Data Science Tutorial 4: Machine Learning

While linear regression

1 Health in the US

We will now use machine learning to match three variables to a training variable. This dataset is contained in the dataset folder or can be downloaded here:

```
http://asecuritysite.com/log/df.csv
```

Example analysis is shown at:

http://asecuritysite.com/bigdata/col?file=df.csv

and you can analyse the training model here:

http://asecuritysite.com/bigdata/ml?file=df.csv

First use **nano** and create a file named data04.py, and add the following lines:

```
import numpy as np
import pandas as pd
import sys

x1="Infant MR"
x2="Heart Disease DR"
x3="Suicide DR"
x4="Cancer DR"

fdata="df.csv"

print "Training data:\t\t",x1,",",x2,",",x3
print "Training against:\t",x4
print "Data set:\t\t",fdata

print "======================"

ver=pd.read_csv(fdata)

from sklearn.cross_validation import train_test_split
from sklearn.ensemble import RandomForestRegressor

type1= ver[x1].dtype
type2= ver[x2].dtype
type3= ver[x3].dtype
type4= ver[x4].dtype

if (type1==object or type2==object or type3==object or type4==object):
        print "One of the data values is an object"
        sys.exit(1)

train, test, y_train, y_test =
train_test_split(ver[[x1,x2,x3]],ver[x4],test_size=0.5, random_state=1)
ind = ver.columns[0]
```

```
model= RandomForestRegressor()

model.fit(train,y_train)

predictions =model.predict(ver[[x1,x2,x3]])

success=0
failure=0
r = float(float(ver[x4].max())-float(ver[x4].min()))
print "Range of values:\t",r
limit=r/5
print "Success limit:\t\t",limit
print "=============================================\n"

c=len(predictions)
print ('%22s %8s %8s %8s %8s' % ("Index","Pred","Actual","Diff","Success"))
print "=============================================="


for x in range(0,c):
        error = abs(predictions[x]-ver[x4][x])

        if (error<=limit):
                str = "Success"
                success=success+1
        else:
                str="Failed!"
                failure = failure+1
        print('%22s %8.2f %8.2f %8.2f %8s' %
(ver[ind][x][:22],predictions[x],ver[x4][x],error,str) )

print ('Success: %3d Fail: %3d' % (success,failure))

print "\n\n\nTraining data:"

print train
print "Training data (y):"

print y_train
```

In this case we are taking 30% of the data, and taking Infant MR, Health Disease DR and Suicide DR, and training against the Cancer DR. The success is +/-20% of the range of values.

Outline five successful values from your run:

Outline which values were not correctly predicted:

Now run training models for the following, and determine the success rate:

Parameter to be trained against	Training variable 1	Training variable 2	Training variable 3	Success rate (%)
Infant MR	Heart Disease DR	Stroke DR	Suicide DR	
Homicide DR	Drug Poisoning DR	Motor Vech DR	Infant MR	
Cancer DR	Heart Disease DR	Stroke DR	Suicide DR	
Suicide DR	Drug Poisoning DR	Motor Vech DR	Infant MR	

2 Alcohol in Scotland

This dataset is contained in the dataset folder or can be downloaded here:

```
http://asecuritysite.com/log/alc.csv
```

A sample of the data is here: http://asecuritysite.com/bigdata/col?file=alc.csv

Using Pandas, determine the correlation between the following:

	Alcohol-related hospital stays	Alcohol-related mortality	Weekly drinkers (pupils age 15)	Common assault	Breach of the Peace
Alcohol-related hospital stays					
Alcohol-related mortality					
Weekly drinkers (pupils age 15)					
Child protection with parental alcohol misuse					
Common assault					
Vandalism					
Breach of the Peace					

Next we will use machine learning to analyse the data for the following:

Parameter to be trained against	Training variable 1	Training variable 2	Training variable 3	Success rate (%)

Alcohol-related hospital stays	Alcohol-related mortality	Weekly drinkers (pupils age 15)	Child protection with parental alcohol misuse	
Alcohol-related hospital stays	Attempted murder & Serious assault	Common assault	Vandalism	
Alcohol-related hospital stays	Breach of the Peace	% people perceiving rowdy behaviour very/fairly	Alcohol treatment waiting times	
Alcohol-related hospital stays	Alcohol-related mortality	Common assault	Breach of the Peace	
Alcohol-related hospital stays	Premise licences in force - On trade	Premise licences in force - Off trade	Breach of the Peace	

Which model fits best, and why do you think this is the case?