

Machine learning

Supervised: labeled data – making predication – generalization, extracting structure- representation

Unsupervised: unlabeled data

Supervised Learning:

- Vector(list) of “Predictors” X
 - Also known as features, independent variables, inputs, regressors, covariants, attributes
- “Response” y
 - Also known as outcome, label, target, dependent variable
 - If y is continuous: Regression

E.g price, blood pressure

If y is categorical (values in a finite, unordered set): Classification

e.g spam/ham, digit 0-9, cancer class of tissue sample

Data is composed of “observations” (predictors and the associated response)

Also known as samples, examples, instances, records

Regression example:

Wage and Age

Wage and year

Wage and Education Level

Example #1

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

Goal: Detect subtle patterns in the data that predicts infection before it occurs

Data (predictors/ features of the data, have some relationship with the variable we are trying to find): 16 vital signs such as heart rate, respiration rate, blood pressure, etc

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Sample response: Did the child develop an infection? True/False

Classification example

Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements. - Supervised

Identify the numbers in a handwritten code –

Given historical stock market data, determine if the market will go up or down.

Supervised

Given historical weather data, predict next week's avrg temp

Regression- supervised

Given historical stock market data, determine tomorrow's trading volume

Supervised regression

Recognize a person's face in an image

Supervised classification

Unsupervised learning

No response variable y , just a set of predictors X

Objective is more open:

Find groups of observations that behave similarly

Find predictors that behave similarly

Find combinations of features that explain the variation in the data

Difficult to evaluate how well you are doing

Data is easier to obtain for unsupervised learning since it can be "unlabeled" (i.e it hasn't been labeled with a response)

Sometimes useful as a preprocessing step for supervised learning

Common techniques: clustering, principal components analysis

Clustering example

Classify US residential neighborhoods into 676 unique segments based on demographic and socioeconomic characteristics

Ersi.com/landing-pages/tapestry