IntroductionAlgorithmsLecture

September 21, 2020

1 Algorithms and Machine Learning

1.1 Target function

One of the most imporant decision we have to make as data scientists is to define what constitutes success.

This is an important concept that often gets overlooked. Results depend on our measure of success, and we should be informed about the choices that we made or the function that we are using has made for us.

Today, we are going to discuss these concepts with an example with real data.

As usual, we start by loading the Python packages that we are going to use in the analysis.

[1]:
```
import numpy as np #Package for numerical multidimensional tables
import pandas as pd #Package for ...
import matplotlib.pyplot as plt #Package for ...
```

Over the years, I have had a discussion with colleagues about if Quentin Tarantino is a good director or merely over-hyped.

To give the discussion some objective dimension, we decided to look at the ratings from Tarantino's movies through the years. The data comes from Rotten Tomatoes and it is saved in the 'RT_Tarantino.csv' file

We first load the data using a pandas data frame. And we print it.

[ ]:
```

```

The dataset inclues ratings, title, year of release, and license cost for all Tarantino movies with a rating in Rotten Tomatoes. The data is presented in descending order through the years.

The simplest way to decide if Tarantino is a good director is to look at the rating of his latest movie.

[ ]:
```

```

This is certainly a valid measurement but it has some drawbacks.

In particular, our whole decision depends on only the latest observation.

Thus, it could change widly from year to year. For instance, we could get a completely different result if we had this discussion in 2008.

[ ]:

As before, we look at the last rating available.

[ ]:

Instead of looking at the last rating we could look at the average rating for the last year.

[ ]:

Or some other year of your choice.

[ ]:

Alternatively, we could consider the average for all his movies.

[ ]:

Considering all the data makes the measurement fairer in the sense that does not depent too much on anyone observation.

Nonetheless, we may miss some important dynamics of his movies. Like, maybe Tarantino is directing better-rated movies now that when he started filming.

As with most methods in statistical analysis, there are trade-offs to all decisions.

1.2 Fitting a Model

Looking at the results through the years, there seems to be an increasing trend in the ratings.

We can plot the data to *see* if our intention is correct.

[ ]:

There does seems to be a slightly increasing trend on the ratings. Notice how the last movies have obtained hight ratings.

We may be able to model this trend using a simple line.

Thus, we try to *fit* the model

$$Rating_i = \alpha + \beta Year_i + \epsilon_i,$$

where α and β are selected according to some fit criteria.

In the linear model above, the ϵ_i allows us to capture random variations in the data that may not be explained using the linear fit.

Thus, one intuitive way to select α and β is to make them minimize the unexplained variations, ϵ_i.

Let us call $\widehat{Rating}_i$ to the rating assigned by the linear model. We may thus look to minimize

$$\min_{\alpha,\beta} \epsilon_i = \min_{\alpha,\beta} Rating_i - \widehat{Rating}_i = \min_{\alpha,\beta} Rating_i - \alpha - \beta Year_i$$

There are two possible drawbacks with the above problem statement: * It is only based on one observation * It will get us undetermined solutions

Why?

- As was the case before with the simple mean, making our decision looking at just one observation may result in wildly different results depending on which observation we are looking at. So instead, we could take some certain combination of more values.
- Notice that if we make α smaller and smaller we get smaller values for ϵ_i.

1.3 Loss Function

Thus, we must decide on a proper **loss function**.

Our choice regarding the number and characteristics of the observations used, and the function used to weight them, produce different results.

The classical linear regression proposes the loss function:

$$\mathcal{L}(\alpha, \beta) = \sum_{i=1}^{N} (Rating_i - \alpha - \beta Year_i)^2,$$

where N is the number of observations used in the computations.

Notice that nothing forces us to choose the square as the weighting function, we could have used absolute value or a higher degree polynomial.

Once we have selected our fit criteria (or loss function), we should decide how are we going to solve it.

In this case, solving it translates to mininize the loss function.

There are several options to look for a solution depending on the complexity of the problem at hand: * Try different combinations of α and β and select the one with the smallest loss. * Start on a (possibly random) combination and use the data itself to point us to a better combination. * Solve the problem analytically.

1.4 Grid Search

We start with the more intuitive way to look for a solution: grid search.

The idea behind grid search is to set up a list or *grid* of values and then *search* for the optimal one.

That is, we set a list of possible values for α and β, evaluate the loss function on them, and select the one that achieves the smallest values

[ ]:

[ ]:

[ ]:

We plot the fitted model using the estimated parameters.

[ ]:

We can try more values easily dy defining a function.

[ ]:

We can then make a list as big as we want to make the grid search.

[ ]:

And we plot the line using the new estimates.

[ ]:

1.5 Gradient Descent

Another possible solution to our minimization algorithm is to start at a (possibly random) value and let the data itself tell us how to improve the fit.

Mathematically, we know that the function decreases in the opposite direction of the derivative. This motivates to look for improvement in our loss function in the direction of the derivative.

Thus, starting from some initial values, we update them as follows:

$$\alpha^{new} = \alpha^{old} - \frac{\mathcal{L}(\alpha, \beta)}{\partial \mathcal{L}(\alpha, \beta)/\partial \alpha},$$
$$\beta^{new} = \beta^{old} - \frac{\mathcal{L}(\alpha, \beta)}{\partial \mathcal{L}(\alpha, \beta)/\partial \beta},$$

and we iterate until our parameters do not change (too much).

For our purposes, we have already obtained the derivatives, they are given by:

$$\frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \alpha} = -2 \sum_{i=1}^{N} (Rating_i - \alpha - \beta Year_i),$$
$$\frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \beta} = -2 \sum_{i=1}^{N} (Rating_i - \alpha - \beta Year_i) Year_i,$$

Which we define in Python below.

[ ]:

With the newly defined functions, we can update the values of our estimates.

[ ]:

And we can plot the results with the last updated values.

[ ]:

More efficiently, we can set a *for* loop to update the values.

[ ]:

And we plot the results with the last update.

[ ]:

1.6 Analytical Solution

For some problems, we may be able to obtain the analytical solution with pen and paper.

Turns out, this is one of those examples.

The solution is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(Rating_i - \overline{Rating})(Year_i - \overline{Year})}{\sum_{i=1}^{N}(Year_i - \overline{Year})^2}$$

$$\hat{\alpha} = \overline{Rating} - \hat{\beta}\overline{Year},$$

where the bar on top means average.

We program this into Python to obtain the analytical solutions.

[ ]:

And we plot the resulting model.

[ ]:

Exercise: What about the license cost to watch the movie?

[ ]:

2 Statistical Analysis

The analysis above, gave us some values for the parameters in the linear model we considered.

In particular, we found a positive value for the effect that $Year$ has on $Rating$; thus, it may seem that Tarantino's movies have been better received on average as times goes by.

Yet, we may be interested in knowing if this result is a fluke (something due to chance given that we *randomly* decided to run the experiment *today*), or if there is perhaps some true better reception to his movies.

Hence, we would like to be able to say if the *positive* effect that $Year$ has on $Rating$ is **statistically significant**.

There are two *cultures* regarding how to analyze if a result is significant: * Make a set of assumptions regarding the data and obtain analytical results * Use the data itself to check for the significance of the results

As before, the methods above have advantages and disadvantages.

2.0.1 Normality Assumption

If we are willing to assume that our data follows a specific probability distribution (typically a Normal distribution), then we know that the estimated parameter follows a Normal distribution. In particular

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2_{\beta})$$

Which tells us that our estimate is going to fall around the *true* value.

Now, if the true value is non-zero, that would make it more likely that our estimate is non-zero.

Moreover, the variance (how far our estimate can fall from the true value) decreases with the sample size. Thus, if our data is large enough, we can be quite certain of our results.

Using the formula above, we can then compute the probability that it is non-zero by pure chance.

If that probability is really small, we may be inclined to say that it is not just a random result but a true non-zero value.

For our purposes, the variance of $\hat{\beta}$ can be obtained by

$$\sigma^2_{\beta} = \frac{\mathcal{L}(\hat{\alpha}, \hat{\beta})/(N-2)}{\sum_{i=1}^{N}(Year_i - \overline{Year})^2}$$

[ ]:

We then plot a Normal distribution with mean 0 and variance given above.

[ ]:

We can check graphically if the zero value is a likely result.

[ ]:

As previously noted, this result relies on the Normality assumption.

We should check if the data seems to follow a Normal distribution.

[ ]:

Does it look Normal?

What about $Year$?

[ ]:

Now, even if the data does not follow a Normal distribution, it can be shown that *asymptotically* it approximates it.

So if the sample is big, the analysis above may still be valid.

2.0.2 Empirical Distributions

For certain exercises, and when the data is far from being Normal, we can use the data itself to get some sense of the statistical validity of our results.

Right now, we are producing just one general estimate of the parameters. As we did before, what if we could construct several estimates and study if they differ too much from each other.

To generate more estimates we need new datasets. The data at hand is of course limited, we do not have an infinite number of Tarantino movies to obtain a bunch of parameter estimates.

One way to generate more estimates from the fixed dataset is to randomly select a subset of the data, solve the model and obtain parameter estimates. If we do this enough times, we can construct an empirical distribution of the parameter.

There are two main ways to generate *new* datasets from the data: * Drop some observations and estimate the model on the rest. We then repeat the experiment dropping another groups of observations. We typically drop just one observation, thus the name **leave-one-out**. * Randomly select a sample of the same size as the original by selection with replacement from the original dataset. The method is called **bootstrap**.

Leave-one-out We start with leave-one-out. We can sequenttially drop one observation, make the analysis in the rest of the data and store the parameter estimates.

[ ]:

[ ]:

We obtain the mean and variance of the estimates.

[ ]:

Bootstrap

[ ]:

[ ]:

We obtain the mean and variance of the estimates.

[ ]:

[ ]: