Data inspection

I decided to begin with exploring into the dictionary. I put all the numerical features on a graph (x is the persons). In order to distinguish between ‘poi’ and ‘non-poi’ I sorted the data to put all ‘poi’s in the beginning of the graph.

Of course in order to see all the features together I had to do scaling. At first I did a simple min-max scaling, but there was not much I could see, and it looked very noisy:

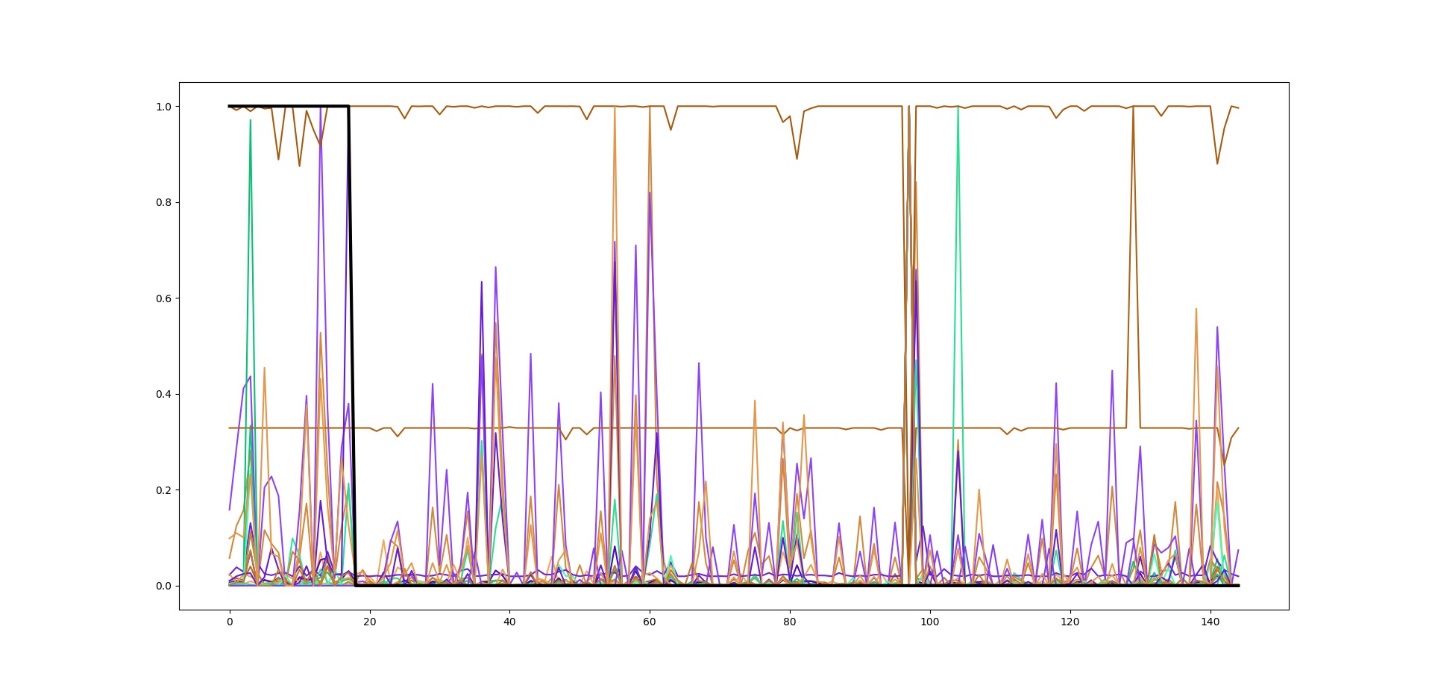


Figure 1.

Features graph, min-max scaled.

The black thick line represents the ‘poi’ feature – or actually the labels.

It’s also possible to see that in case there is an outlier point in some feature, it badly effects the dynamic range of the graph, so nothing can be seen. Furthermore I realized that the actual values of the scaled data is not interesting, and what I really want to see is the “activity” over the feature, which means the differences in values of feature between different people.

So I changed the scaler to use a normal scaler, which means that all the graphs’ means are zero, and standard deviations (and variances) are ones.

This was the picture that I got:

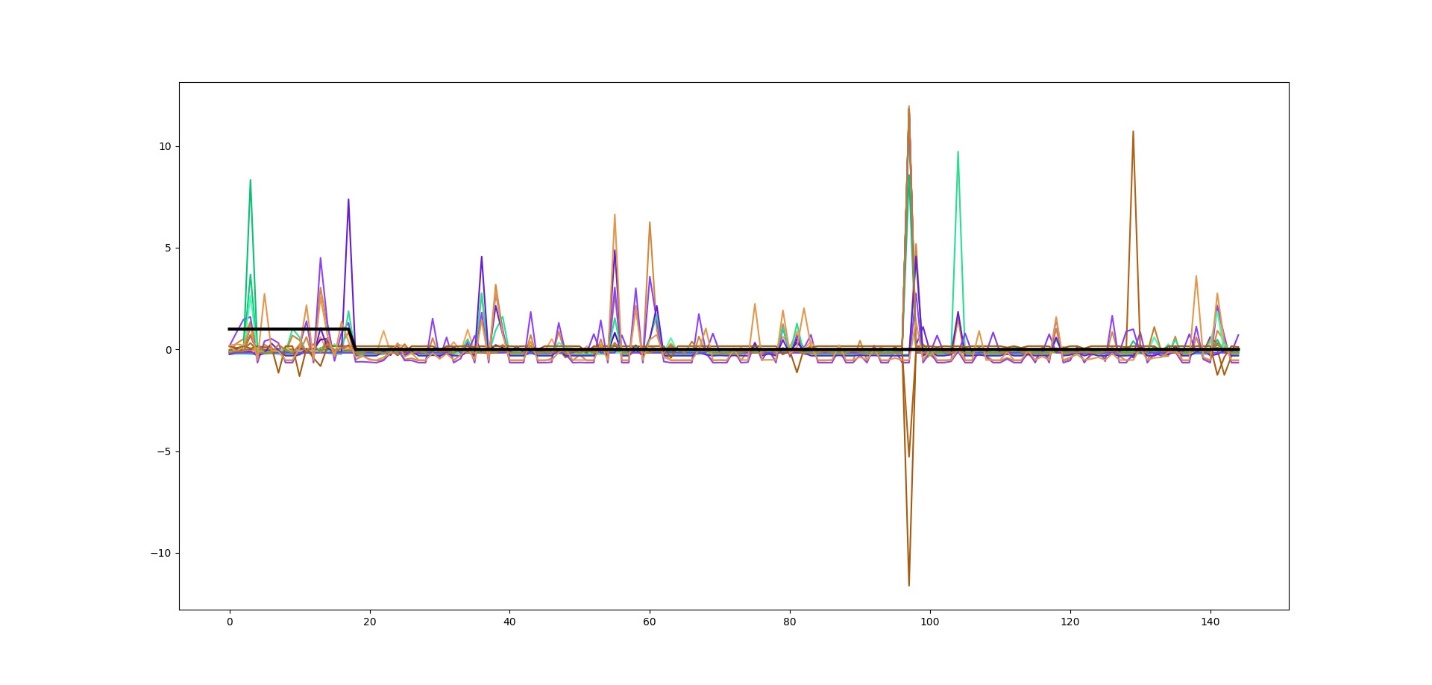


Figure 2.

Features graph, normally scaled.

The awkward phenomenon at the 97’th person immediately reminded me of the terrible outlier we cleaned lesson 7! So I removed the ‘TOTAL’ entry from the dictionary, and that was the new result:

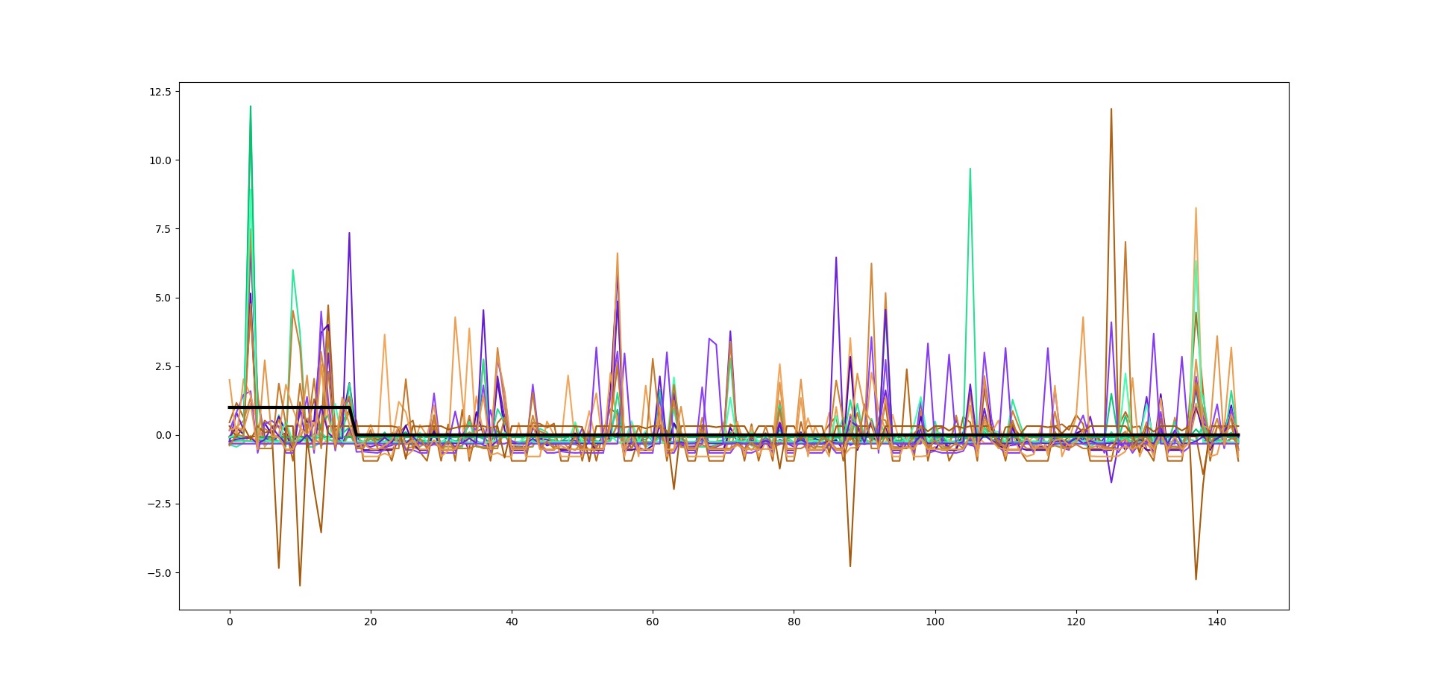


Figure 3.

Features graph, normally scaled. ‘TOTAL’ deleted!

Well, things seems to warm up a little. We can already see some of the information in the features.

It’s now obvious that the normal scaling that I chose is much more effective than the min-max, as it already gave us one very important result (removing the outlier). Still at this point I checked again what do I get using the min-max scaler, after removing ‘TOTAL’. That was the result:

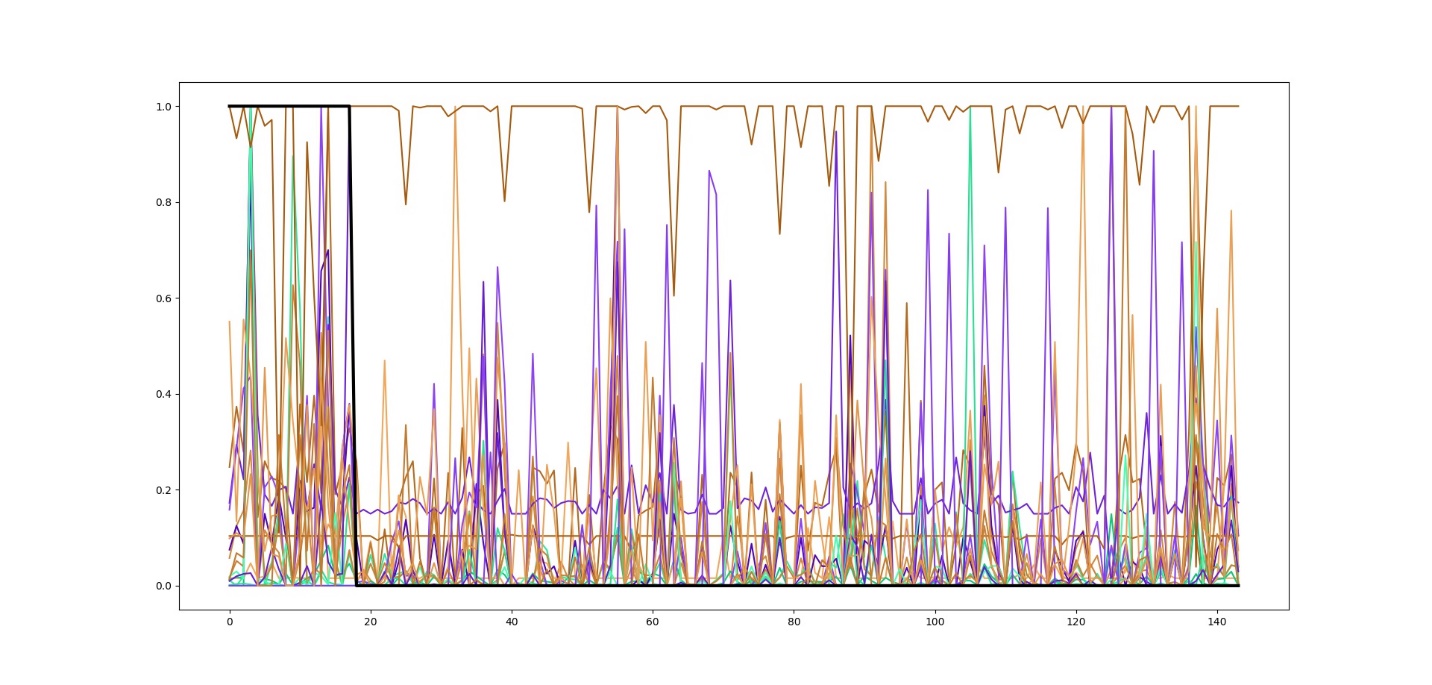


Figure 4.

Features graph, min-max scaled. ‘TOTAL’ deleted!

This graph is indeed better from Figure 1, by several means, but still I decided to keep working with the normal scale.

So let’s look back at figure 3, and particularly at those irregular high picks that are seen there. There are several of them, which are higher than 10 or even getting close to 12. This is very unusual for a data with standard deviation of 1. Of course this can occur in a random data ether because it’s far from being normally distributed or there is simply not enough data. Yet we are usually not expecting to get samples in random data with value grater then 3-4 standard deviation, and certainly not 12!

Let check what that means - one point at the height of 12 in a data of 144 points (like we have) with variance of 1. What will happen to the variance if we take this point out of the data set? Looking at basic moments’ calculations, and under assumption of zero mean (µ=0) the new variance of the other 143 points will be around:

Equation 1.

Variance of data-set after excluding one point x0.

Were:

: Number of samples in original data-set

: The original data-set variance

: Number of samples in original data-set

And the result that we get:

: The variance of the new data-set (without x0)

More details and detailed proof of Equation 1, are presented in Appendix 1.

The meaning of this is that in our conditions, if we take this sample (x0) out of the data, and look on the left 143 points, we will get a standard deviation of about 0!! That means that there is almost no information that we can conclude from this feature for all the other 143 points.

So now there are 2 possibilities for proceeding with this:

1. Ether we throw this feature and not use it (because it gives no information),
2. Or that the irregular point (x0) is an outlier that must be deleted/Ignored/fixed.

Therefore let’s look at all the points which are farther from zero (µ=0) in more than 7 standard deviations… I isolated all the features that has sample values higher (or lower) than 7 (or -7).

And these are the test cases:

|  |  |  |
| --- | --- | --- |
| name | feature | Scaled sample value |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table 1.

Irregular features’ values.

Now, After cleaning some more noise from our data, here is the updated graph:

Figure 5.

Features graph, normally scaled. Irregular samples handled.

As we can see in the graph, although the mean, and the variance of all the features are the same, there are still some features that are “more active” then the others, and so probably contain more information than the others (not sure that this information is correlated with our labels, but it’s worth investigating…).

So in order to

Appendix 1.

Reduced data set variance.

Given a N-length data-set of samples { xi : i = 0..N-1 }, with zero mean: µ=0, and unit variance: σN2 .

Let us look at a the reduced (N-1)-length data-set { xi : i = 1..N-1 }, which is the original data-set with only one point x0 excluded (index 0 is arbitrary, and does not limit the generality).

What is the variance of the new reduced data-set σN-12 ?

Remembering basic statistic class:

So by using the two framed equations from above, we get:

And hence we got Equation 1 ☺.