Multivariate data analysis

R Workshop

Mohamad Ghassany & Altay Ozaygen

24/02/2017

Overview

Machine Learning

The aim of ML is to build computer systems that can adapt to their environments and learn form experience.

Application examples:

- effective web search
- social networks recognize friends from photos or suggest friends
- email spam detection
- handwriting recognition
- understanding the human genome
- predict possibility for a certain disease on basis of clinical measures
- fraud detection
- drive vehicles
- recommendations (eg, Amazon, Netflix)

Machine Learning

Automatically learn programs by **generalizing from examples**. As more data becomes available, more ambitious problems can be tackled.

Machine Learning is a branch of artificial intelligence and an interdisciplinary field of CS, statistics, math and engineering.

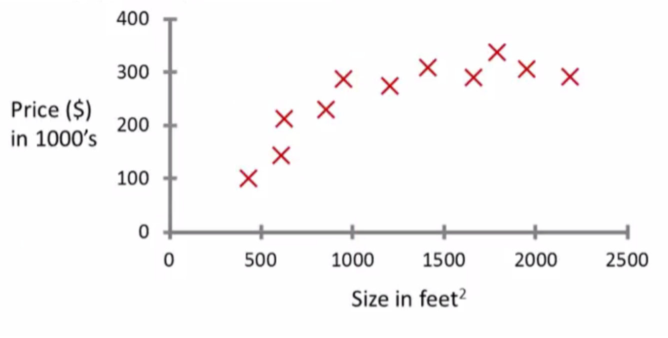
In general, any machine learning problem can be assigned to one of two broad classifications:

Supervised Learning and Unsupervised Learning

Supervised Learning

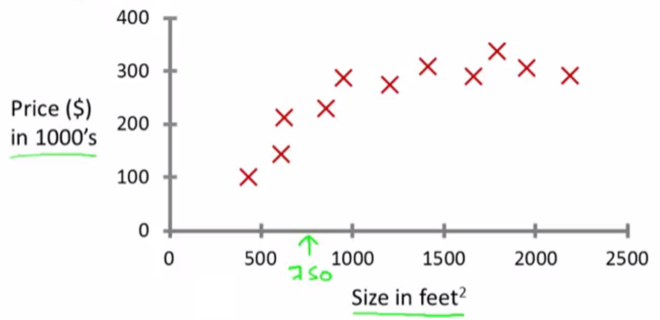
Example: House price prediction

Let's say we want to predict housing prices. We plot a data set and it looks like this



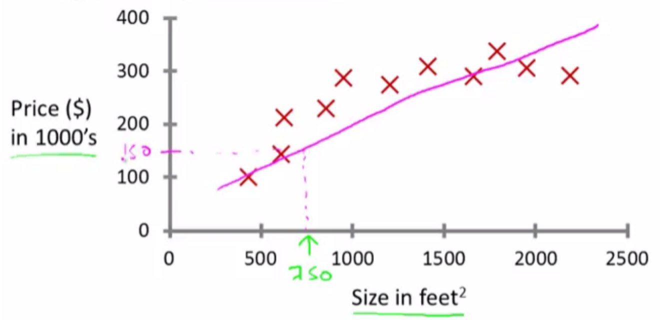
Example: House price prediction

Let's say we own a house that is, say 750 square feet and hoping to sell the house and we want to know how much we can get for the house.



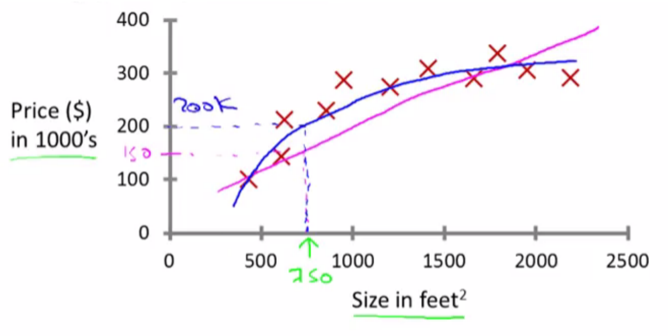
Example: House price prediction

A learning algorithm can for example **"fit"** a straight line to the data and, based on that, it looks like maybe the house can be sold for maybe about $150,000.



Example: House price prediction

There might be a better learning algorithm! Maybe a *quadratic function* to this data.



If we do that, and make a prediction here, then it looks like maybe we can sell the house for closer to $200,000.

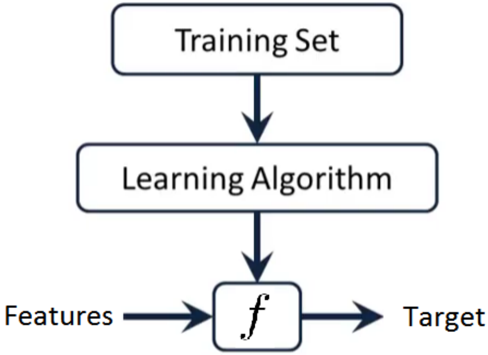
Supervised Learning: Definition

The term supervised learning refers to the fact that we gave the algorithm a data set in which the **"right answers"** (know as **labels**) were given.

Notations:

- The size of the house is the **input** variable. Typically denoted by X.
- The inputs go by different names, such as *predictors*, *independent variables*, *features*, *predictor* or sometimes just *variables*.
- The house price is the **output** variable, and is typically denoted using the symbol Y.
- The output variable is often called the *response*, *dependent variable* or *target*.

Supervised Learning: Model



- Supervised Learning refers to a set of approaches for estimating f.
- f is also called ***hypothesis*** in Machine Learning.

Regression and Classification

Regression:

- The example of the house price prediction is also called a **regression** problem.
- A regression problem is when we try to predict a **quantitative (continuous)** value output. Namely the price in the example.

Classification:

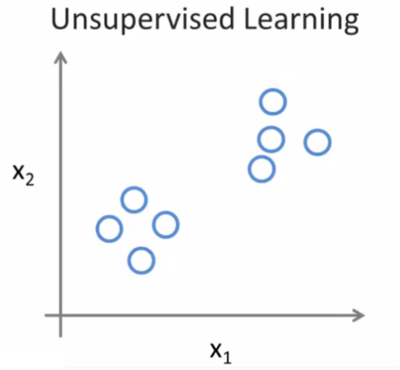
- The process for predicting **qualitative (categorical, discrete)** responses is known as classification.
- Methods: Logistic regression, Support Vector Machines, etc..

Unsupervised Learning

Unsupervised Learning: "No labels"

In Unsupervised Learning, we're given data that doesn't have any **labels**.

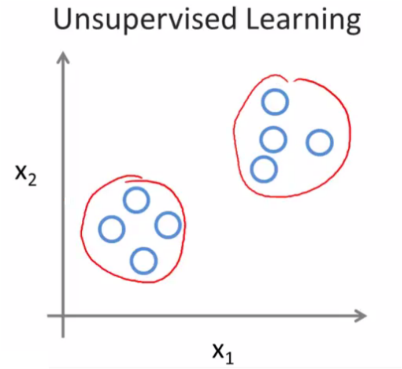
For example:



Question: Can you find some structure in the data?

Unsupervised Learning: Structure

Given this data set, an Unsupervised Learning algorithm might decide that the data lives in two different clusters.



This is called a **clustering** algorithm.

Unsupervised Learning: Example

One example where clustering is used is in Google News (news.google.com)

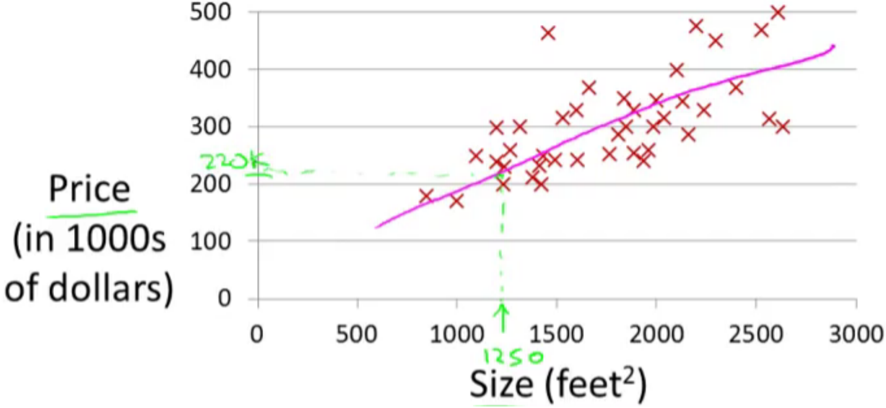
Additional readings

Additional readings:

- The Elements of Statistical Learning (by Friedman, Tibshirani and Hastie)
- Pattern Recognition and Machine Learning (by Bishop)
- Andrew Ng.'s Machine Learning course on Coursera

Supervised Learning – Predictive Models

Linear Regression



Linear Regression - Model

Simple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$

Linear Regression - Model

Simple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$

- Prediction: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Linear Regression - Model

Simple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- Prediction: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- The coefficients minimize: $RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

Linear Regression - Model

Simple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- Prediction: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- The coefficients minimize: $RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- Coefficients:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Linear Regression - Model

Multiple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$

Linear Regression - Model

Multiple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$
- Matrix notation: $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}$

Linear Regression - Model

Multiple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$
- Matrix notation: $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}$
- Coefficients: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

Some important questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Linear Regression - Hypothesis testing

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

vs

$$H_1 : \text{There is some relationship between } X \text{ and } Y$$

Linear Regression - Hypothesis testing

H_0 : There is no relationship between X and Y

vs

H_1 : There is some relationship between X and Y

$$H_0 : \beta_i = 0 \quad \forall \ i$$

vs

$$H_1 : \exists \ i \quad s.t. \quad \beta_i \neq 0$$

Linear Regression – Example

	Coefficient	Std. error	t-statistic	p-value
Constant	2.939	0.3119	9.42	<0.0001
X_1	0.046	0.0014	32.81	<0.0001
X_2	0.189	0.0086	21.89	<0.0001
X_3	-0.001	0.0059	-0.18	0.8599

In this table we have the following model

$$Y = 2.939 + 0.046X_1 + 0.189X_2 - 0.001X_3$$

Note that for each individual predictor a t-statistic and a p-value were reported. These p-values indicate that X_1 and X_2 are related to Y, but that there is no evidence that X_3 is associated with Y, in the presence of these two.

Application on R

1. Tutorial on European Union dataset.
2. Application on the “Boston” data set.

Lien: http://mghassany.com/seminaireR/R_Workshop.html