

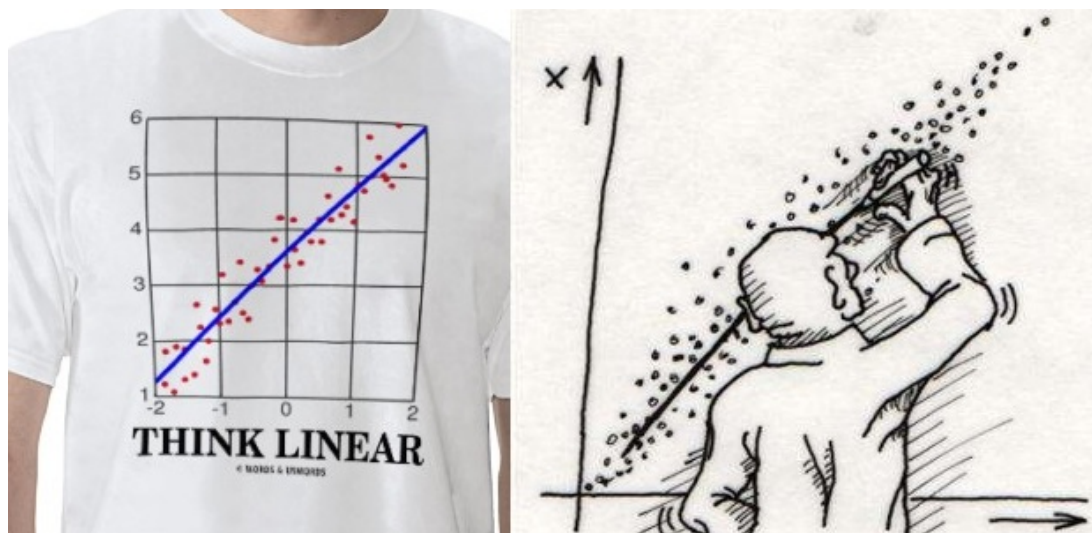


DataAspirant

To express my passion on data mining field

LINEAR REGRESSION

Posted on October 2, 2014 under DATAMINING



Introduction to Linear Regression:

Linear Regression means predicting scores of one variable from the scores of second variable. The variable we are predicting is called the **criterion variable** and is referred to as **Y**. The variable we are basing our predictions is called the **predictor variable** and is referred to as **X**. When there is only one predictor variable, the prediction method is called simple regression. The aim of linear regression is to finding the best-fitting straight line through the points. The best-fitting line is called a **regression line**.

$$h_0(x) = A_0 + A_1 x$$

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

The above equation is **hypothesis equation**

where:

$h_{\theta}(x)$ is nothing but the value Y (which we are going to predicate) for particular x (means Y is a linear function of x)

θ_0 is a **constant**

θ_1 is the **regression coefficient**

x is the value of the independent variable

Properties of the Linear Regression Line

Linear Regression line has the following properties:

1. The line minimizes the sum of **squared differences between observed values** (the y values) and predicted values (the $h_{\theta}(x)$ values computed from the regression equation).
2. The regression line passes through the **mean of the x values** (x) and through the mean of the Y values ($h_{\theta}(x)$).
3. The regression constant (θ_0) is equal to the y **intercept** of the regression line.
4. The regression coefficient (θ_1) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

Goal of Hypothesis Function

Goal of Hypothesis is to choose **θ_0 and θ_1** , so that $h_{\theta}(x)$ is close to Y for our **training data**, while choosing θ_0 and θ_1 we have to consider the cost function(

$J(\theta)$) where we are getting low value for cost function($J(\theta)$).

The below function is called as cost function, cost function ($J(\theta)$) is nothing but just a [Squared error function](#).

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2.$$

Let's Understand Linear Regression with Example

Before going to explain linear Regression let me summarize the things we learn

Hypothesis: $h_{\theta}(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}} J(\theta_0, \theta_1)$

dataaspirant.wordpress.com

Have some function $J(\theta_0, \theta_1)$

Want $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Outline:

- Start with some θ_0, θ_1
- Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$

until we hopefully end up at a minimum

dataaspirant.wordpress.com

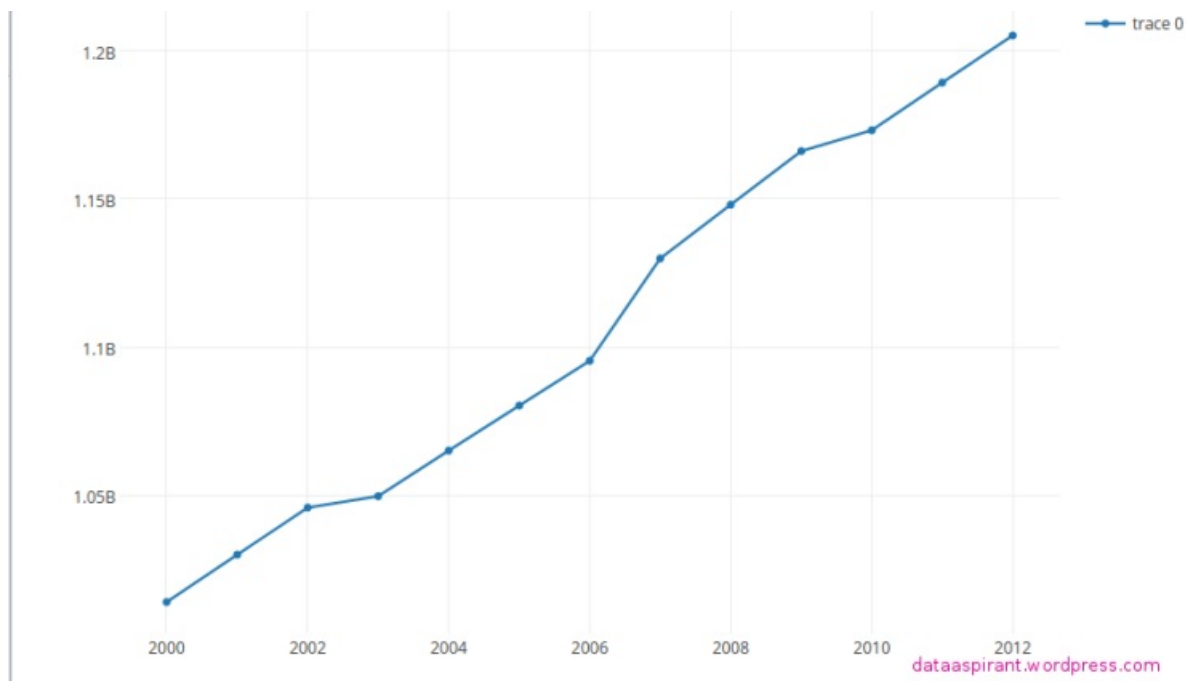
Suppose we have data some thing look's like this

No.	Year	Population
1	2000	1,014,004,000
2	2001	1,029,991,000
3	2002	1,045,845,000
4	2003	1,049,700,000
5	2004	1,065,071,000
6	2005	1,080,264,000
7	2006	1,095,352,000
8	2007	1,129,866,000
9	2008	1,147,996,000
10	2009	1,166,079,000
11	2010	1,173,108,000
12	2011	1,189,173,000
13	2012	1,205,074,000

Now our task is to answer the below questions

No.	Year	Population
1	2014	?
2	?	2,205,074,000

Let me draw a graph for our data



Python Code for graph

```
import plotly.plotly as py from plotly.graph_objs import *
```

```
py.sign_in("username", "API_authentication_code")
```

```
from datetime import datetime
```

```
x = [
```

```
datetime(year=2000,month=1,day=1),  
datetime(year=2001,month=1,day=1),  
datetime(year=2002,month=1,day=1),  
datetime(year=2003,month=1,day=1),  
datetime(year=2004,month=1,day=1),  
datetime(year=2005,month=1,day=1),  
datetime(year=2006,month=1,day=1),  
datetime(year=2007,month=1,day=1),  
datetime(year=2008,month=1,day=1),  
datetime(year=2009,month=1,day=1),  
datetime(year=2010,month=1,day=1),  
datetime(year=2011,month=1,day=1),  
datetime(year=2012,month=1,day=1)
```

```
]
```

```
data = Data([
```

```
Scatter(
```

```
x = x,
```

```
y = [
```

```
1014004000,
```

```
1029991000,
```

```
1045845000,
```

```
1049700000,
```

```
1065071000,
```

```
1080264000,
```

```
1095352000,
```

```
1129866000,
```

```
1147996000,
```

```
1166079000,
```

```
1173108000,
```

```
1189173000,
```

```
1205074000]
```

```
)
```

])

```
plot_url = py.plot(data, filename='DataAspirant')
```

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

- Now what we will do is we will find the **most suitable value** for our θ_0 and θ_1 using hypotheses equation.
- Where x is nothing but the **years**, and the $h_{\theta}(X)$ is the **prediction value** for our hypotheses.
- Once we done finding θ_0 and θ_1 we can find any value.
- Keep in mind we first find the θ_0 and θ_1 for our **training data**.
- Later we will use these θ_0 and θ_1 values to do **prediction for test data**.

Don't think too much about how to find θ_0 and θ_1 values, in **coming posts** i will explain how we can find θ_0 and θ_1 values with nice example and i will explain the **coding** part too.

Thank's for Reading!!!!

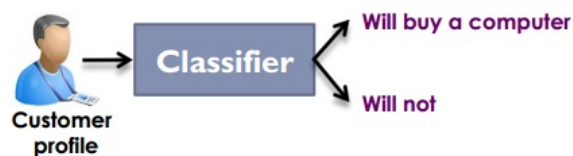
Don't miss any post like our Facebook page **below**

[Leave a comment](#)



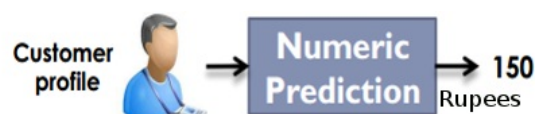
CLASSIFICATION AND PREDICTION

Posted on [September 27, 2014](#) under [DATAMINING](#)



Classification

dataaspirant.wordpress.com



Prediction

Formal Classification and prediction

Definition:

- Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.
- Such analysis can help to provide us with a better understanding of the data at large.
- classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions.

Let's Understand Classification a morsel more:

- The goal of data classification is to organize and categorize data in distinct classes.
- A model is first created based on the data distribution.
- The model is then used to classify new data.
- Given the model, a class can be predicted for new data.
- In general way of saying classification is for discrete and nominal values.

Let's Understand Prediction a morsel more:

- The goal of prediction is to forecast or deduce the value of an attribute based on values of other attributes.
- A model is first created based on the data distribution.
- The model is then used to predict future or unknown values.

Summarization of Classification and Prediction:

- If forecasting **discrete** value (Classification)
- If forecasting **continuous** value (Prediction)

Understanding Classification and prediction in DataAspirant way:



Classification:

- Suppose from your past data (**train data**) you came to know that your best friend liked above movies.
- Now one new movie (**test data**) released and hopefully you want know your best friend like it or not.
- If you strongly conformed about chances of liking that movie by your friend, you can take your friend to movie this weekend.
- If you clearly observe the problem it is just about your friend **like or not**.
- Finding solution to this type of problems is called as **classification** this is because we are classifying the things to there belongings (**yes or no, like or dislike**)
- Keep in mind here we are forecasting **discrete** value(classification) and the other thing this classification in belongs to **Supervised learning**.
- This is because you are **learning** this from your **train data**.
- Mostly classification is **binary classification** means we have to predict is our output belongs to class 1 or class 2 (**class 1 : yes, class 2: no**)

- We can use classification for predicting more classes too. Like
(suppose colors: RED, GREEN, BLUE, YELLOW, ORANGE)

Prediction:

- Suppose from your past data (**train data**) you came to know that your best friend liked above movies and you also know how many times each particular movie was seen by your friend.
- Now one new movie (**test data**) released same like above, now your are going to find how many times this present newly released movie will your friends see is it , 5 times, 6 times, 10 times anything.
- If you clearly observe the problem it is about finding the **count**, some times we can say this as **predicting the value**.
- keep in mind here we are forecasting **continuous value** (Prediction) and the other thing this prediction is also belongs to **Supervised learning**.
- This is because you are **learning** this from you **train data**.

THANK'S FOR READING

Leave a comment



SUPERVISED AND UNSUPERVISED LEARNING

Posted on September 19, 2014 under DATAMINING

Supervised Learning



Unsupervised Learning





Wiki Supervised Learning Definition

Supervised learning is the Data mining task of inferring a function from **labeled training data**. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the **supervisory signal**). A **supervised learning algorithm** analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for **unseen instances**. This requires the learning algorithm to generalize from the training data to unseen situations in a “reasonable” way.

Wiki Unsupervised Learning Definition

In Data mining, the problem of **unsupervised learning** is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution.

Let's learn supervised and unsupervised learning with an real life example





- suppose you had a basket and it is filled with some different kinds of fruits, your task is to arrange them as groups.
- For understanding let me clear the names of the fruits in our basket.
- we are having four types of fruits they are



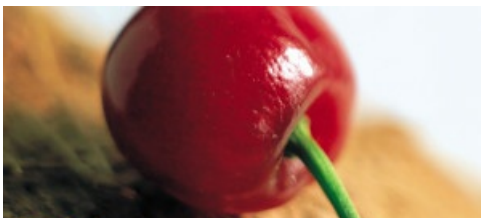
APPLE



BANANA



GRAPE



Supervised Learning :

- You already learn from your previous work about the physical characters of fruits.
- So arranging the same type of fruits at one place is easy now.
- Your previous work is called as **training data** in data mining.
- so you already learn the things from your train data, this is because of **response variable**.
- Response variable mean just a **decision variable**.
- You can observe response variable below (**FRUIT NAME**) .

No.	SIZE	COLOR	SHAPE	FRUIT NAME
1	Big	Red	Rounded shape with a depression at the top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry
3	Big	Green	Long curving cylinder	Banana
4	Small	Green	Round to oval,Bunch shape Cylindrical	Grape

- Suppose you have taken an new fruit from the basket then you will see the size , color and shape of that particular fruit.
- If size is Big , color is Red , shape is rounded shape with a depression at the top, you will conform the fruit name as apple and you will put in apple group.
- Likewise for other fruits also.
- Job of groping fruits was done and happy ending.
- You can observe in the table that a column was labeled as “**FRUIT NAME**” this is called as response variable.
- If you learn the thing before from training data and then applying that knowledge to the test data(for new fruit), This type of learning is called as **Supervised Learning**.

- **Classification** come under Supervised learning.

Unsupervised Learning :

- suppose you had a basket and it is filled with some different types fruits, your task is to arrange them as groups.
- This time you don't know any thing about that fruits, honestly saying this is the first time you have seen them.
- so how will you arrange them.
- What will you do first???
- You will take a fruit and you will arrange them by considering physical character of that particular fruit. suppose you have considered color.
- Then you will arrange them on considering base condition as **color**.
- Then the groups will be some thing like this.
- RED COLOR GROUP: apples & cherry fruits.
- GREEN COLOR GROUP: bananas & grapes.
- so now you will take another physical character such as **size** .
- RED COLOR AND BIG SIZE: apple.
- RED COLOR AND SMALL SIZE: cherry fruits.
- GREEN COLOR AND BIG SIZE: bananas.
- GREEN COLOR AND SMALL SIZE: grapes.
- job done happy ending.
- Here you didn't know learn any thing before ,means no train data and no response variable.
- This type of learning is know unsupervised learning.
- clustering comes under unsupervised learning.



DATA MINING.....!

Posted on September 16, 2014 under Uncategorized



Ancient story about Datamining

In the 1960s, statisticians used terms like “Data Fishing” or “Data Dredging” to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term “Data Mining” appeared around **1990** in the database community.

Data mining in Technical words

Data mining is a process of extracting specific information from data and presenting relevant and usable information that can be used to solve problems. There are different kinds of services in this process like text mining, web mining, audio and video mining, pictorial data mining and social

network data mining.

Why Data mining is hot cake Topic for this generation

Data mining is young and promising field for present generation this is because of its spacious applications. In general way of saying it has an attracted a great deal of attention in the information industry and in society, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. This is the reason why data mining is also called as **knowledge discovery from data**.

Understanding of data mining with buying apples example



Before going to explain data mining with this fresh apples, let me say some interesting facts about apples.

Nutritions: According to the United States Department of Agriculture, a typical apple serving weighs **242 grams and contains 126 calories** with

significant dietary fiber and modest vitamin C content, with otherwise a generally low content of essential nutrients.

Toxicity of apple seeds: The seeds of apples contain small amounts of amygdalin, a sugar and cyanide compound known as a cyanogenic glycoside. Ingesting small amounts of apple seeds will cause no ill effects, but in extremely large doses can cause adverse reactions. There is only one known case of fatal cyanide poisoning from apple seeds; in this case the individual chewed and swallowed one cup of seeds. It may take several hours before the poison takes effect, as cyanogenic glycosides must be hydrolyzed before the cyanide ion is released.

Now we will step into our example.

Suppose your family members want to meet some one how is suffering from pancreatic cancer. we all know that the consumption of apples could help to reduce pancreatic cancer by up to 23 per cent. so your father asked you to bring apples from shop which is near to your house. Also your father **learn you** how to buy apples by give some **set of rules**.

Rules for buying apples

- Big size apples are having less taste than small size apples.
- Dark red apples are not fresh ones.
- Light red apples are fresh ones.
- Green apples are good for health.

On seeing this list of rules you can pick the apples which you want to buy, Your family members want to give this apples to an unhealthy person so you will pick green apples mostly. So when you went to shop you will pick small size apples which are in green color. End of the story to selecting apples which are good for health.

Non Data mining Algorithm

```
selecting apples Algorithm
if( selected_apple == small (in size ))
{
    if(selected apple == green ( in color ) ){
```

```
select apple
}

else {
    don't select apple
}

}
```

Comparing with data mining

- You will randomly select an apple from the shop (**training data**)
- make a table of all the physical characteristics of each apple, like color, size(**features**)
- tasty apples, apple which are good for health(**output variables**)
- If you went to other shop and buy the apples (**test data**)

you can now buy apples with great confidence, without worrying about the details of how to choose the best apples. And what's more, you can make your algorithm improve over time (**reinforcement learning**), so that it will improve its accuracy as it reads more training data, and modifies itself when it makes a wrong prediction. But the best part is, you can use the same algorithm to train different models, one each for predicting the quality of apples, oranges, bananas, grapes, cherries and watermelons, and keep all your loved ones happy.

This type of learning is called as **supervised learning** in data mining ,in next post i will give you clear picture of difference between supervised learning and unsupervised learning with real life examples.

THANK'S FOR READING

[Leave a comment](#)



DATAASPIRANT



DataAspirant

पसंद करें

84 लोगों को DataAspirant पसंद है.



Facebook सामाजिक प्लग-इन

RECENT POSTS

[Linear Regression](#)

[Classification and Prediction](#)

[Supervised and Unsupervised learning](#)

[DATA MINING.....!](#)

FOLLOW DATAASPIRANT

Enter your email address to follow this blog and receive notifications of new posts by email.

Follow US

ARCHIVES


[October 2014](#)

[September 2014](#)

RECENT COMMENTS

[Create a free website or blog at WordPress.com.](#) | [The Book Lite Theme.](#)

2

 Follow

Follow
"dataaspirant"

Get every new post
delivered to your Inbox.

Enter your email address

Sign me up

Powered by WordPress.com