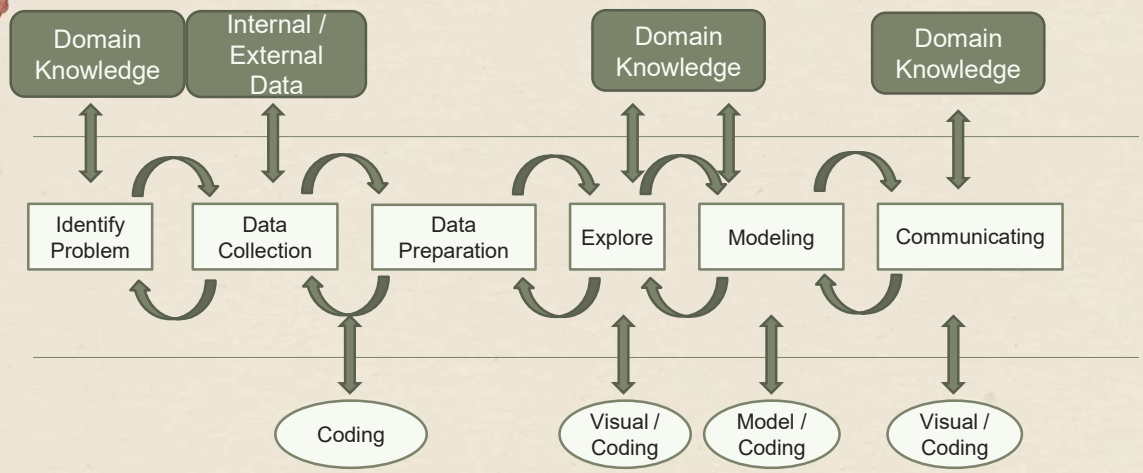
Data Mining and Machine Learning for Business

DSCI 5240

1

What is data mining/business analytics?



- Generating actionable business insights from data
- An iterative process

2

What is machine learning?

"A field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks." (Wikipediae)

"ML is the process of training a piece of software, called a model, to make useful predictions from data." – Google AI

Why machine learning?

Using data, machine learning can:

- Find new ways for previously complicated tasks
- Make it easy to scale tasks
- Make unprogrammable tasks possible

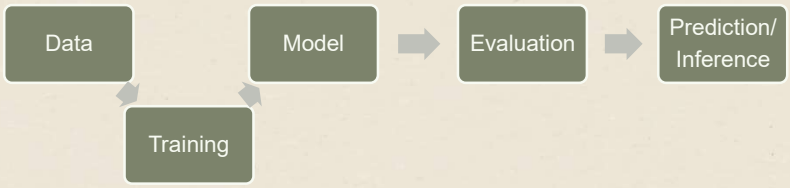
3

Supervised learning

- The correct answers (labels: Y) are given in the data
- Train a model to find the mapping (model) between features (Xs) and Y using data
- Analogy:
  - Students are given past questions and correct answers
  - Students learn by studying the questions and answers
  - Student can answer new questions based on training

4

Key components of Supervised Learning



5

Data

Features

Label

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23

Example

Features: attributes, fields, independent variables, predictors, explanatory variables

Label: dependent variable, target, response

Features

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure
2021-09-09	49.71N	82.16W	74	20	3	N	18.6
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94

Example

6

Data - Notation

In general, we will let x_{ij} represent the value of the jth variable for the ith observation, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. Throughout this book, i will be used to index the samples or observations (from 1 to n) and j will be used to index the variables (from 1 to p). We let $\mathbf{X}$ denote a $n \times p$ matrix whose (i, j)th element is x_{ij}. That is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

You can visualize X as a spreadsheet of numbers with *n* rows and *p* columns

7

Data - Notation

We use y_i to denote the ith observation of the variable on which we wish to make predictions, such as `wage`. Hence, we write the set of all n observations in vector form as

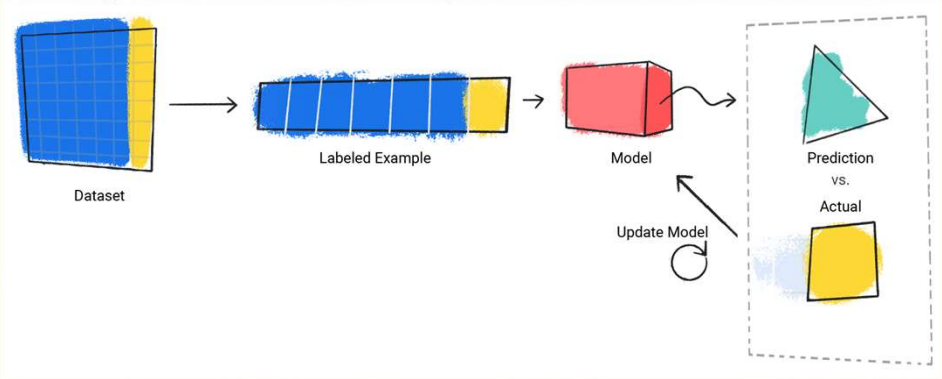
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Then our observed data consists of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each x_i is a vector of length p. (If $p = 1$, then x_i is simply a scalar.)

8

Training

- To find the best solution for predicting Y from Xs
- By comparing prediction with actual values, the model keeps updating itself
- Finally, it finds the best model to make predictions



Source: Google AI

9

Model

A mathematical mapping between features (Xs) to output labels (Y)
Most often seen models:

- Regression
- Logistic regression
- Tree-based methods
- Support vector machine
- Neural Networks
- Generative models
- …

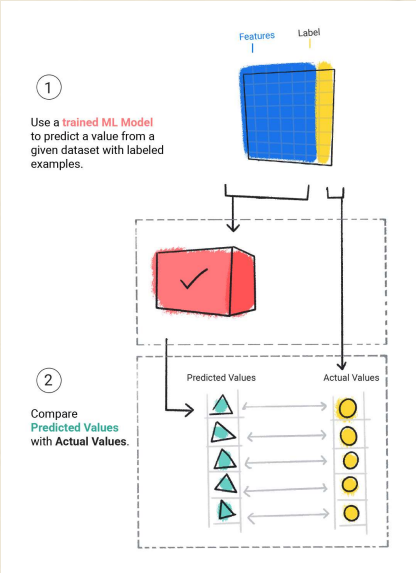
We configure

- Parameters
- Features

10

Evaluation

Evaluate the performance of models to find the best one



11

Prediction vs. Inference

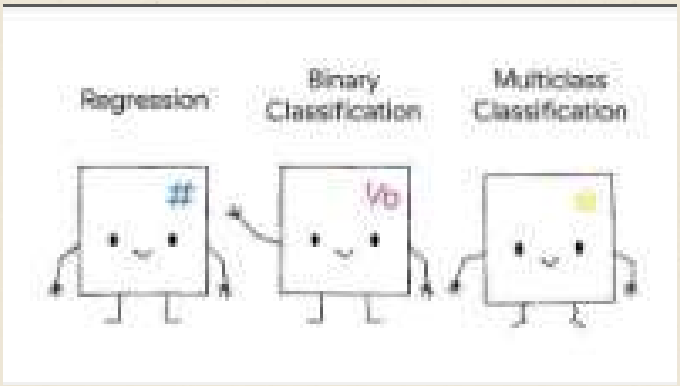
- Prediction: main goal is to make accurate prediction
  - E.g., A direct marketing campaign that only cares about to make accurate prediction about which customer will respond
- Inference: to understand how Y changes when X changes
  - E.g., the problem of advertising budget allocation may involve questions like: which media generate the biggest boost in sales? How much increase in sales is associated with a given increase in TV advertising?
- Focus of projects: prediction, inference or both
- Different methods may be appropriate for different focus

12

Regression vs classification

In a **regression** problem, Y is quantitative (e.g., continuous variables such as price, wage)

In a **classification** problem, Y is qualitative, taking values in a finite set (e.g., respond/not respond, win/lose, choice of brands, cancer types)



13

Unsupervised learning

- No correct answers (labels), just a set of features measured on a set of samples
- Objective is more fuzzy – find groups of samples that behave similarly, features that behave similarly, etc.
- Difficult to know how well you are doing
- The most common use
  - Clustering: cluster data into similar groups based on features

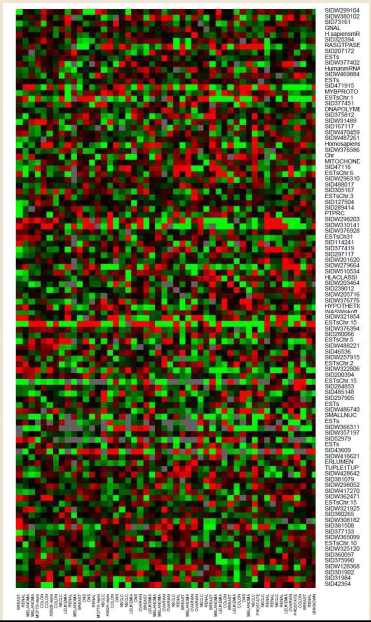
14

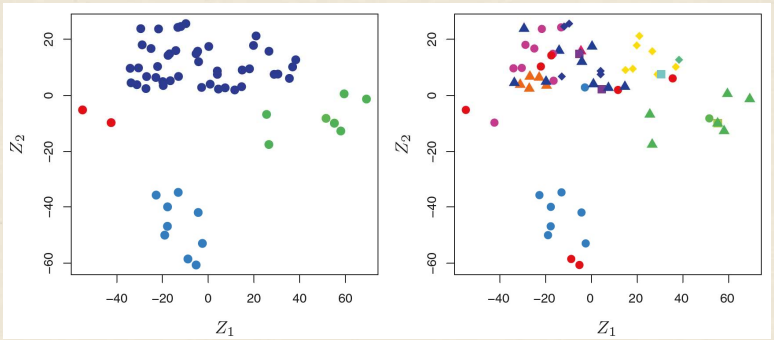
Gene Expression

- The data
  - 64 cell lines
  - Each cell line has 6830 gene expression measurements
  - Associated with 14 cancer types
- We need to
  - Identify whether there are groups, or clusters among the cell lines based on their gene expression measurements

15

Example – Gene Expression





16

Our philosophy

- Optimizing algorithms and theoretical properties of statistical models are not our primary interest – it is not necessary to have the skills to construct the machine
- However, it is important to understand the idea behind each method, in order to know when and how to use them
- One has to understand simpler methods first to grasp more sophisticated ones
- To apply methods to real-world problems

17

Get to know Python

18