

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 from 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.

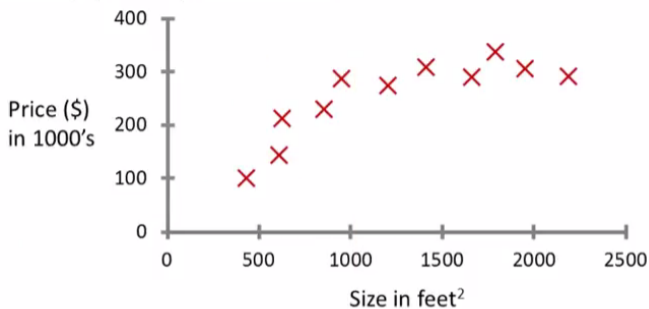
In general, any machine learning problem can be assigned to one of two broad classifications:

Supervised Learning and **Unsupervised Learning**

Supervised Learning

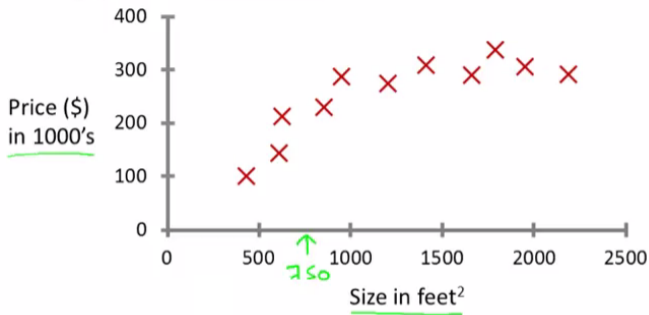
Example: House price prediction

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



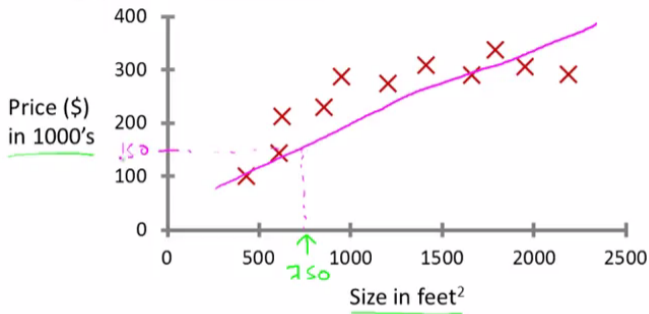
Example: House price prediction

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.



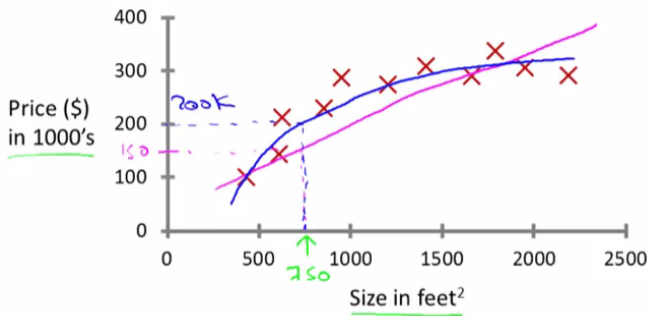
Example: House price prediction

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.

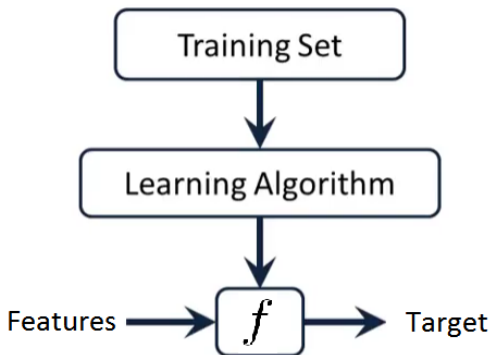
Supervised Learning: Definition

The term supervised learning refers to the fact that we gave the algorithm a data set in which the “**right answers**” (known 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:

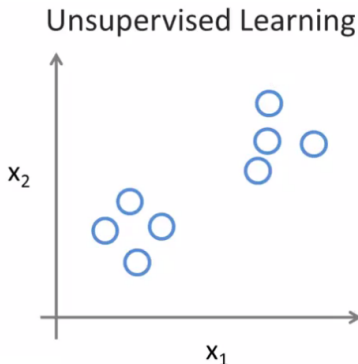
- The process for predicting **qualitative (categorical, discrete)** responses is known as classification.
- Methods: Logistic regression, Support Vector Machines, etc..

Unsupervised Learning

Unsupervised Learning: “No labels”

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

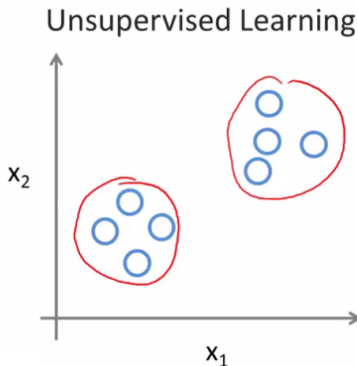
For example:



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)

The screenshot shows the Google News homepage. At the top is the Google search bar with the text "Edition France" and a "Connexion" button. Below the search bar, there are several news articles and a sidebar.

Actualités - Edition France

A la une

Peillon et Valis présentent leurs programmes
 Le Monde - il y a 2 heures
 Les deux candidats à la primaire à gauche ont exposé les grandes lignes de leurs projets, mardi 3 janvier. Le Monde | 03.01.2017 à 11h54 • Mis à jour le 03.01.2017 à 12h35 | Par Bastien Bonnefoux, Manuel Vallis, sur le marché de Noël de Strasbourg, le ...
 Autres: Manuel Vallis »

Un nourrisson est mort après une prise de vitamine D
 Le Monde - il y a 4 heures
 Des malaises ont déjà été signalés après l'administration d'Uvestérol. Des investigations doivent déterminer si le décès survenu en décembre est imputable à ce médicament. Le Monde | 02.01.2017 à 21h02 • Mis à jour le 03.01.2017 à 12h06 | Par Paul ...
 Citée à de nombreuses reprises : Un nourrisson décède après la prise de vitamine D | France-Santé
 Articles les plus référencés : Prescrire - Tous les articles en Une "En fin avec Uvestérol" pour éviter les malaises chez les nouveau-nés, 1er mai ... Prescrire
 Editorial : Vitamine D : les pédiatres prudents après le décès d'un nourrisson
 La Croix

Autres: Ergocalcérol »

Foot - Transfert - Le PSG officialise l'arrivée de Julian Draxler (Wolfsburg)
 L'Equipe.fr - il y a 2 heures
 Le PSG a annoncé la signature de Julian Draxler, ce mardi, en provenance de Wolfsburg. Le milieu offensif allemand s'est engagé jusqu'en 2021. Passager sur Facebook Twitter Google+ 0 partages. Football - Transferts Football - Julian Draxler a signé un ...
 Editorial : Mercato: le PSG actif, l'OM attendu | La Croix
 Articles de fond : Batshuayi, Dybala, Payet : les 8 infos mercato qui vous ont échappé ce lundi | Eurosport.fr
 Autres: Paris Saint-Germain Football Club » Julian Draxler »

Autres: Manuel Vallis »

Autres: Ergocalcérol »

Autres: Julian Draxler »

Connectez-vous pour recevoir l'actualité des sujets qui vous intéressent.
 En savoir plus

Articles récents

Notre-Dame-des-Landes : l'évacuation de la ZAD, c'est maintenant ?
 Le Point - il y a 35 minutes

Météo à Arcueil, Île-de-France

Aujourd'hui mer. 3°-1° jeu. 6°-3° ven. 2°-3°

Arcueil, Île-de-France »

Paris: 14 000 habitants ont quitté la capitale depuis 2009
 MCE Ma Chaire Étudiante - il y a 2 heures
 Théâtre du Châtelet : la Britannique Ruth Mackenzie prend la direction
 Le Point - il y a 50 minutes
 Où vivre en sécurité en région parisienne ?
 Une étude inédite au cas par cas
 20minutes.fr - il y a 1 heure

Le choix des rédactions

EUROSPORT

La LI cherche des renforts à prix raisonnable ? Le onze des bonnes pioches
 L'Équipe.fr
 Draxler débarque officiellement au PSG
 Cyril Morin
 Le PSG va prolonger Cavani mais peine à trouver un nouveau port de...
 Pierre-Alexandre Crois
 Griezmann : "Ne me posez plus de questions sur mon avenir"
 Eurosport
 Le mercato EN DIRECT : Arsenal pense à Planck
 Fabien Bonie

Téléchargez l'application Eurosport.fr

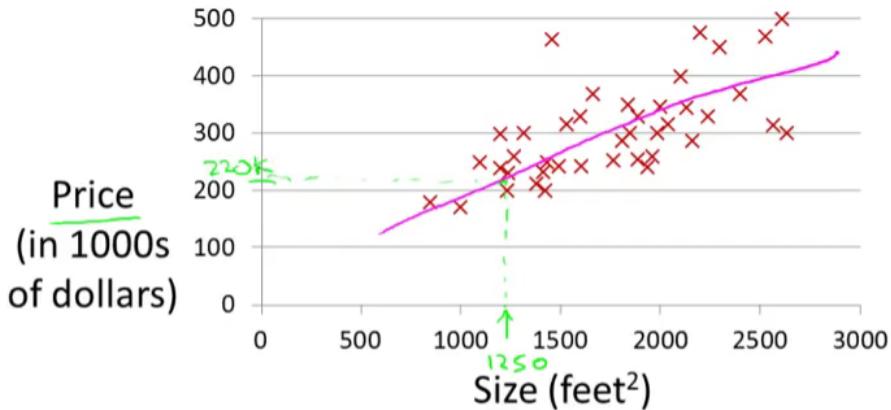
Additional readings

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: $\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1}$

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: $\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1}$
- Coefficients: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$

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

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

Application on the “Boston” data set.