Data Science

Week 1
Introduction

Instructor Introduction

Course Instructor

Name: Yanlei GU

Email: guyanlei@fc.ritsumei.ac.jp

Courses: Systems Biology- Exercises,
Engineering Mathematics 2, Computer Architecture,
Applied informatics (Deep learning),
Embedded system, **Data Science**

Student Attainment Objectives

At the conclusion of the course, students should be able to:

- Understand data science project lifecycle
- To carry out basic modeling and analysis (Using Python)
- Understand and apply basic machine learning and data analysis algorithms
- Visualize data and perform exploratory data analysis
- Apply the learned knowledge for a data science project

Course Schedule

Week 1: Real-World Problems and Data Science Solutions
Week 2: Getting Start with Data Manipulation
Week 3: Exploratory Data Analytics (1) - Statistics
Week 4: Exploratory Data Analytics (1) – Visualization
Week 5: Test 1 (tentative) and Introduction to Predictive Modeling
Week 6: Fitting a Model to Data (1) - Classification
Week 7: Fitting a Model to Data (2) - Neural Networks
Week 8: Course Review and Q&A
Week 9: Fitting a Model to Data (3) - Regression
Week 10: Test 2 (tentative) and Fitting a Model to Data (4) - Clustering
Week 11: Decision Analytic Thinking: What Is a Good Model?
Week 12: Visualizing Data and Model Performance
Week 13: Representing and Mining Text
Week 14: Project Example of Data Science
Week 15: Course Summary and Final test

Grade Evaluation Method

- Attendance and activity in class
- Continuous assessment of assignments
- In-class tests

Important Note

- In the case of BCP level 1-2:
  - -The Number of face-to-face class sessions: 6
  - -The Number of web-based class sessions: 9
- In the case of BCP level 3-4:
  - -All the classes will be the web-based class sessions

- We plan to use Zoom to conduct the web-based class.

- We are planning to provide live-stream by using Zoom (same link) in face-to-face classes.

- If there are any changes, we will announce on Manaba +R

Important Note

- Consultations.

Office Hours: By appointment, e-mail: guyanlei@fc.ritsumei.ac.jp.
Note: Contact me if you are having any difficulties with the material. The sooner the better.

- Attendance.

Students are responsible for all material covered in this class.
Students who miss more than 5 classes without a legitimate reason will automatically receive an "F" for the course.

- Professional ethics.

The behavioral and ethical standards of Ritsumeikan University will be observed in all aspects of this course. Specifically, academic dishonesty (e.g. copying assignments or the like) will result in a grade F for the corresponding assignment, and in many cases - in a failing grade (F) for the course.

Other requirements

- **Device**

In class, students need to use your own PC to program.

- **Programming Language**

We will mainly use Python for programming.

Outline

- Introduction for Data Science
- From Real World Problems to Data Mining Tasks
- Fundamentally Different Types of Tasks
- Supervised Versus Unsupervised Methods
- Data Mining and Its Results

What is Data Science?

- Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data (textual data).

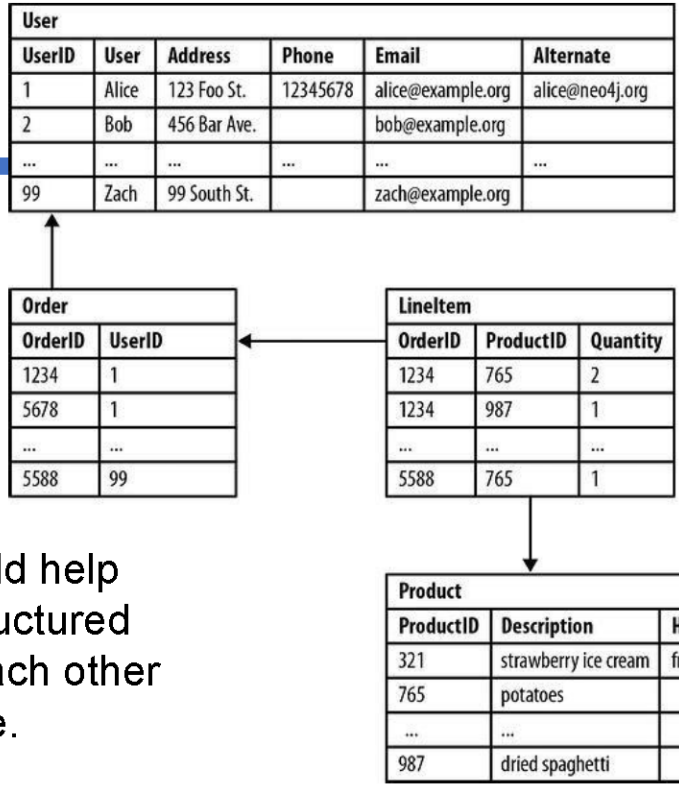
- Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems".

From Wikipedia, Data science, https://en.wikipedia.org/wiki/Data_science

Structured Data

- Structured data is most often categorized as quantitative data, and it's the type of data most of us are used to working with. Think of data that fits neatly within fixed fields and columns in relational databases and spreadsheets.
- Examples of structured data include names, dates, addresses, credit card numbers, stock information, geolocation, and more.
- Structured data is highly organized and easily understood by machine language. Those working within relational databases can input, search, and manipulate structured data relatively quickly. This is the most attractive feature of structured data.

Visualization of Structured Data



User

UserID	User	Address	Phone	Email	Alternate
1	Alice	123 Foo St.	12345678	alice@example.org	alice@neo4j.org
2	Bob	456 Bar Ave.		bob@example.org	
...	...	...	...	...	...
99	Zach	99 South St.		zach@example.org	

Order

OrderID	UserID
1234	1
5678	1
...	...
5588	99

LineItem

OrderID	ProductID	Quantity
1234	765	2
1234	987	1
...	...	...
5588	765	1

Product

ProductID	Description	Handling
321	strawberry ice cream	freezer
765	potatoes	
...	...	
987	dried spaghetti	

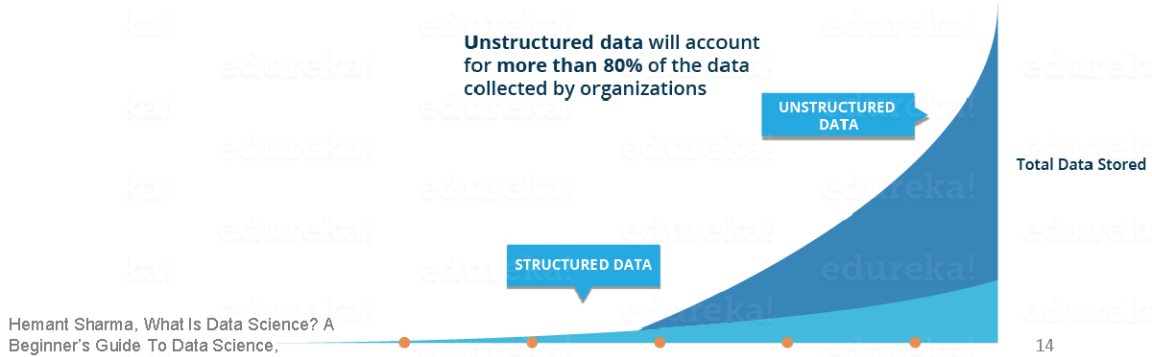
The picture should help visualize how structured data relates to each other within a database.

Unstructured Data

- Unstructured data is most often categorized as qualitative data, and it cannot be processed and analyzed using conventional tools and methods.
- Examples of unstructured data include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery – the list goes on and on.
- Unstructured data is difficult to deconstruct because it has no pre-defined model, meaning it cannot be organized in relational databases.

Why Data Science?

- Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured.
- The data trends in the figure given below shows that by 2020, more than 80 % of the data are unstructured.



Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Why Data Science?

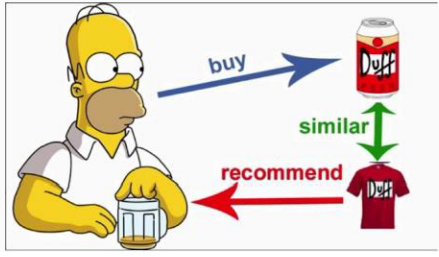
- This data is generated from different sources like financial logs, text files, multimedia forms, sensors, and instruments.
- Simple off-the-shelf tools are not capable of processing this huge volume and variety of data.
- This is why we need to learn Data Science to create more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of data for our own project.

Data Science in our Daily Life

- Let's take weather forecasting as an example. Data from ships, aircraft, radars, satellites can be collected and analyzed to build models.
- These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help you to take appropriate measures beforehand and save many precious lives.

Data Science in our Daily Life

- How about if you could understand the precise requirements of your customers from the existing data like the customer's past browsing history, purchase history, age and income.
- No doubt you had all this data earlier too, but now with the vast amount and variety of data, you can train models more effectively and recommend the product to your customers with more precision.



https://towardsdatascience.com/build-your-own-recommender-system-within-5-minutes-30dd40388fbf

Data Science in our Daily Life

- How about if your car had the intelligence to drive you home? The self-driving cars collect data from sensors, including radars, cameras, and lasers to create a map of its surroundings.
- Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn – making use of advanced machine learning algorithms.

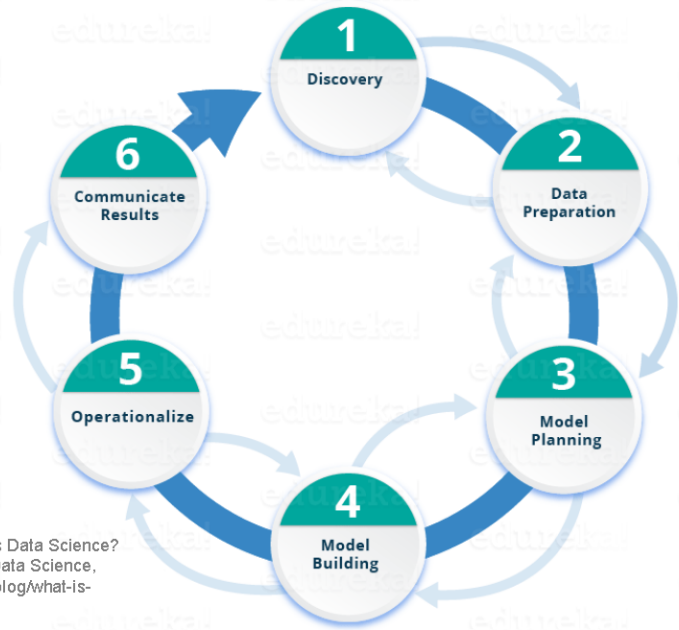
Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

What does a Data Scientist do?

- Data scientists are those who crack complex data problems with their expertise in certain scientific disciplines.
- They work with several elements related to mathematics, statistics, computer science, etc.
- They make a lot of use of the latest technologies in finding solutions and reaching conclusions that are crucial for an organization's growth and development.
- Data Scientists present the data in a much more useful form as compared to the raw data available to them from structured as well as unstructured forms.

Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Lifecycle of Data Science



Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Lifecycle of Data Science

Phase 1—Discovery:

- Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget.
- Here, you assess if you have the required resources present in terms of technology and data.
- In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.

Phase 2—Data preparation:

- You need to explore and preprocess data prior to modeling.
- The properties and defects of the data should be discovered.

Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Lifecycle of Data Science

Phase 3—Model planning:

- Here, you will determine the methods and techniques to draw the relationships between variables.
- These relationships will set the base for the algorithms which you will implement in the next phase.
- You will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

Phase 4—Model building:

- In this phase, you will develop datasets for training and testing purposes.
- Here you need to consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing).
- You will analyze various learning techniques like classification, association and clustering to build the model.

Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Lifecycle of Data Science

Phase 5—Operationalize:

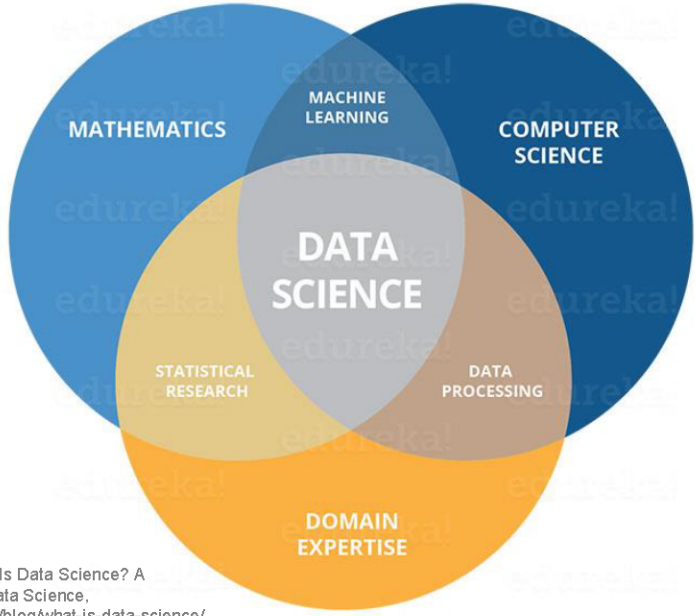
- we will run a small pilot project to check if our results are appropriate. We will also look for performance constraints if any. If the results are not accurate, then we need to replan and rebuild the model.
- In addition, sometimes a pilot project is also implemented in a real-time production environment.

Phase 6—Communicate results:

- Now it is important to evaluate if you have been able to achieve your goal that you had planned in the first phase.
- So, in the last phase, you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.

Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Skills of Data Scientist



Hemant Sharma, What Is Data Science? A Beginner's Guide To Data Science, https://www.edureka.co/blog/what-is-data-science/

Outline

- Introduction for Data Science
- From Real World Problems to Data Mining Tasks
- Fundamentally Different Types of Tasks
- Supervised Versus Unsupervised Methods
- Data Mining and Its Results

From Real World Problems to Data Mining Tasks

- Each data-driven decision-making problem is unique, comprising its own combination of goals, desires, and constraints.
- A critical skill in data science is the ability to decompose a real world problem into pieces such that each piece matches a known task for which tools are available.
- Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel.

Example: Predicting Customer Churn

- Assume you just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms.
- They are having a major problem with customer **retention** in their wireless business.
- 20% of cell phone customers leave when their contracts expire, and it is getting increasingly difficult to acquire new customers. Customers switching from one company to another is called *churn*.
- You have been called in to help understand the problem and to devise a solution. Attracting new customers is much more expensive than retaining existing ones, so a good deal of marketing budget is allocated to prevent churn. Marketing has already designed a special retention offer.

Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

Converting “Predicting Customer Churn” Problem to Data Mining Task

- For example, our telecommunications churn problem is unique to MegaTelCo: there are specifics of the problem that are different from churn problems of any other telecommunications firm.
- However, a subtask that will likely be part of the solution to any churn problem is to estimate from historical data the probability of a customer terminating her contract shortly after it has expired.
- Once the MegaTelCo data have been assembled into a particular format, this probability estimation fits the mold of one very common data mining task.

Outline

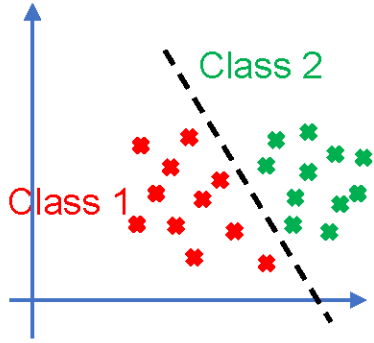
- From Real World Problems to Data Mining Tasks
- Fundamentally Different Types of Tasks
- Supervised Versus Unsupervised Methods
- Data Mining and Its Results

Fundamentally Different Types of Tasks

- Despite the large number of specific data mining algorithms developed over the years, there are only a handful of fundamentally different types of tasks these algorithms address.
  - ✓ Classification
  - ✓ Regression
  - ✓ Clustering
  - ✓ Similarity matching
  - ✓ Co-occurrence grouping
  - ✓ Profiling
  - ✓ Link prediction
  - ✓ Data reduction
  - ✓ Causal modeling

Classification -1

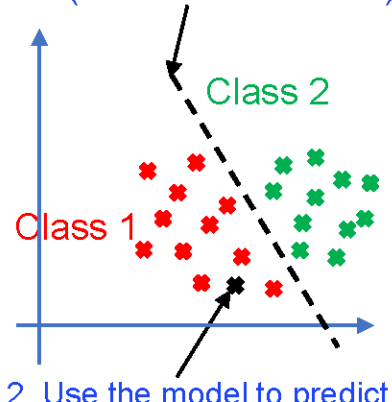
- Classification and class probability estimation attempt to predict, for each individual in a population, which of a (small) set of classes this individual belongs to.
- Usually the classes are mutually exclusive. An example classification question would be:
- "Among all the customers of MegaTelCo, which are likely to respond to a given offer?" In this example the two classes could be called will respond and will not respond.



Classification -2

- For a classification task, a data mining procedure produces a model that, given a new individual, determines which class that individual belongs to.
- A closely related task is *scoring* or class *probability estimation*. A scoring model applied to an individual produces, instead of a class prediction, a score representing the probability (or some other quantification of likelihood) that that individual belongs to each class.
- E.g. In our customer response scenario, a scoring model would be able to evaluate each individual customer and produce a score of how likely each is to respond to the offer.

1. Get the Model (Criterion for classification) from the data (of class 1 and class 2)

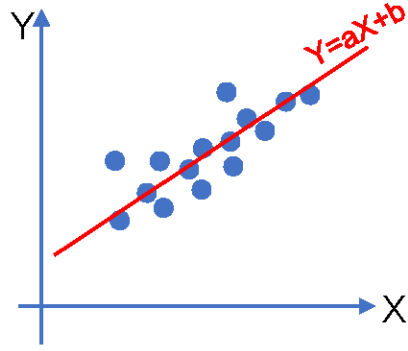


2. Use the model to predict the label for the new individual: class 1 or class 2

Regression -1

- Regression ("value estimation") attempts to estimate or predict, for each individual, the numerical value of some variable for that individual.
- An example regression question would be: "How much will a given customer use the service?"
- The property (variable) to be predicted here is *service usage*, and a model could be generated by looking at other, similar individuals in the population and their historical usage.

Regression: Get the model (red line) based on blue points



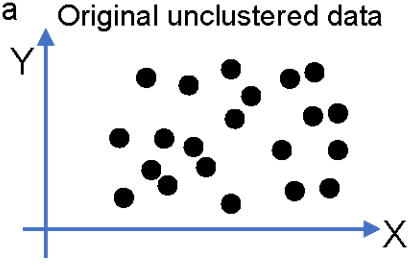
Use the model (Y=aX+b) to estimate Y for a given new X (real Y is unknown)

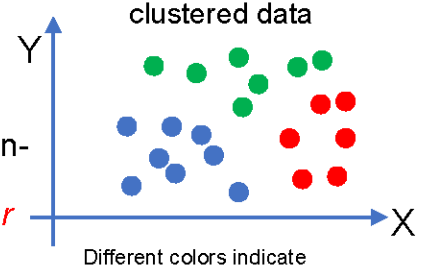
Regression -2

- A regression procedure produces a model that, given an individual, estimates the value of the particular variable (e.g. Y) specific to that individual (e.g. X).
- Regression is related to classification, but the two are different.
- Informally, classification predicts *whether* something will happen, whereas regression predicts *how much* something will happen. The difference will become clearer as the class progresses.

Clustering

- Clustering attempts to *group* individuals in a population together by their similarity, but not driven by any specific purpose.
- An example clustering question would be: "Do our customers form natural groups or segments?"
- Clustering is useful in preliminary domain exploration to see which natural groups exist because these groups in turn may suggest other data mining tasks or approaches.
- Clustering also is used as input to decision-making processes focusing on questions such as: *What products should we offer or develop? How should our customer care teams (or sales teams) be structured?*





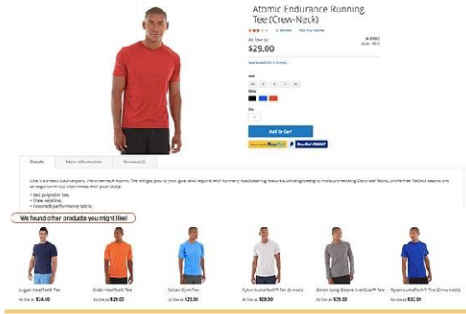
Different colors indicate different clusters

Similarity Matching -1

- Similarity matching attempts to *identify* similar individuals based on data known about them. Similarity matching can be used directly to find similar entities.
- For example, IBM is interested in finding companies similar to their best business customers, in order to focus their sales force on the best opportunities. They use similarity matching based on "firmographic" data describing characteristics of the companies.

Similarity Matching -2

- Similarity matching is the basis for one of the most popular methods for making product recommendations (finding people who are similar to you in terms of the products they have liked or have purchased).
- Similarity measures underlie certain solutions to other data mining tasks, such as classification, regression, and clustering.



Co-occurrence Grouping -1

- Co-occurrence grouping (also known as frequent itemset mining, association rule discovery, and market-basket analysis) attempts to find *associations* between entities based on transactions involving them.
- An example co-occurrence question would be: *What items are commonly purchased together?*
- While clustering looks at similarity between objects based on the objects' attributes, co-occurrence grouping considers similarity of objects based on their appearing together in transactions.

Co-occurrence Grouping -2

- Co-occurrence of products in purchases is a common type of grouping known as market-basket analysis. Some *recommendation* systems also perform a type of affinity grouping by finding, for example, pairs of books that are purchased frequently by the same people ("people who bought X also bought Y").
- The result of co-occurrence grouping is a description of items that occur together. These descriptions usually include statistics on the frequency of the co-occurrence.

What books are commonly purchased together?

Profiling -1

- Profiling (also known as behavior description) attempts to characterize the typical behavior of an individual, group, or population.
- An example profiling question would be: “What is the typical cell phone usage of this customer segment?”
- Behavior may not have a simple description; profiling cell phone usage might require a complex description of night and weekend airtime averages, international usage, text minutes, and so on. Behavior can be described generally over an entire population, or down to the level of small groups or even individuals.

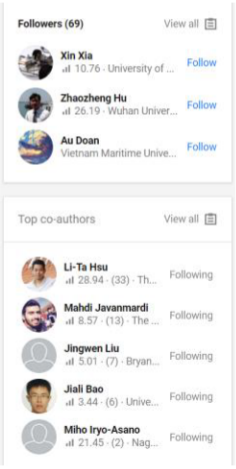
Profiling -2

- Profiling is often used to establish behavioral norms for anomaly detection applications such as fraud detection and monitoring for intrusions to computer systems (such as someone breaking into your iTunes account).
- For example, if we know what kind of purchases a person typically makes on a credit card, we can determine whether a new charge on the card fits that profile or not. We can use the degree of mismatch as a suspicion score and issue an alarm if it is too high.

Link Prediction

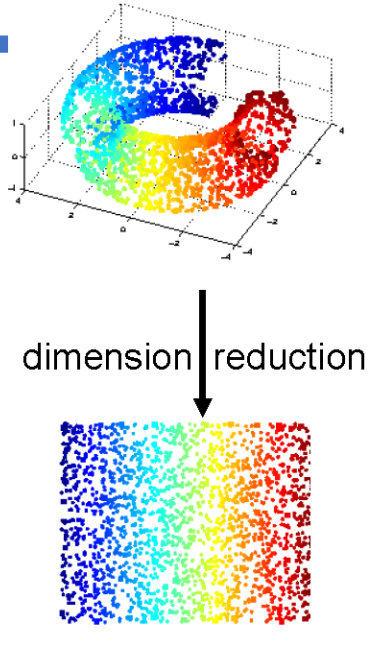
- Link prediction attempts to predict connections between data items, usually by suggesting that a link should exist, and possibly also estimating the strength of the link.
- Link prediction is common in social networking systems: “Since you and Wang share 10 friends, maybe you’d like to be Wang’s friend?” Link prediction can also estimate the strength of a link.

Friend Prediction in Research Gate



Data reduction

- Data reduction attempts to take a large set of data and replace it with a smaller set of data that contains much of the important information in the larger set. The smaller dataset may be easier to deal with or to process. Moreover, the smaller dataset may better reveal the information.
- For example, a massive dataset on consumer movie-viewing preferences may be reduced to a much smaller dataset revealing the consumer taste preferences that are latent in the viewing data.



Causal Modeling -1

- Causal modeling attempts to help us understand what events or actions actually *influence* others.
- For example, consider that we use predictive modeling to target advertisements to consumers, and we observe that indeed the targeted consumers purchase at a higher rate subsequent to having been targeted. Was this because the advertisements influenced the consumers to purchase? Or did the predictive models simply do a good job of identifying those consumers who would have purchased anyway?

Causal modeling -2

- Techniques for causal modeling include those involving a substantial investment in data, such as randomized controlled experiments (e.g., so-called “A/B tests”), as well as sophisticated methods for drawing causal conclusions from observational data.
- Both experimental and observational methods for causal modeling generally can be viewed as “counterfactual” analysis: they attempt to understand what would be the difference between the situations—which cannot both happen —where the “treatment” event (e.g., showing an advertisement to a particular individual) were to happen, and were not to happen.

Outline

- From Real World Problems to Data Mining Tasks
- Fundamentally Different Types of Tasks
- Supervised Versus Unsupervised Methods
- Data Mining and Its Results

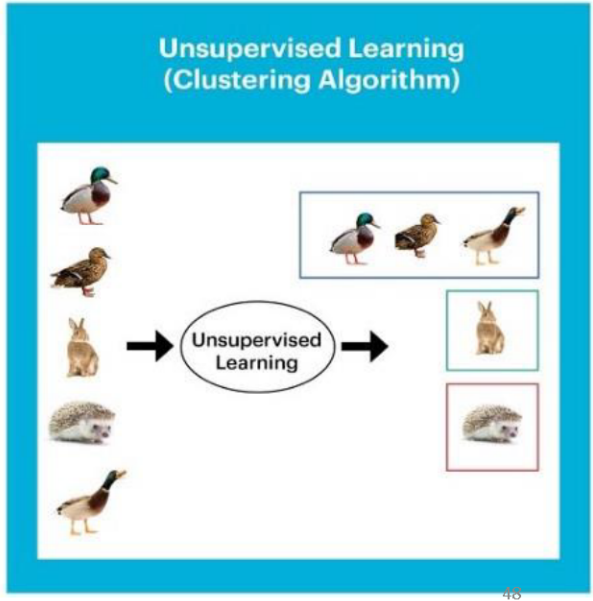
Supervised Versus Unsupervised Methods

- Consider which of these types of tasks might fit our churn-prediction problem. Often, practitioners formulate churn prediction as a problem of finding segments of customers who are more or less likely to leave.
- This segmentation problem sounds like a classification problem, or possibly clustering, or even regression. To decide the best formulation, we first need to introduce some important distinctions.

→Supervised Versus Unsupervised Methods

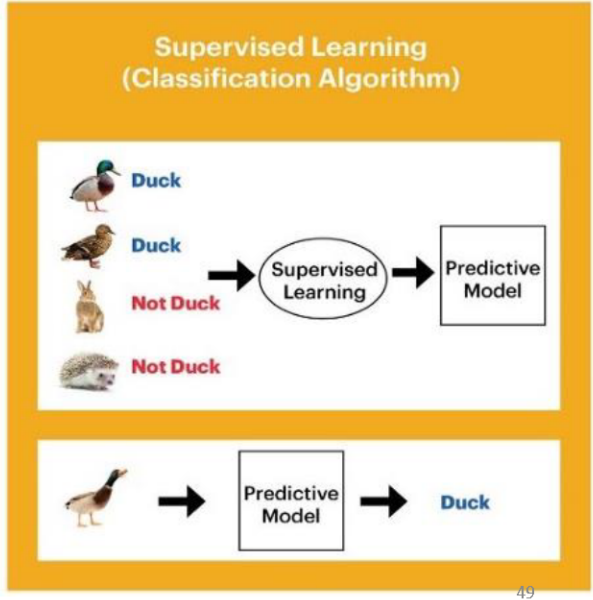
No Specific Purpose or Target: Unsupervised

- Consider two similar questions we might ask about a customer population.
- The first is: "Do our customers naturally fall into different groups?" Here no specific purpose or *target* has been specified for the grouping.
- When there is no such target, the data mining problem is referred to as *unsupervised*.



With Specific Purpose or Target: supervised

- Contrast this with a slightly different question: "Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contracts expire?"
- Here there is a specific target defined: will a customer leave when her contract expires?
- In this case, segmentation is being done for a specific reason: to take action based on likelihood of churn. This is called a supervised data mining problem.



Label for Supervised

- Technically, another condition must be met for supervised data mining: there must be *data* on the target. It is not enough that the target information exist in principle; it must also exist in the data.
- For example, it might be useful to know whether a given customer will stay for at least six months, but if in historical data this retention information is missing or incomplete (if, say, the data are only retained for two months) the target values cannot be provided. Acquiring data on the target often is a key data science investment.
- The value for the target variable for an individual is often called the individual's *label*, emphasizing that often (not always) one must incur expense to actively label the data.

Examples of Supervised and Unsupervised Methods

- Classification, regression, and causal modeling generally are solved with supervised methods.
- Similarity matching, link prediction, and data reduction could be either.
- Clustering, co-occurrence grouping, and profiling generally are unsupervised.

Outline

- From Real World Problems to Data Mining Tasks
- Fundamentally Different Types of Tasks
- Supervised Versus Unsupervised Methods
- Data Mining and Its Results

Data Mining and Its Results

- There is another important distinction pertaining to mining data: the difference between

(1) mining the data to find patterns and build models → Training phase, and (2) *using* the results of data mining → Test phase.

- Students often confuse these two processes when studying data science.
- The use of data mining results should influence and inform the data mining process itself, but the two should be kept distinct.

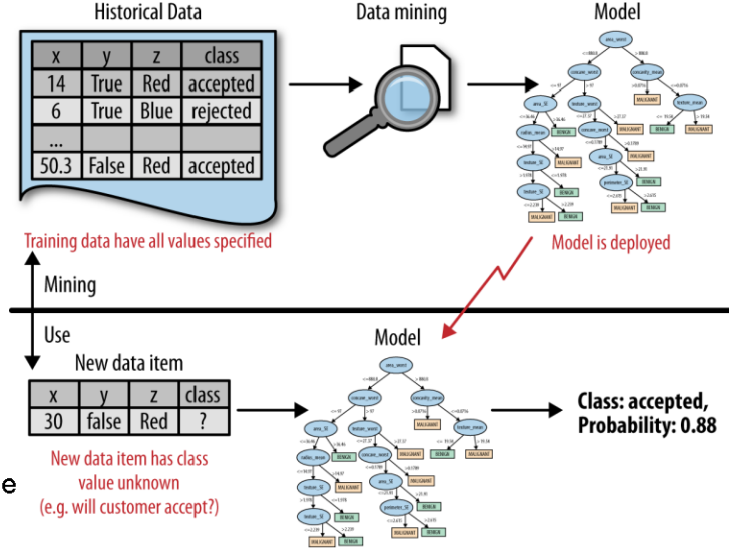
Produce Model in Data Mining (training), then Use Model (test)

- In our churn example, consider the deployment scenario in which the results will be used. We want to use the model to predict which of our customers will leave.
- Specifically, assume that data mining has created a class probability estimation model *M*. Given each existing customer, described using a set of characteristics, *M* takes these characteristics as input and produces a score or probability estimate. This is the *use* of the results of data mining.
- The data mining produces the model *M* from some other, often historical, data.

Example of Data Mining and Its Results

The upper half of the figure illustrates the mining of historical data to produce a model. Importantly, the historical data have the target ("class") value specified.

The bottom half shows the result of the data mining in use, where the model is applied to new data for which we do not know the class value. The model predicts both the class value and the probability that the class variable will take on that value.



Data mining versus the use of data mining results.

Summary

- Introduction
- From Real World Problems to Data Mining Tasks
- Fundamentally Different Types of Tasks
- Supervised Versus Unsupervised Methods
- Data Mining and Its Results

Reference

- Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., 2013.

Data Science

Week 2

Getting Start with Data Manipulation