Predictive Analytics

(Course: Introduction to Data Science)

Tirtharaj Dash

BITS Pilani, K.K. Birla Goa Campus

tirtharaj@goa.bits-pilani.ac.in

October 23, 2020

Data Analytics

- Data analytics is the science of analyzing raw data in order to make conclusions about that information.
- The techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
- Data analytics help a business optimize its performance.

Types of Data Analytics I

- **Descriptive analytics** describes what has happened over a given period of time.
  - Have the number of views gone up?
  - Are sales stronger this month than last?
- **Diagnostic analytics** focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing.
  - Did the weather affect beer sales?
  - Did that latest marketing campaign impact sales?

Types of Data Analytics II

- **Predictive analytics*** moves to what is likely going to happen in the near term.
  - What would happen to the sales if we have a hot summer this year?
  - What is the probability that the summer will be hot this year?
- **Prescriptive analytics** suggests a course of action.
  - If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.

*We will restrict our study to predictive analytics only

Predictive Analytics

- Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyze current and historical facts to make predictions about future or otherwise unknown events.
- Predictive analytics and machine learning are not same! PA is one of the most common enterprise applications of machine learning.

*In this course, we will restrict to some standard machine learning models (with minimal mathematical- and probabilistic foundation).

Machine Learning (ML) I

Arthur Samuel (1959) Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998) Well-posed learning problem: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Machine Learning (ML) II

Example: Spam classification

Suppose your computer program watches which email you do or do not mark as 'spam' and based on that it learns how to better filter spams.

Can you identify T, E and P in this setting?

Machine Learning (ML) III

- T Classifying emails as spam or not spam.
- E Watching us label emails as spam or not spam.
- P The number (or fraction) of emails classified as spam or not spam.

ML algorithms

The main two types of algorithms are based on

- Supervised learning
- Unsupervised learning

Third type: Reinforcement learning

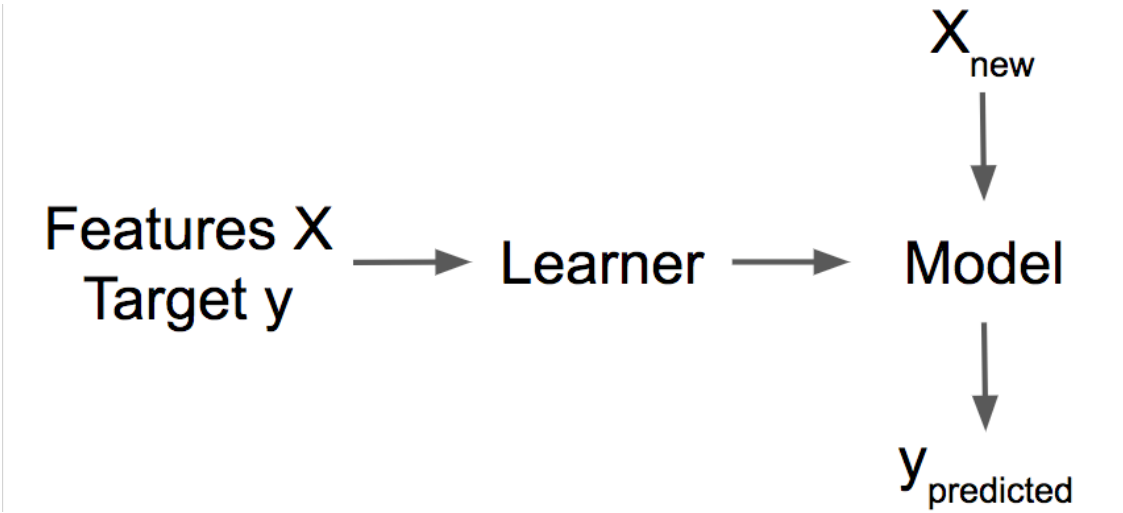
Supervised learning I

- It builds a mathematical model of a set of data that contains both the inputs and the desired outputs.
- The data is known as training data, and consists of a set of training examples.
- Each training example has one or more inputs and the desired output, also known as a supervisory signal.
- In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix.

Supervised learning II

- Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs.
- An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data.
- An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.

Supervised learning III



A supervised machine learning set-up: a learner builds a model with many such (X, y) pairs. The built model can be used to predict an output for a new data instance represented by a feature vector X_{new}.

Supervised learning IV

Types of supervised learning: classification, regression.

Classification Classification algorithms are used when the outputs are restricted to a limited set of values.
Example: classifying type of cancer

Regresion Regression algorithms are used when the outputs may have any numerical value within a range.
Example: prediction of house price

Unsupervised learning I

- These algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points.
- The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data.

Unsupervised learning II

- One common example of unsupervised learning is: cluster analysis (or clustering).
- It is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar.