Data Mining - Introduction

AW

Lecture Overview

1. Data Mining, what is it?
2. Terminology and Settings
3. Types of Data
4. Views of Data

What are we trying to achieve?

GOAL:

Discover information patterns in large data sets

METHOD:

Take a sample of data objects:

- database records
- transactions
- strings (e.g. DNA)
- graphs (e.g. social networks)

and design an algorithm that does one of the folowing

- Assigns a category to a data object
- Assigns real value to a data object
- Groups homogeneous objects
- Finds outlier objects
- Finds non-relevant attributes of an obect

Task examples

- data = text:
  - classify texts (e.g. spam vs. not spam)
  - group (cluster) texts that are similar to each other (e.g. news that describe same events)
  - estimate number of 'likes'
- data = financial records (including stocks)
  - determine fraudulent transactions
  - estimate credit limit
  - find transaction types that usually happen together
- data = social network graphs
  - discover similar groups
  - find group leaders

Data Mining Tasks

Predictive tasks:

Use known values of some variables to predict unknown (or future) values of other variables.

Descriptive tasks:

Find human-interpretable patterns that describe the data.

Data Mining Tasks

Tasks:

- Classification [Predictive]

  assign a category to each item
- Clustering [Descriptive]

  partition data into homogenous (w.r.t. to some measure of similarity) regions
- Association Rule Discovery [Descriptive]

  Find out co-ocurences of data objects
- Regression (i.e. finding a relation between true variable(s) and observed data) [Predictive]

  predict a real value for each item
- Best representation [Desciptive]

  find the transformation of data space to another data space where interesting properties of data are more explicit
- Dimensionality reduction/Feature selection [Desciptive]

  find lower-dimensional space preserving interesting properties of the data

Objectives of Data Mining

Design algorithms:

- Efficient and accurate
- Can deal with large-scale problems
- Can handle a variety of different problems

Answer theoretical questions:

- what patterns can be discovered, under what conditions?
- are there any guarantees?
- How good are our data mining algorithms?

Lecture Overview

1. Data Mining, what is it?
2. Terminology and Settings
3. Types of Data
4. Views of Data

Terminology

Instance: unlabled item/data object

Example: labeled item/object/instance of the data.

Features: attributes associated to an item, often represented as a vector (e.g., word counts). A collection of attributes describe an example.

Labels: category (classification) or real value (regression) associated to an item.

Data: records, points in $\mathbb{R}^n$, graphs, strings

- *training* data (typically labeled)
- *test* data (labeled, but labels not seen)
- *validation* data (labeled, for tuning parameters)

Data Mining Scripts

Settings:

- batch: learner receives full (training) sample, which he uses to make predictions for unseen points.
- on-line: learner receives one sample at a time and makes a prediction for that sample.

Queries:

- active: the learner can request the label of a point.
- passive: the learner receives labeled points.

Lecture Overview

1. Data Mining, what is it?
2. Terminology and Settings
3. Types of Data
4. Views of Data

Data Matrices

$$\begin{aligned}\text{Examples (data objects)} &= \text{set of values of features}\\ &= \text{tuple of attribute values}\end{aligned}$$

An attribute takes values in its domain.
Domains could be:

- natural numbers or integers
- real numbers
- finite sets of symbols (letters, colors, etc.)
- strings in $\{0,1\}^*$
- $\ldots$

Attributes

Examples

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Such data set is an $m \times n$ matrix, where m is the number of examples (i.e. rows are examples, and n is the number of attributes (i.e. columns are attributes)

Data lists

Big number of attributes, and in examples most of the attributes are undefined - better represnt by lists.

Typical examples - transactions:

- a set of products purchased by a customer during one shopping trip
- a set of stocks sold/bought in one transaction
- changes made in resource allocations
- ...

TID	*Items*
1	Bread, Coke, Milk 2
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graphs and Ordered Data

Graphs – sets of nodes V and binary relation $E \subseteq V \times V$ (symmetric or asymmetric).

Examples: web graph, social network graph, chemical graphs (molecules)

Graphs and Ordered Data

Graphs – sets of nodes V and binary relation $E \subseteq V \times V$ (symmetric or asymmetric).

Examples: web graph, social network graph, chemical graphs (molecules)

Ordered Data – sequences of atomic symbols

Examples: DNA sequence in A,C,T,G alphabet, temperature sequences in time, space coordinates of an object in time, etc.

Lecture Overview

1. Data Mining, what is it?
2. Terminology and Settings
3. Types of Data
4. Views of Data

Geometric View of Data

Suggested by data matrix format: each example is a point (vector) in $\mathbb{R}^n$ where n is the number of attributes.

Example:

	Y_1 sepal length	Y_2 sepal width	Y_3 petal length	Y_4 petal width	Z class
$\overline{\mathrm{x}}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\overline{\mathrm{x}}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\overline{\mathrm{x}}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\overline{\mathrm{x}}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\overline{\mathrm{x}}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\overline{\mathrm{x}}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\overline{\mathrm{x}}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\overline{\mathrm{x}}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{\mathrm{x}}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\overline{\mathrm{x}}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

$$\vec{x}_1 = \begin{pmatrix} 5.9 \\ 3.0 \\ 4.2 \\ 1.5 \end{pmatrix}, \vec{x}_2 = \begin{pmatrix} 6.9 \\ 3.1 \\ 4.9 \\ 1.5 \end{pmatrix}, \ldots .$$

Geometric View of Data

Suggested by data matrix format: each example is a point (vector) in $\mathbb{R}^n$ where n is the number of attributes.

Example:

	Y_1 sepal length	Y_2 sepal width	Y_3 petal length	Y_4 petal width	Z class
$\overline{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\overline{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\overline{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\overline{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\overline{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\overline{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\overline{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\overline{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\overline{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Spaces: input space $Y = \mathbb{R}^4$, output space $Z = \{setosa, versicolor, virginica\}$, cassifiier is a map $C : Y \to Z$.

Probabilistic View of Data

The set of all (possible, existing) examples - sample space. Events are outcomes. All events are equally probable.

Example: All irises in the world

Attribute is a random variable i.e. a map of sample space to domain of the attribute. Can be either continuous or discrete.

Example: Sepal width of an iris.

Important distinction between random variable Y_i and its value x^i_j in j^{th} example:

- An attribute is a theoretical function. It has not yet been observed, but it has the potential to take different values with certain probabilities
- Observed value of an attribute in an example is a sampled value from the domain of values that attribute can take

Attribute Distribution

An attribute Y is a random variable, so each of its values happen with some probability, i.e. it has a distribution.

The distribution of a random variable is the collection of possible values of random variable along with their probabilities:

Mass distribution in discrete case :

$$\Pr_Y(x) = \Pr(Y = x)$$

Cumulative distribution in continuous case:

$$F_Y(x) = \Pr(Y < x) = \int_{\infty}^{x} f(t)d(t)$$

where $f(t)$ is probability density function of Y.