CS5785 Modern Analytics

Homework 1

September 29, 2014

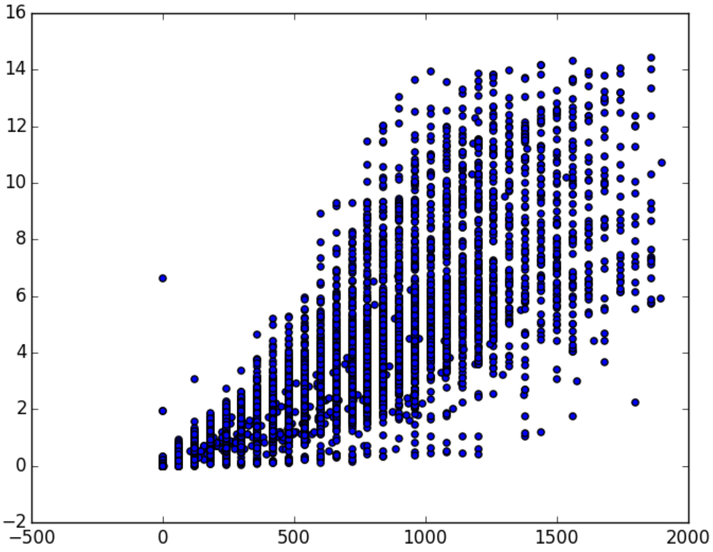
Fanxing Meng

Programming Exercises

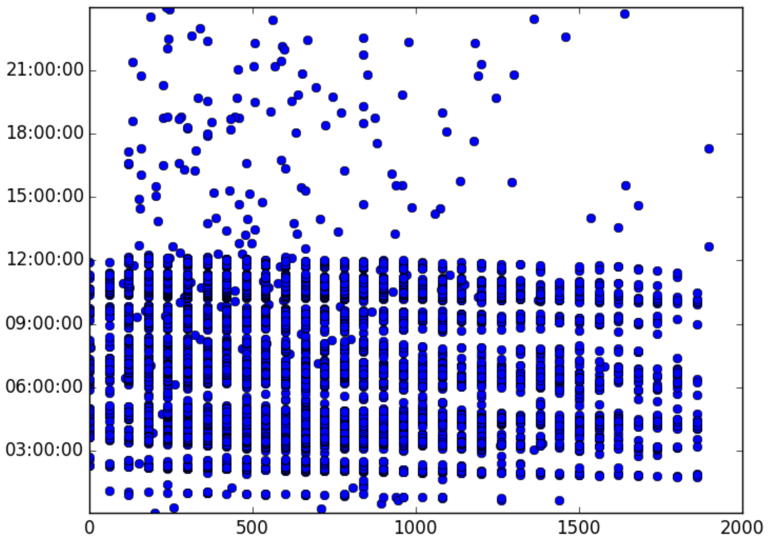
1. Crazy Taxi

1. Minimum: 0, maximum: 5389.36. Outliers are considered to be data points 3 sigma away from the mean, which turns out to be greater than 851.23. Although this distribution is not Gaussian, data points 3 sigma away will constitute at a maximum of 1% (actually 0.27%), which will not significantly decrease the credibility of the data set. In fact, all outliers result from missing one or more coordinates, creating values around 5000.

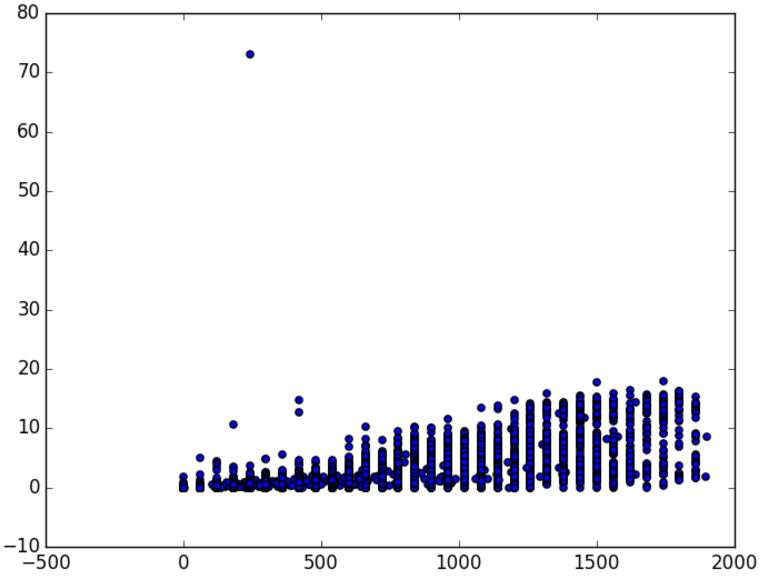
2. Trip distance vs trip time:



Pickup time vs trip time:

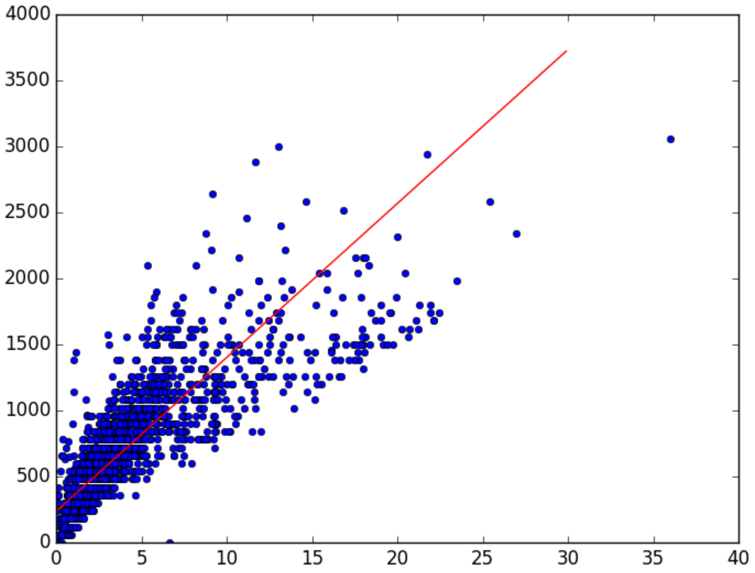


Distance between pickup and dropoff vs trip time:



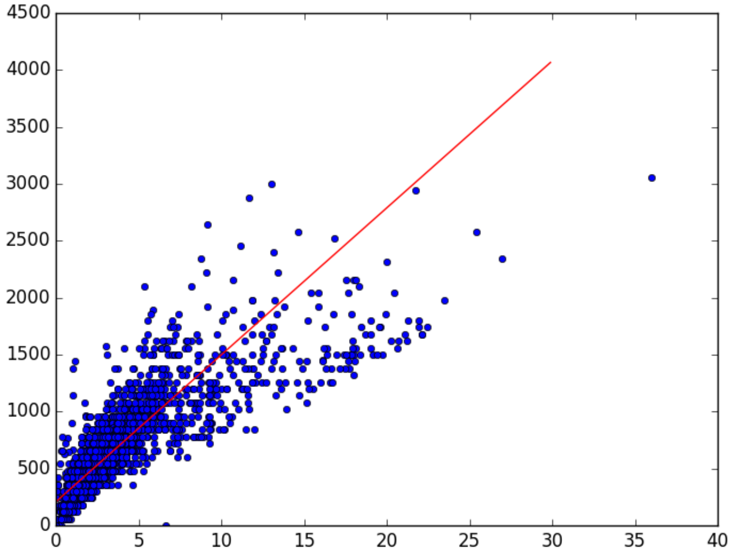
3. Using 3 sigma as point selection criteria, the OLS of the test set is 248.67, while the TLS is 2.13 due to the large slope. If the test data is also cleaned, then OLS will become 200.34, and TLS will be 1.72.

   Below is the overlay diagram showing the fitted line $y = 116.598\ x + 235.641$ in which r = 0.837.



4. By using the 3 sigma selection criteria again on the already cleaned dataset, the OLS and TLS became 268.59 and 2.08, respectively. The OLS worsens but the TLS value improves. Once again, if test data is cleaned, OLS increases to 203.05 while TLS decreases to 1.57.

   Here the line becomes $y = 129.072\ x + 206.975$, with r = 0.836.



5. Even the AWS r3.8xlarge instance is incapable of finishing the run on trip_data_2. Computing the OLS, TLS, and correlation coefficient still need more computation resources.

6. From the graph of pickup time vs trip time, it is natural to divide the time into two zones: 02:00 to 12:00 where data points are dense and shows good uniformity, and 12:00 to 2:00 where data points are sparse and mainly fall on the short travel time side.

7. Normalization of trip distance:

$$nrm(d_i) = \frac{d_i \text{-} d_{avg}}{s_d}$$

8. Same as above, computing the OLS,TLS, and correlation coefficient for the linear system is too much for the AWS machine.

2. The Titanic Disaster

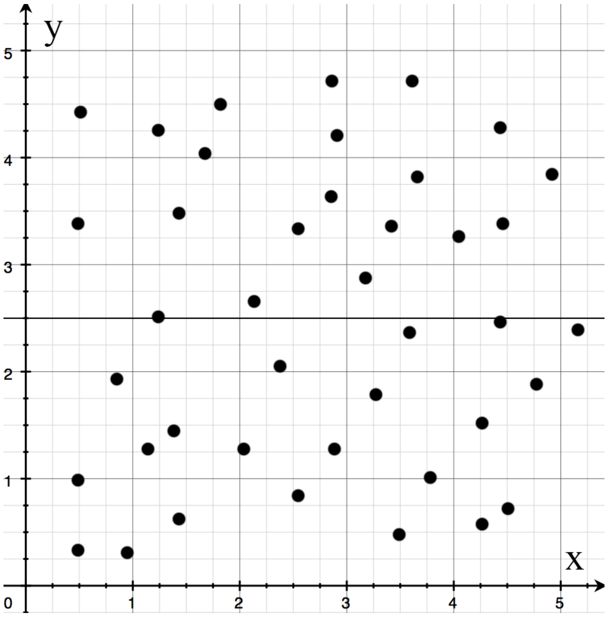
The features chosen in the first step are Pclass, Sex, Age, Sibsp, Parch, Fare, and Embarked, to preserve as much quantitative and categorical information as possible. Data cleaning includes changing the string representation of sex, embarked to integers, changing empty age fields to the mean or median age of all passengers, limiting the fare to a maximum of $50, and limiting the number of Sibsp and Parch to 2 to enhance linearity.

Using LogisticRegression from sklearn.linear_model, the score is 0.75598, while using Logit from statsmodels.api, the score is only 0.73684. Both are not as good as the models mentioned in the tutorial. Further refinement may be achieved based on regularization and parameter selection techniques.

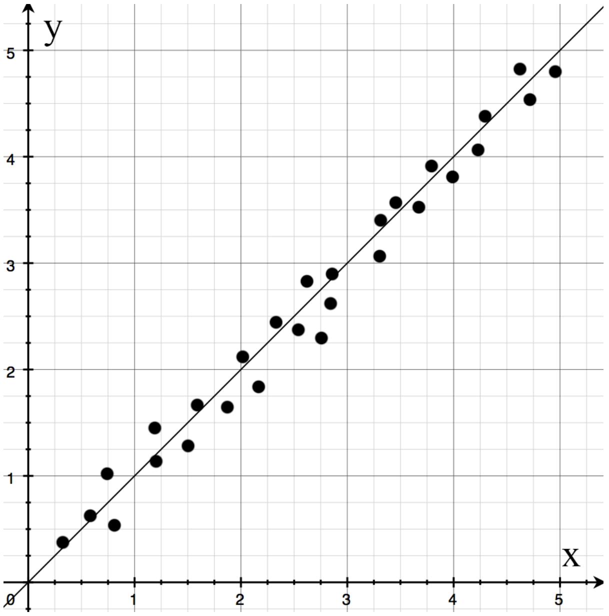
Written Exercises

1. Pearson's correlation coefficient

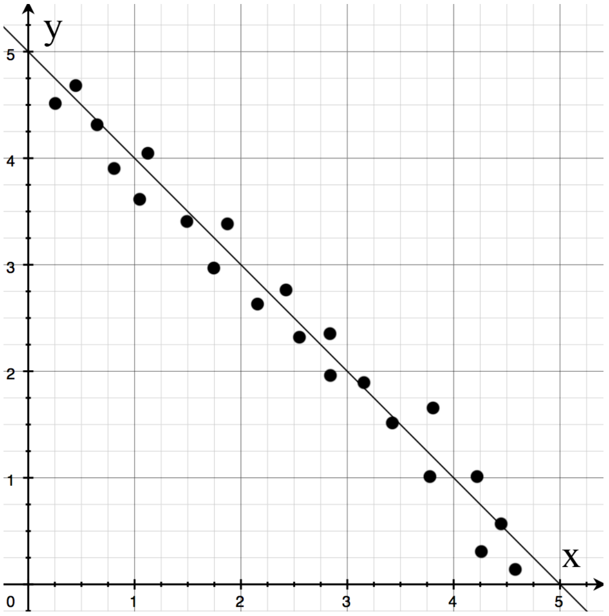
1. Correlation coefficient value close to 0
   r = 0 means no correlation, which can be represented by the evenly distributed points, for which no meaningful linear relationship can be found.



2. Correlation coefficient value close to 1
   r = 1 means positive correlation, which can be represented by the fitted line having positive slope.



3. Correlation coefficient value close to -1
   r = -1 means negative correlation, which can be represented by the fitted line having negative slope.

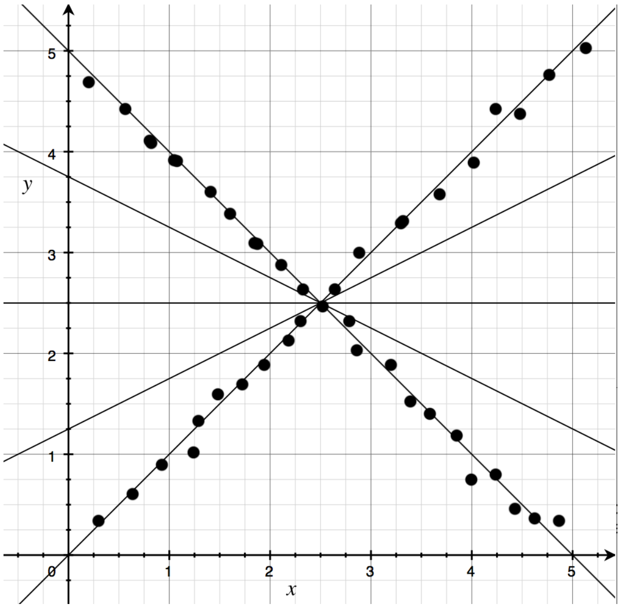


2. L1 and L2 norms

1. Most data along a line with one large outlier: L1 regression will fit most data points well, largely ignoring the single outlier; while L2 regression will tilt significantly towards the outlier, not matching the majority of the data points.

2. If data points are somewhat symmetrical with respect to the horizontal axis, then infinitely many L1 regression will be optimal since the sum of the residual will be the same no matter how the line rotates around the crossing; while L2 regression will result in the horizontal line due to the following inequality

$$x^2 \geq (ax)^2 + [(1\text{-}a)x]^2 \geq 2\times\left(\frac{x}{2}\right)^2, a \in [0,1]$$



3. It should have only one point of minimum instead of an infinite number of possibilities all resulting in the same minimum residual.