  - The goal is the estimated value of beta coefficients: b0, b1, and b2.
- In data mining algorithms, DATA tells you the best function f(X1, X2) that predicts Y.
  - f(X1, X2) can be a very complicated function and is determined by various kinds of novel ideas of algorithms.
  - In most data mining algorithms, we do not even know the details of f(X1, X2).

Recap: Traditional Regression Methods

- Linear Regression for continuous dependent variable (DV)
- Logistic Regression or Probit Regression for categorical DV
- Time series model (ARIMA, GRACH)
- Panel Regression for panel data (fixed effect and random effect)
- Poisson regression model for "counts" DV
- Duration model for survival analysis
- Estimation methods: least square estimation, maximum likelihood estimation, and General Moment Methods.
- Non parametric regressions

Modern Econometrics for Causality

Modern econometrics focused on how to establish causality between X and Y without conducting randomized experiments ⇔ theoretically the best method. Three main methods:

1. Instrumental variable regression
2. Differences-in-Differences
3. Regression Discontinuity

One related subject

4. Propensity score matching

Accuracy or Causality?

- Which approach is better depends on how the decision maker uses the results.
- Accuracy is more important when the decision-making is based on the predicted value of Y.
  - Fraud "detection" ⇔ we try to identify a fraudulent case based on the predicted value of Y=f(X).
  - Predicting other countries macroeconomic statistics ⇔ we care about "Y" and we cannot manipulate Y by adjusting X, which is some kind of policy.
- Causality is more important when we try to adjust X to affect Y.
  - Fraud "prevention" ⇔ we try to prevent a fraudulent case by changing the values of X.
  - Predicting Singapore's macroeconomic statistics (GDP, unemployment, or inflation) and the goal is to optimize Y by changing the values of X (gov budget, exchange rate, interest rate, …etc.).

Example: Accuracy or causality?

- In Search Engine Optimization (how Google ranks a web page), the most important factor is the number of quality URL links pointing to your site.
- However, the correlation between the number of social shares and Google ranking is higher than the number of backlinks (Why?)
- Given this empirical fact, if you create **fake** Google+ shares, you may not be able to increase ranking a lot and will waste efforts.
- If you create natural Google plus shares, it may be helpful to some extent. But the true most impactful factor is still the number of backlinks.

Example: Accuracy or causality?

- In marketing, you may need to evaluate the effects of an advertisement or promotion event (X) on profit or sales (Y). This is also more about causality than predicting Y alone. This kind of analysis may also be more suitable for traditional econometrics than data mining.
- In finance, many people are interested in predicting stock prices. Prediction accuracy could be more important. But without knowing causality, pattern may not repeat itself.
- In healthcare, the effect of any drug or medical device must go through very rigorous experiments several times to establish causality.

Example: Accuracy or causality?

- **"Books in the Home Are Strongly Linked to kids' Academic Achievement" even after controlling** parents' occupations and education level and family wealth. the effect was consistently found in all kinds of countries…"

- **Should SG gov. just buy books and give it to all family, or simply force all family to have books at home for building a smart nation?**

- **What if we use this as a criteria for selecting kids into schools?**

5: Terminologies of Data Mining

First, let's define and clarify key terminologies that will be used throughout the semester, especially if this is your first class in data mining.

1. Unit of analytics
2. Labels = Class in classification algorithms
3. Features = Attributes
4. Training set and Test set
5. Multi-class classification vs Multi-label classification

Unit of Analytics

- The phrase **unit of observation** is used to describe the smallest entity with measured properties of interest for a study. For example, one person? One team? One department? One firm? One industry? One city? One USA state? One country?
  - Deciding the unit of analysis is an important step in open-ended data analytics projects.
  - Aggregation at which level is one research topic.
  - There exist various hierarchical methods.
- **In CS books and papers, we use "Examples" to mean**: Instances of the unit of observation for which properties have been recorded
  - Also called "tuples" = "records" = "observations" = "samples" in references from various disciplines.

Key Terminologies

- **Labels**: This means the categorical "dependent variable" for data mining algorithms to predict.
- **Features**: properties or attributes of examples that may be useful for learning
  - Also called, = dimensions (CS) = attributes (CS) = independent variables (Stat/BIZ) = explanatory variables (Stat/BIZ) = predictors (Stat/BIZ) = "factors" in other disciplines.
- **Training Set**: The set of examples for building a predictive model in supervised learning.
- **Test Set**: The set of examples for evaluating prediction accuracy.

Features

- Features can have many data types: numeric, categorical, binary, or ordinal (L/M/S), or fancy ones like text, audio, video, geo-location, or social network connections.

features

year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

examples

Class = Labels

- In classification, the target feature to be predicted is a categorical feature and is also known as the **class**, and is divided into categories called **levels**.
- A class can have two or more levels, and the levels may or may not be ordinal.
  - Non-ordinal: Gender => male vs female
  - Ordinal: small, medium, large
- Standard classification is also called **multiclass** classification. This is different from multi-label classification in which each record may belong to multiple types.

6: Overview of Lazy Learning

- One of the most intuitive algorithm for prediction.

Underlying assumption and motivation of this method are

"Birds of a feather flock together"

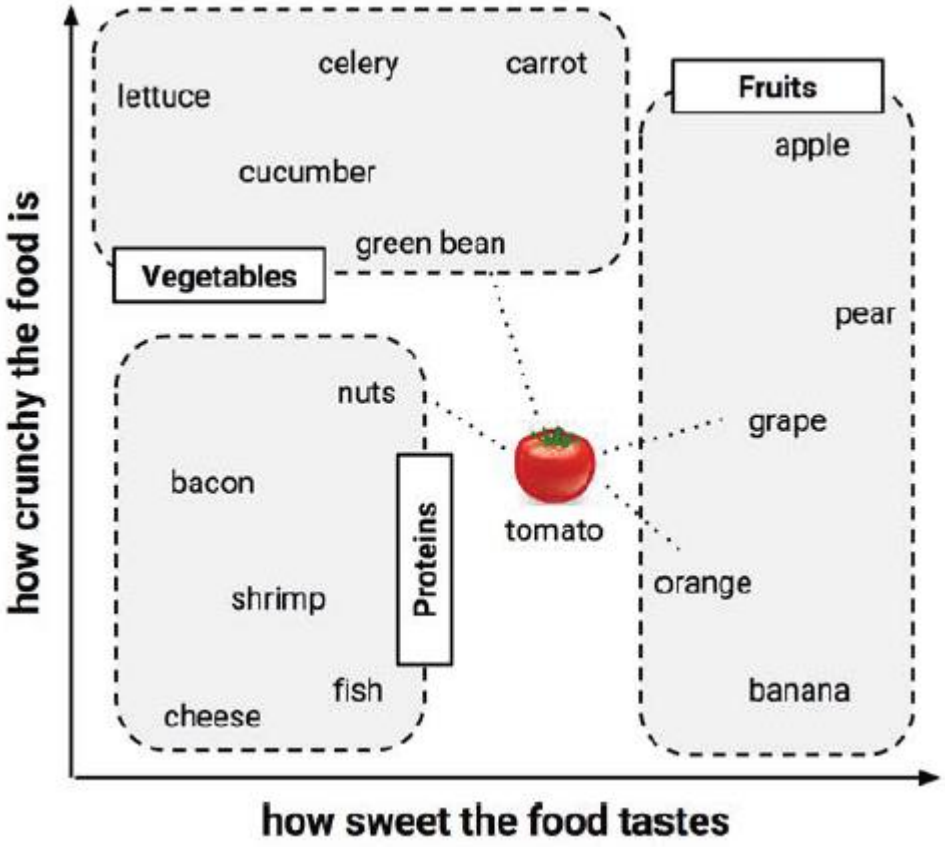
1. We want to classify an example X.
2. We find K most similar records in our training set (K Nearest Neighbor algorithm).
3. We use the labels of the K most similar records to predict the label of X.

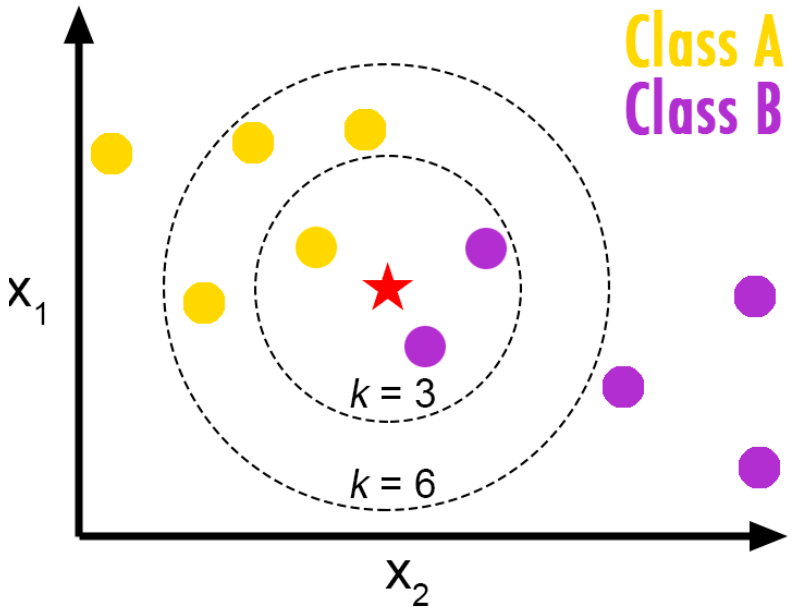
Applications of Lazy Learning

Good applications are those cases in which "Similarity of X" strongly implies similarity of Y and also you have a **dense enough** training dataset.

- Computer vision applications, including optical character recognition and facial recognition in both still images and video
- In marketing, we predict the preference of a customer by similar customers.
  - Predicting whether a person will enjoy a movie or music recommendation

Illustrations of the KNN Idea





Strengths and Weakness

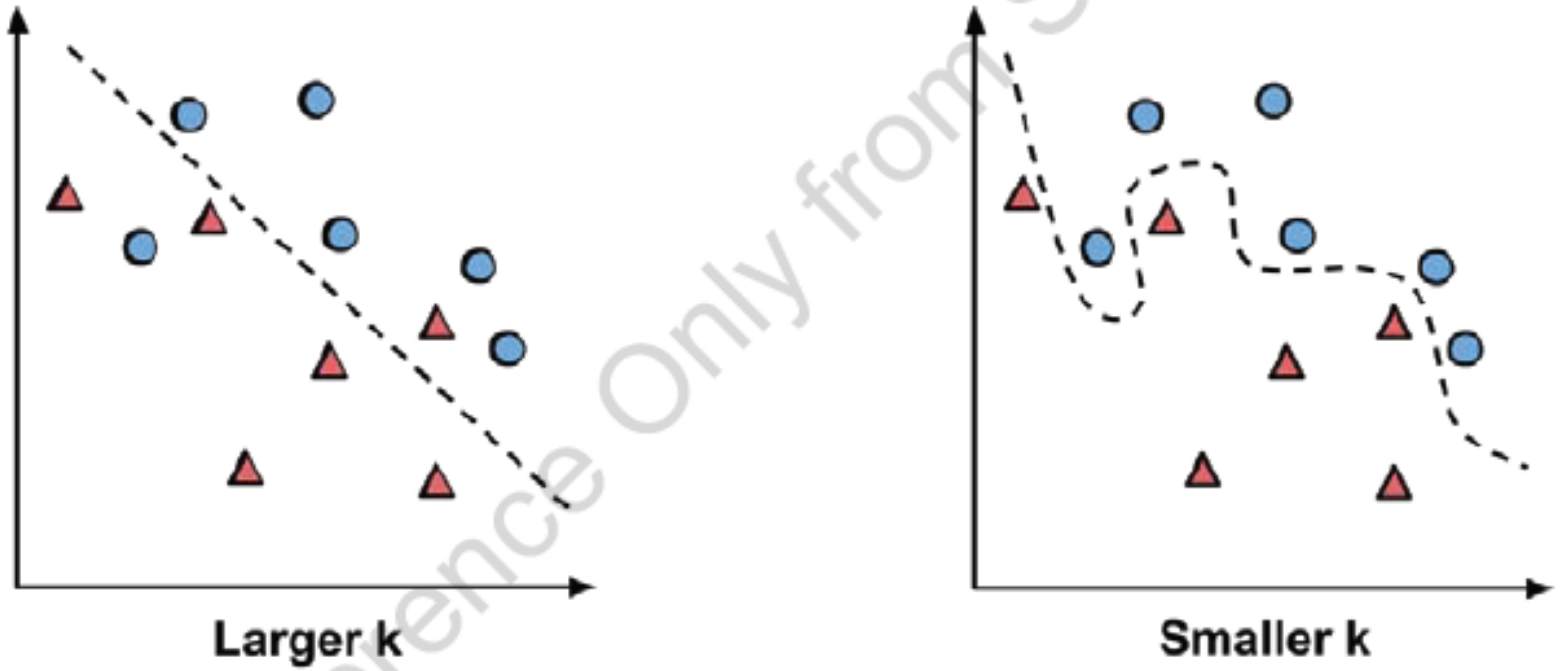
Strengths	Weaknesses
• Simple and effective • Makes no assumptions about the underlying data distribution • Fast training phase	• Does not produce a model, limiting the ability to understand how the features are related to the class • Requires selection of an appropriate k • Slow classification phase • Nominal features and missing data require additional processing

- "Fast training": no training phase ⇔ no model
- Although simple, KNN has its benefits and is a simple solution in practice for a wide variety of problems.

Technical Challenges of KNN

1. What is the distance function?
   - Euclidean distance is a straightforward and most common choice.
   - Other distance functions are fine.
2. How to decide K?
   - You have to try different values and compare the performance.
   - Too few: unstable.
   - Too many: you use the non-similar cases for prediction
3. Numerical attributes need to be standardized.
4. Categorical attributes need to be converted to dummy variables (binary variables for indicating each category = oneHot in CS.)

Larger/Smaller K ⇔ Under/Overfitting



- A small K implies a complicated decision boundary whereas a larger K implies a smooth boundary.
- We do not know which boundary better represents the unobservable true boundary.
- But the complicated boundary may be more likely to be an overfitting case and the linear boundary is more likely to be an under-fitting case.

Standardization of Numeric Features

There are two commonly used standardization formula.

1. Min-Max normalization (0 to 1)

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

2. Z-score standardization (same as in statistics, roughly from -3 to +3)

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - \text{Mean}(X)}{\text{StdDev}(X)}$$

Transforming Nominal Features

- The coding of nominal features is straightforward if you are familiar with dummy variables in statistics/regression
- For example, for a feature with 3 (n) categories (hot, medium, or cold), we create 2 (=n-1) dummy variables.

$$\text{hot} = \begin{cases} 1 & \text{if x = hot} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{medium} = \begin{cases} 1 & \text{if x = medium} \\ 0 & \text{otherwise} \end{cases}$$

- In data mining, the tradition is to use OneHot packages that will create a binary variable for each level of a categorical variable, different from regression
- Since the range is 0 to 1, this will be consistent with the min-max normalization.