G12-MDM-DataScienceAndTheFoundationsOfScience

December 16, 2019

1 Data Science and the Foundations of Science

This is the summary of our class on Thu. Dec 12 2019 on G12 Mathematics of Data Management.

As we mentioned at the beginning of our course, data management is but a proxy for what nowadays is called Data Science.

Currently Data Science is an extremely hot topic in our society, both because it's a hot job market and because companies are eager to tap into the potential information hidden in the vast amounts of data gathered in our digital world.

Today we asked the question of the relation of Data Science to Science. More specifically:

- **Is Science simply but Data Science?** This is the idea that scientist need to do experiments and analyze data in order to obtain the *right* model and reach an understading. Usual claims in this sense start like *'Statistics has shown that...'*

- **Does Statistics and its methods provide the answers to all our questions?**
- **Does the data contain all the information there is to be found?**

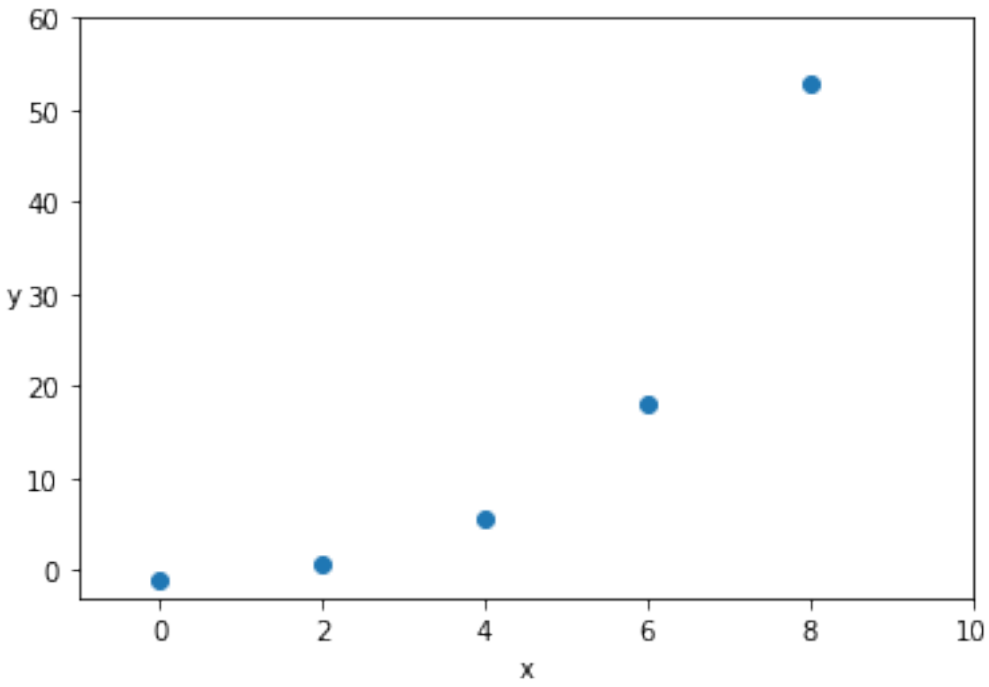
We will see that the answer to these three questions is a conclusive **NO**.

2 Toy example: Fitting a model to some data

[50]:
```
import matplotlib.pyplot as plt
import numpy as np
import scipy.optimize as spopt
```

[51]:
```
x =np.array([0, 2, 4, 6, 8])
y = np.array([-1, 3/4, 11/2, 18, 53])
plt.plot(x,y, 'o')
plt.xlim(-1,10)
plt.ylim(-3,60)
plt.ylabel('y',rotation=0)
plt.xlabel('x')
print("Figure1: We gathered some data that follows a yet unknown relation \
and would like to find a 'good' model of it. \
We try four different models. See below for a plot and their corresponding R^2␣
 ↪values.\
Which model describes best the underlying relation between x and y?")
```

Figure1: We gathered some data that follows a yet unknown relation and would like to find a 'good' model of it. We try four different models. See below for a plot and their corresponding R^2 values.Which model describes best the underlying relation between x and y?



[52]:
```
#Model 1: Linear: y = a0 + a1 * x
def y1(x,a0,a1):
    return a1*x+a0

#Model 2: Quadratic: y = a0 + a1 * x + a2 * x^2
def y2(x,a0,a1,a2):
    return a2*x**2+a1*x+a0

#model 3: Exponential :   y  = a0 + a1 * e^(a2*x)
def y3(x,a0,a1,a2):
    return a0+a1*np.exp(a2*x)

#model 4: Quartic:        y = a0 + a1 * x + a2 * x^2 + a3 * x^3 + a4 * x^4
def y4(x,a0,a1,a2,a3,a4):
    return a4*pow(x,4)+a3*pow(x,3)+a2*x**2+a1*x+a0
```

[58]:
```
#Starting point for the optimization process to find the coefficients
p1=(1,1)
p2=(1,1,1)
p3=p2
p4=(1,1,1,1,1)
```

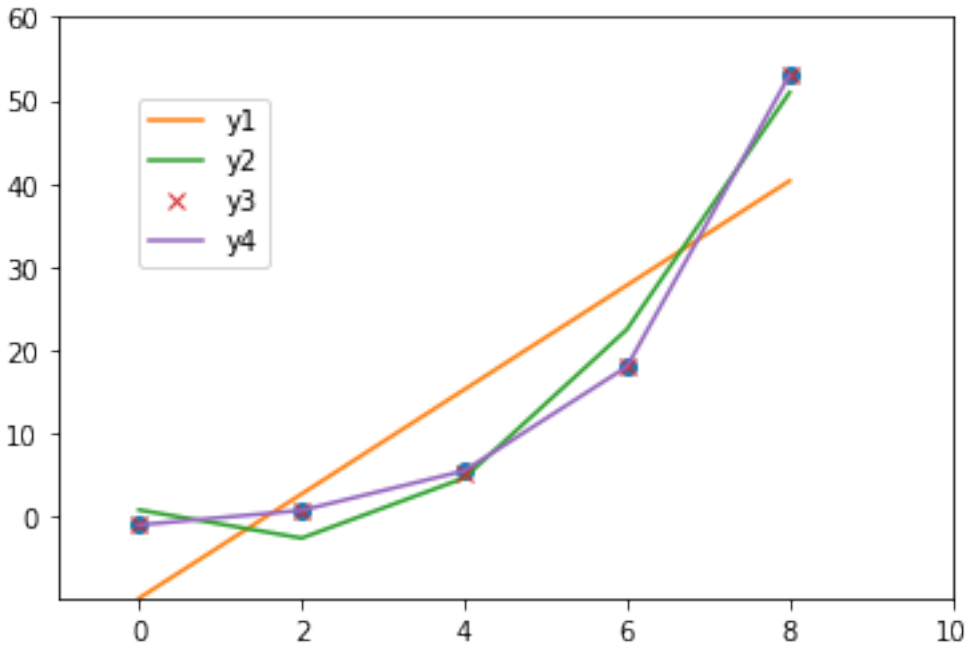
```
'''
opt == the values of the coefficients after the model fit
cov == estimated covariances among those coefficients
'''
opt1, cov1 = spopt.curve_fit(y1,x,y,p0=p1)
opt2, cov2 = spopt.curve_fit(y2,x,y,p0=p2)
opt3, cov3 = spopt.curve_fit(y3,x,y,p0=p3)
opt4, cov4 = spopt.curve_fit(y4,x,y,p0=p4)

plt.plot(x,y,'o')
plt.plot(x,y1(x,*opt1),label='y1')
plt.plot(x,y2(x,*opt2),label='y2')
plt.plot(x,y3(x,*opt3),'x',label='y3')
plt.plot(x,y4(x,*opt4),label='y4')
plt.legend(loc=(1/11,4/7))
plt.xlim(-1,10)
plt.ylim(opt1[0],60)

def r2(x,y,f,par):
    se=st=0
    ym = y.mean()
    for i in range(len(x)):
        xi = x[i]
        yi = y[i]
        yhat = f(xi,*par)
        se += (yhat - ym)**2
        st += (yi - ym)**2
    return se/st

print("Model\t\tR^2\t\t\tParameters [a0,a1,...]:")
print("y1: Linear\t",r2(x,y,y1,opt1),opt1)
print("y2: Quadratic\t",r2(x,y,y2,opt2),opt2)
print("y3: Exponential\t",r2(x,y,y3,opt3),opt3)
print("y4: Quartic\t",r2(x,y,y4,opt4),opt4)
```

```
Model           R^2                     Parameters [a0,a1,...]:
y1: Linear       0.7835945304724022 [-9.8     6.2625]
y2: Quadratic    0.9802929213630669 [ 0.80714286 -4.34464286  1.32589286]
y3: Exponential  0.9999866477108664 [-1.90218251  0.96440218  0.50519354]
y4: Quartic      1.0000000000000002 [-1.         -0.33333333  0.92708333
-0.21354167  0.02604167]
```



3 Discussion

If we follow just the bare data and statistics we calculated, the best fit is that of a quartic model with $R^2 = 1.000000$.

The second best fit to the data is that of the exponential with $R^2 = 0.999987$.

Clearly, that's a tie: the difference between both R^2 is meaningless given the size of the data (just 5 points).

We could also say that the quadratic model provides quite good a fit explaining 98% of the variability in y.

While the linear model is the worse of all four, notice it is still explaining slightly over 78% of the variability in the data!

3.0.1 Questions

We see then that these statistics still leave some **questions open**:

Does this mean, the ***law*** that governs these data is a quartic polynomial?

Do we have any criterion to choose between the quartic model and the exponential one?

3.0.2 Overfitting and Underfitting

The quartic model requires 5 parameters. The data consists in five points. Clearly, if we have to choose five arbitrary parameters to explain five points, we may very well simply list those points as a "model". Hence, the ***quartic model is not informative*** -despite its perfect R^2 score!

To provide some perspective let's consider the Standard Model of Physics. This model explains all Electromagnetic, radioactive and nuclear phenomena -everything except gravity. In

other words, it explains why your tiles look white, why they are solid at room temperature, why the sky look blue or redish, why water boils at 100^oC, how a computer works, how your wifi works,... In summary, it basically explains almost all of your common live physical experiences. Yet, ***the Standard Model contains only 19 free parameters to adjust***!

In this sense the Standar Model is way more informative in relation to natural phenomena than this quartic model is in relation to those five data points.

We say that the quartic model is *OVERFITTING* the data.

In the other extreme lies our linear model. It only requires two parameters, but clearly it's too *coarse grained* thus missing to reproduce many details of the data.

We say tha the linear model is *UNDERFITTING* the data.

4 Conclusions: Occam's Razor

The data was generated using an exponential relation $y = -2 + e^{\frac{x}{2}}$ and rounding off the values obtained. The rounding step plays the role of noise in the data

Hence, it is clear that **no amount of data can reveal the *true* (whatever that may mean) underlying law explaining the data**.

There is no such thing as ***true explanation***, only models that fit better with our current knowledge we have so far about nature, including the data we want to explain as well as other models explaining other phenomena.

One guiding principle is thus **Occam's razor**: Among competing models with equal explicative power, we must choose the one requiring the least number of assumptions (parameters).

Hence, the art of Science is like building a house of cards or, if you want, like Jenga, distilling out the slimmest tower that still stands.

There is no Science with just data. And, clearly, there is no Science with just model building in our heads. **We need both, the data AND our imagination for building mathematical models *and* theories!**

Science is thus the result of a delicate balance of intertwining data and models.

5 Assignment

1. Consider the following set of 20 data points. Using Desmos or any similar software, find what you think would be a reasonably good model for it. Provide the expression of your model, the value of the parameters obtained from the regression and the value of R^2.

t	Temp
2	-1
3	1
4	5
5	11
6	17
7	24
8	30
9	35
10	39

t	Temp
11	40
12	40
13	38
14	34
15	28
16	22
17	15
18	9
19	4

2. Now you are told some additional information, namely, that the previous data represents seasonal changes in temperature along some period of time. How does your best model change? Provide the expression of your model, the value of the parameters obtained from the regression and the value of R^2.

[ ]: