Elastic Net Rare Event Prediction Experiment Over SCF Data

José Manuel Navarro

Wednesday, September 17, 2014

Input Data

The model input data are 77 categorical variables indicating the presence (1) or absence (0) of a certain event in a five or thirty minute window before a given time.

Output Data

The output data is a categorical variable that predicts the appearance of a certain event in a five or thirty minute window after a given time.

Motivation and Algorithm

Previous experiments with logistic regression and the same input data yielded accuracy results of up to a 75%, but presented two main problems:

- **A large number of possible features**: there are 77 kinds of events present in the system. Our previous approach to selecting valid features was trying each event's performance one on one. This approach was cumbersome and slow.
- **Testing period didn't usually contain the target event**: due to the scarcity of target events, which were usually less than a 1% of all available instances, most of the times the testing phase was not completed, as no target event appeared on it.

We employed three different techniques to overcome these problems:

1. **Elastic Net Logistic Regression**: a technique proposed by Zou and Hastie [1], the Elastic Net combines lasso and ridge regularization methods in a linear weighted way to exploit both methods' advantages. Summed up, standard logistic regression optimized through Mean Squared Error minimizes the following error function

$$E = \frac{1}{n}\sum_{i=1}^{N}(y - \hat{y})^2 \tag{1}$$

where N is the amount of samples, y is the real output and $\hat{y}$ is the predicted output. Elastic net adds two terms to that equation, pondered by a parameter α. The resulting error function to minimize is

$$E = \frac{1}{n}\sum_{i=1}^{N}(y - \hat{y})^2 - (\alpha\sum_{j=1}^{k} w_i + (1-\alpha)\sum_{j=1}^{k} w_i^2) \tag{2}$$

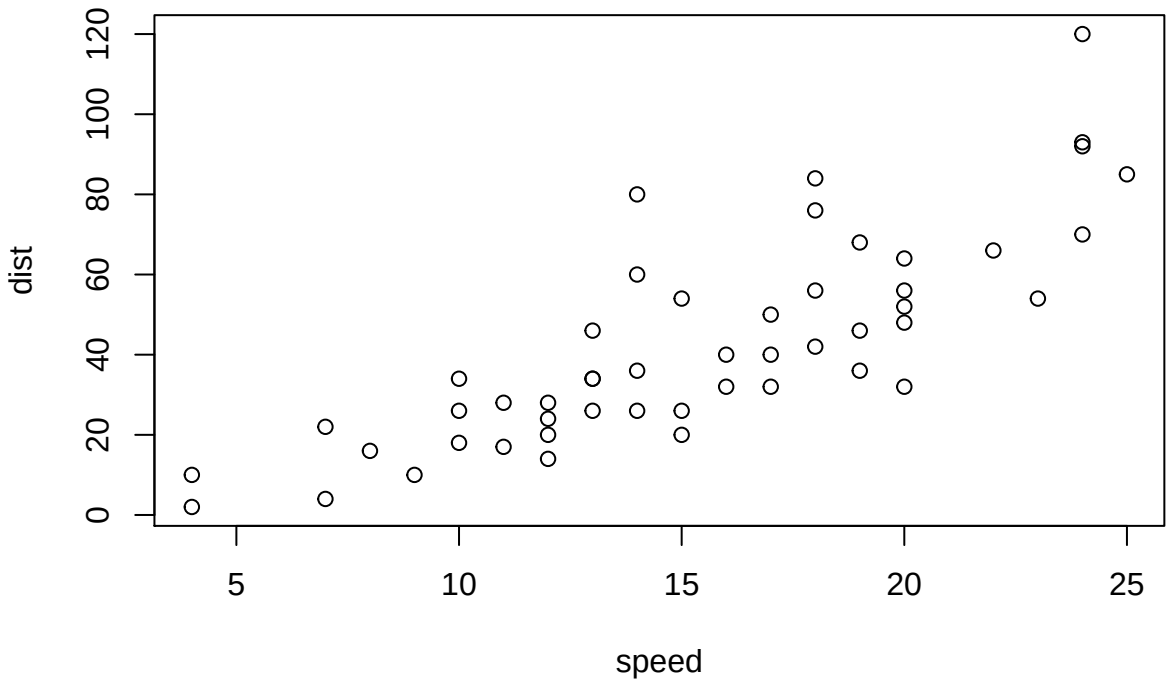
2. **Rare Events Prediction Techniques**
3. **Obvservations Sampling and Reordering**

You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

You can also embed plots, for example:



References

1. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 67(2), 301-320.