Introduction to Data Analytics

ITE 5201

Lecture7-Linear Regression

Instructor: Parisa Pouladzadeh
Email: parisa.pouladzadeh@humber.ca

www.udemy. /course/python-for-data-science-and-machine-learning.com

HUMBER

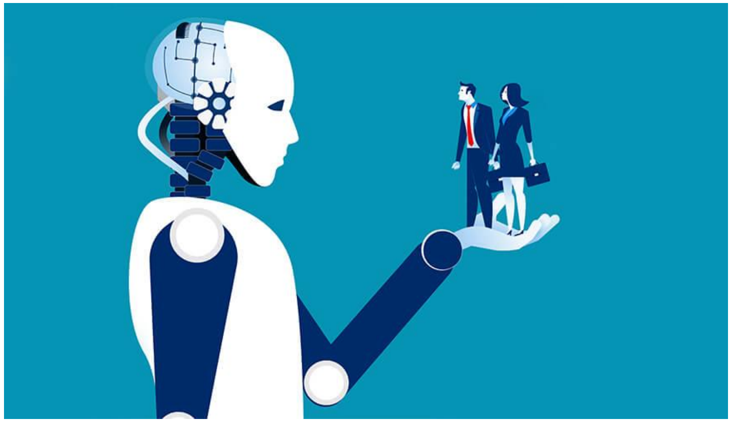
Introduction

Machine learning is making great strides

- Large, good data sets
- Compute power
- Progress in algorithms

Many interesting applications

- commericial
- scientific



Machine learning tasks

Supervised learning

- regression: predict numerical values
- classification: predict categorical values, i.e., labels

Unsupervised learning

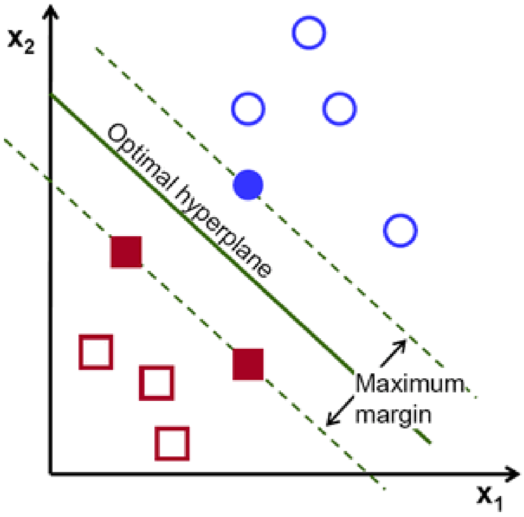
- clustering: group data according to "distance"
- association: find frequent co-occurrences
- link prediction: discover relationships in data
- data reduction: project features to fewer features

Machine learning algorithms

Regression:
regression, Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...

Classification:
Naive Base, , Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...

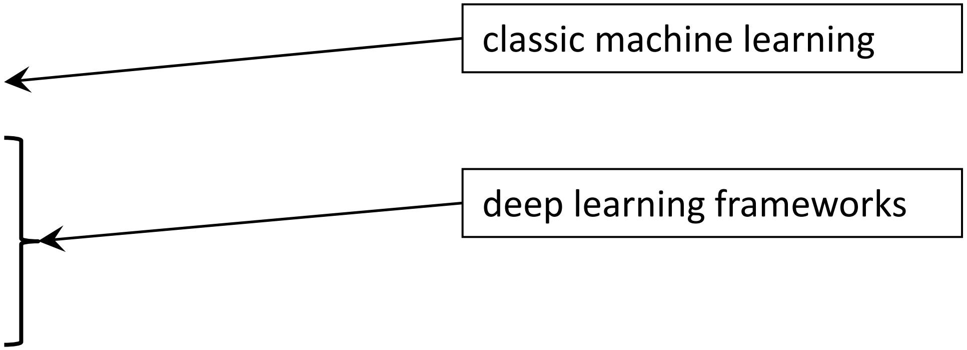
Clustering:
k-Means, Hierarchical Clustering, ...



Frameworks

Programming languages
- Python
- R
- C++
- ...

Many libraries
- scikit-learn
- PyTorch
- TensorFlow
- Keras
- …



classic machine learning

deep learning frameworks

scikit-learn

Nice end-to-end framework

- data exploration
- data preprocessing (+ pandas)
  - cleaning/missing values
  - normalization
- training
- testing
- application

"Classic" machine learning only

Supervised learning: methodology

Select model, e.g., random forest, (deep) neural network, ...

Train model, i.e., determine parameters

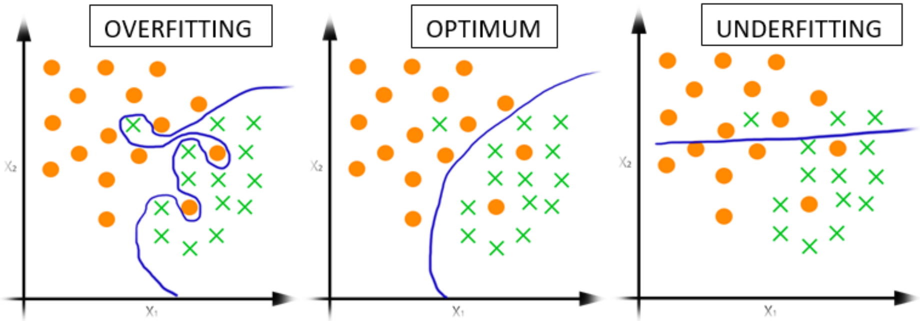
- Data: input + output
  - training data → determine model parameters
  - validation data → to avoid overfitting

Test model

- Data: input + output
  - testing data → final scoring of the model

Production

- Data: input → predict output



Metrics

It is extremely important to use quantitative metrics for evaluating a machine learning model

For classification

- Accuracy/Precision/Recall/F1-score, ROC curves,…

For regression

- Normalized RMSE, Normalized Mean Absolute Error (NMAE)

Precision and recall

Suppose that $y = 1$ in presence of a **rare class** that we want to detect

Precision *(How much we are precise in the detection)*

Of all patients where we predicted $y = 1$, what fraction actually has the disease?

$$\frac{\text{True Positive}}{\#\ \text{Predicted Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall *(How much we are good at detecting)*

Of all patients that actually have the disease, what fraction did we correctly detect as having the disease?

$$\frac{\text{True Positive}}{\#\ \text{Actual Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Confusion matrix

		Actual class	
		1 (p)	0 (n)
Predicted class	1 (Y)	True positive (TP)	False positive (FP)
	0 (N)	False negative (FN)	True negative (TN)

Trading off precision and recall

Logistic regression: $0 \leq h(\boldsymbol{x}) \leq 1$

- Predict 1 if $h(\boldsymbol{x}) \geq 0.5$
- Predict 0 if $h(\boldsymbol{x}) < 0.5$

These thresholds can be different from 0.5!

At different thresholds, correspond different confusion matrices!

Suppose we want to predict $y = 1$ (disease) only if very confident

- Increase threshold → Higher precision, lower recall

Suppose we want to avoid missing too many cases of disease (avoid false negatives).

- Decrease threshold → Higher recall, lower precision

F1-score

It is usually better to compare models by means of one number only. The $\mathrm{F1-score}$ can be used to combine precision and recall

	Precision(P)	Recall (R)	Average	F_1 Score
Algorithm 1	0.5	0.4	0.45	0.444
Algorithm 2	0.7	0.1	0.4	0.175
Algorithm 3	0.02	1.0	0.51	0.0392

The best is Algorithm 1

Algorithm 3 predict always 1

Average says not correctly that Algorithm 3 is the best

$$\mathrm{Average} = \frac{\mathrm{P}+\mathrm{R}}{2} \qquad \mathrm{F}_1\mathrm{score} = 2\frac{\mathrm{PR}}{\mathrm{P}+\mathrm{R}}$$

- $\mathrm{P} = 0 \text{ or } \mathrm{R} = 0 \Rightarrow \mathrm{F}_1\mathrm{score} = 0$
- $\mathrm{P} = 1 \text{ and } \mathrm{R} = 1 \Rightarrow \mathrm{F}_1\mathrm{score} = 1$

Regression Analysis

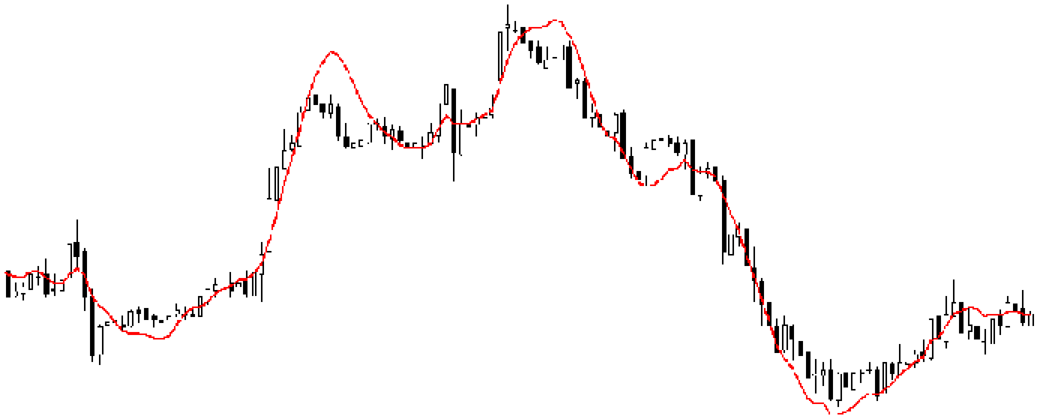
Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).

This technique is used for forecasting, time series modelling and finding the casual effect relationship between the variables.

For example, relationship between rash driving and number of road accidents by a driver.

Regression analysis

- Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.



Why do we use Regression Analysis?

Regression analysis estimates the relationship between two or more variables. Let’s understand this with an easy example:

Let’s say, you want to estimate growth in sales of a company based on current economic conditions.

You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy.

Using this insight, we can predict future sales of the company based on current & past information.

Types of regression techniques

- There are various kinds of regression techniques available to make predictions.
- These techniques are mostly driven by three metrics
  - number of independent variables
  - type of dependent variables
  - shape of regression line

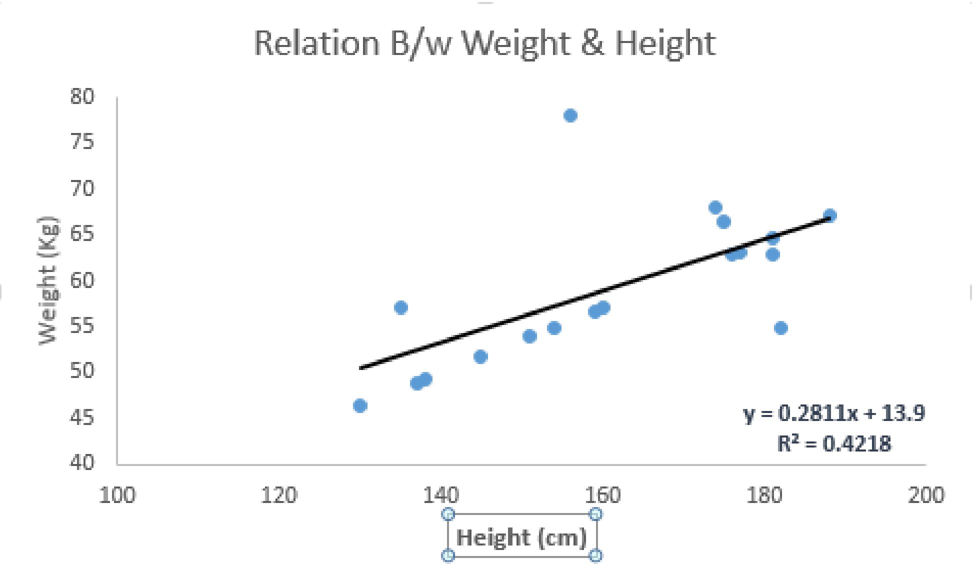
Linear regression

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line.

It is represented by an equation Y=a+b*X + e,

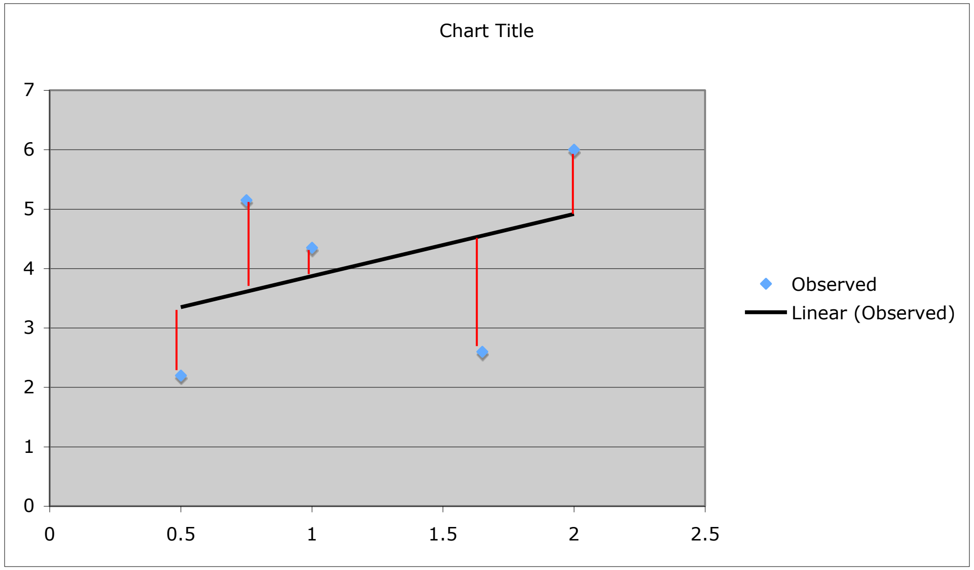
- where a is intercept
- b is slope of the line
- e is error term.
- This equation can be used to predict the value of target variable based on given predictor variable(s).

Linear Regression Ex.



Residual Error

- Observed value - Predicted value



Sum-squared Error (SSE)

$$SSE = \sum_{y} (y_{observed} - y_{predicted})^2$$

$$TSS = \sum_{y} (y_{observed} - \bar{y}_{observed})^2$$

$$R^2 = 1 - \frac{SSE}{TSS}$$

What is R-Squared

It’s a statistical measure between 0 and 1 which calculates how similar a regression line is to the data it’s fitted to.

- If it’s a 1, the model 100% predicts the data variance
- if it’s a 0, the model predicts none of the variance.

MSE, RMSE

The essential step in any machine learning model is to evaluate the accuracy of the model. The Mean Squared Error, Mean absolute error, Root Mean Squared Error, and R-Squared

MSE

MSE:

MSE (Mean Squared Error) is the average squared error between actual and predicted values.

Squared error, is a row-level error calculation where the difference between the prediction and the actual is squared.

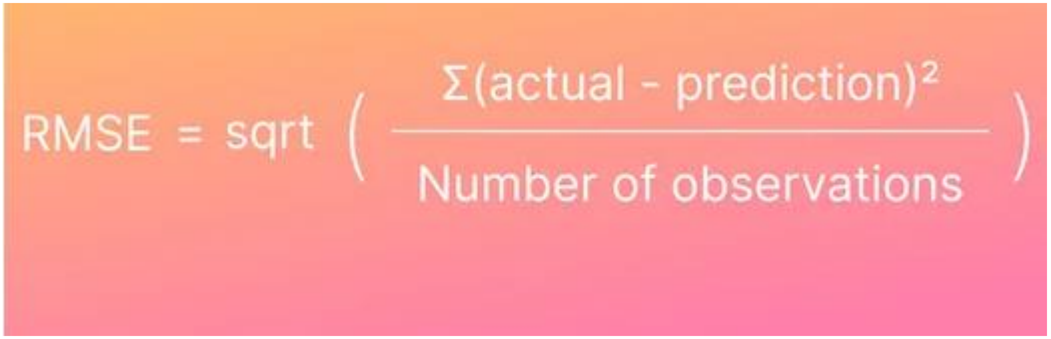
The main draw for using MSE is that it squares the error, which results in large errors being punished or clearly highlighted.

$$\text{MSE} = \frac{\Sigma(\text{actual} - \text{prediction})^2}{\text{Number of observations}}$$

RMSE

Root Mean Squared Error (RMSE) is the square root of the mean squared error (MSE) between the predicted and actual values.

A benefit of using RMSE is that the metric it produces is in terms of the unit being predicted. For example, using RMSE in a house price prediction model would give the error in terms of house price, which can help end users easily understand model performance.



$$\text{RMSE} = \text{sqrt}\left(\frac{\Sigma(\text{actual} - \text{prediction})^2}{\text{Number of observations}} \right)$$

R squared compared to RMSE

RMSE (or MSE) is the measure of goodness of predicting the validation/test values, while R^2 is a measure of goodness of fit in capturing the variance in the training set.

R Square is not only a measure of Goodness-of-fit, it is also a measure of how much the model (the set of independent variables you selected) explain the behavior of your dependent variable.

So, both R Square (and Adjusted R Square) and the Standard Error are extremely useful in assessing the statistical robustness of a model. And, as indicated they have completely different practical application. One measures the explanatory power of the model. The other one allows you to build Confidence Intervals. Both, very useful but different stuff.

Linear model assumptions

Linear regression analysis is based on six fundamental assumptions:

- The dependent and independent variables show a linear relationship between the slope and the intercept.
- The independent variable is not random.
- The value of the residual (error) is zero.
- The value of the residual (error) is constant across all observations.
- The value of the residual (error) is not correlated across all observations.
- The residual (error) values follow the normal distribution.

Multiple linear regression

In Multiple linear regression multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX1 + cX2 + dX3 + \epsilon$$

Simple and Multiple linear regression

- The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.
- Now, the question is “How do we obtain best fit line?”.