[Lecture Note] Overview of the Data Mining Process

MGS3701: Data Mining

Chungil Chae

Wed, 26 February 2025

In this chapter, we give an overview of the steps involved in data mining, starting from a clear goal definition and ending with model deployment. The general steps are shown schematically in Figure 2.1. We also discuss issues related to data collection, cleaning, and preprocessing. We introduce the notion of data partitioning, where methods are trained on a set of training data and then their performance is evaluated on a separate set of validation data, as well as explain how this practice helps avoid overfitting. Finally, we illustrate the steps of model building by applying them to data.(Shmueli et al., 2017)

https://www.youtube.com/watch?v=7rs0i-9nOjo

Introduction

- In this chapter, we introduce the variety of methods sometimes referred to as data mining.
- Data mining focuses on what has come to be called predictive analytics, the tasks of
  - *Classification*
  - *Prediction*
  - *Pattern discovery*

Core Idea in Data Mining

Classification

- Classification is the most basic form of data analysis.
  - e.g., application of a loan, credit card transaction
- To examine data where the classification is
  - unknown or
  - will occur in the future
- Similar to classification is develop rules

Prediction

- Prediction is to predict **the value of a numerical variable** (e.g., amount of purchase) rather than a class (e.g.,purchaser or nonpurchaser).
- the value of a continuous variable.

Assiciation Rules and Recommendation Systems

- **What goes with what**
- Association rules, or affinity analysis, is designed to find such **general associations patterns** between items in large databases.
  - grocery stores: product placement, weekly promotional offers, bundling products.
  - hospital database: which symptom is followed by what other symptom to help predict future symptoms for returning patients.
  - Online recommendation systems: collaborative filtering

> **Collaborative filtering**
>
> Collaborative filteringa method that uses individual users' preferences and tastes given their historic purchase, rating, browsing, or any other measurable behavior indicative of preference, as well as other users' history.

> **Association rules vs Collaborative filtering**
>
> In contrast to association rules that generate rules general to an entire population, collaborative filtering generates "what goes with what" at the individual user level. Hence, collaborative filtering is used in many recommendation systems that aim to deliver personalized recommendations to users with a wide range of preferences.

Predictive Analytics

- Classification, prediction, and to some extent, association rules and collaborative filtering constitute the analytical methods employed in predictive analytics.
- The term predictive analytics is sometimes used to also include data pattern identification methods such as clustering.

Data Reduction and Dimension Reduction

- Data mining algorithms is often improved
  - when the number of variables is limited, and
  - when large numbers of records can be grouped into homogeneous groups.

- For example,
  - rather than dealing with thousands of product types, an analyst might wish to group them into a smaller number of groups and build separate models for each group.
  - Or a marketer might want to classify customers into different "personas," and must therefore group customers into homogeneous groups to define the personas.
- This process of consolidating a large number of records (or cases) into a smaller set is termed data reduction. Methods for reducing the number of cases are often called clustering.
- Reducing the number of variables is typically called dimension reduction.
  - Dimension reduction is a common initial step before deploying data mining methods, intended to improve predictive power, manageability, and inter-pretability.

Data Exploration and Visualization

- Exploration is aimed at
  - understanding the global landscape of the data, and
  - detecting unusual values.
- Exploration is used for **data cleaning** and **manipulation** as well as for **visual discovery** and **hypothesis generation**.
- Methods for exploring data include looking at various data aggregations and summaries
- (Both numerically and graphically) looking at each variable separately as well as looking at relationships among variables.
- The purpose is to **discover patterns** and **exceptions**.
- Data Visualization or Visual Analytics - Exploration by creating charts and dashboards
- For numerical variables,
  - we use histograms and boxplots to learn about the distribution of their values,
    - to detect outliers (extreme observations), and to
    - find other information that is relevant to the analysis task.
- For categorical variables,
  - we use bar charts. We can also look at scatter plots of pairs of numerical variables
    - to learn about possible relationships,
    - the type of relationship,
    - to detect outliers.
- Visualization can be greatly enhanced by adding features such as color and interactive navigation.

Supervised and Unsupervised Learning

- A fundamental distinction among data mining techniques is between supervised and unsupervised methods.

Supervised Learning Algorithms

- Supervised learning algorithms are those used in **classification** and **prediction**.
  - We must have data available in which the value of the outcome of interest (e.g., purchase or no purchase) is known, "labeled data"
  - These training data are the data from which the classification or prediction algorithm "learns," or is "trained," about the relationship between predictor variables and the outcome variable.
  - Once the algorithm has learned from the training data, it is then applied to **another sample of labeled data (the validation data)** where the *outcome is known but initially hidden*, to see how well it does in comparison to other models.
  - If many different models are being tried out, it is prudent to save a third sample, which also includes known outcomes (the test data) to use with the model finally selected to predict how well it will do.
  - The model can then be used to classify or predict the outcome of interest in new cases where the outcome is unknown.

> **Linear Regression as Supervised Machine Learning Algorithm**
>
> - Simple linear regression is an example of a supervised learning algorithm (although rarely called that in the introductory statistics course where you probably first encountered it).
>   - The Y variable is the (known) outcome variable and the X variable is a predictor variable.
>   - A regression line is drawn to minimize the sum of squared deviations between the actual Y values and the values predicted by this line.
>   - The regression line can now be used to predict Y values for new values of X for which we do not know the Y value.

Unsupervised Learning Algorithms

- Unsupervised learning algorithms are those used where there is no outcome variable to predict or classify.
- Hence, there is no "learning" from cases where such an outcome variable is known.
- Association rules, dimension reduction methods, and clustering techniques are all unsupervised learning methods.
- Supervised and unsupervised methods are sometimes used in conjunction.

> **Note**
>
> For example, unsupervised clustering methods are used to separate loan applicants into several risk-level groups. Then, supervised algorithms are applied separately to each risk-level group for predicting propensity of loan default.

The Steps in Data Mining

Data Mining Steps

1. **Develop an understanding of the purpose of the data mining project.** How will the stakeholder use the results? Who will be affected by the results? Will the analysis be a one-shot effort or an ongoing procedure?
2. **Obtain the dataset to be used in the analysis.** This often involves sampling from a large database to capture records to be used in an analysis. How well this sample reflects the records of interest affects the ability of the data mining results to generalize to records outside of this sample. It may also involve pulling together data from different databases or sources The databases could be internal (e.g., past purchases made by customers) or external (credit ratings). While data mining deals with very large databases, usually the analysis to be done requires only thousands or tens of thousands of records.
3. **Explore, clean, and preprocess the data.** This step involves verifying that the data are in reasonable condition. How should missing data be handled? Are the values in a reasonable range, given what you would expect for each variable? Are there obvious outliers? The data are reviewed graphically: for example, a matrix of scatterplots showing the relationship of each variable with every other variable. We also need to ensure consistency in the definitions of fields, units of measurement, time periods, and so on. In this step, new variables are also typically created from existing ones. For example, "duration" can be computed from start and end dates.
4. **Reduce the data dimension, if necessary.** Dimension reduction can involve operations such as eliminating unneeded variables, transforming variables (e.g., turning "money spent" into "spent > \$100" vs. "spent ≤ \$100"), and creating new variables (e.g., a variable that records whether at least one of several products was purchased). Make sure that you know what each variable means and whether it is sensible to include it in the model.
5. **Determine the data mining task.** (classification, prediction, clustering, etc.). This involves translating the general question or problem of Step 1 into a more specific data mining question.
6. **Partition the data (for supervised tasks).** If the task is supervised (classification or prediction), randomly partition the dataset into three parts: training, validation, and test datasets.
7. **Choose the data mining techniques to be used.** (regression, neural nets, hierarchical clustering, etc.).
8. **Use algorithms to perform the task.** This is typically an iterative process—trying multiple variants, and often using multiple variants of the same algorithm (choosing different variables or settings within the algorithm). Where appropriate, feedback from the algorithm's performance on validation data is used to refine the settings.
9. **Interpret the results of the algorithms.** This involves making a choice as to the best algorithm to deploy, and where possible, testing the final choice on the test data to get an idea as to how well it will perform. (Recall that each algorithm may also be tested on the validation data for tuning purposes; in

this way, the validation data become a part of the fitting process and are likely to underestimate the error in the deployment of the model that is finally chosen.)

10. **Deploy the model.** This step involves integrating the model into operational systems and running it on real records to produce decisions or actions. For example, the model might be applied to a purchased list of possible customers, and the action might be "include in the mailing if the predicted amount of purchase is > $10." A key step here is "scoring" the new records, or using the chosen model to predict the outcome value ("score") for each new record.

SEMMA

The foregoing steps encompass the steps in SEMMA, a methodology developed by the software company SAS:

- **Sample**: Take a sample from the dataset; partition into training, validation, and test datasets.
- **Explore**: Examine the dataset statistically and graphically.
- **Modify**: Transform the variables and impute missing values.
- **Model**: Fit predictive models (e.g., regression tree, neural network).
- **Assess**: Compare models using a validation dataset.

IBM SPSS Modeler (previously SPSS-Clementine) has a similar methodology, termed CRISP-DM (CRoss-Industry Standard Process for Data Mining). All these frameworks include the same main steps involved in predictive modeling.

Other Models

- **KDD Model**: Knowledge Discovery in Databases (KDD) is a systematic process that seeks to identify valid, novel, potentially useful, and ultimately understandable patterns from large amounts of data. In simpler terms, it's about transforming raw data into valuable knowledge.
- **CRISP-DM**: CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It is a cyclical process that provides a structured approach to planning, organizing, and implementing a data mining project. The process consists of six major phases: Business Understanding, Data Understanding, Data, Preparation, Modeling, Evaluation, Deployment

https://www.youtube.com/watch?v=q_okDS2RtzY

Preliminary Steps

Organization of Datasets

Predicting Home Values in the West Roxbury Neighborhood

Loading and Looking at the Data in R

Sampling from a Database

Oversampling Rare Events in Classification Tasks

Preprocessing and Cleaning the Data

Predictive Power and Overfitting

Overfitting

Creation and Use of Data Partitions

Building a Predictive Model

Using R for Data Mining on a Local Machine

Automating Data Mining Solutions

House management

More to Read

Assignment

- What to do
- Requirement
  - PDF format
  - file name should be include your student id and name
    - stuID_name_title.pdf (e.g. 1111111_ChungilChae_SelfIntroduction.pdf)
- Due date

- by DATE 11:59PM
- NO LATE SUBMISSION ALLOWED!!!!

Reference

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data mining for business analytics: Concepts, techniques, and applications in r*. John Wiley & Sons.

