Dimensionality Reduction Based on Penalized Regression Methods

Alessandro d'Agostino
a.y. 2018/19

September 15, 2019

Abstract

This project shows a *dimensionality reduction* method for regression based on the selection of the most significant features in a dataset for the regression purpose. The selection is made through different penalized regression methods (Ridge, Lasso and ElasticNet) and looking at those features that receive the highest regression coefficient as the most significant. A detailed description of the process and the results of two different *applications* on real datasets follow below: the algorithm gave positive results in both the cases.

The Method

Given a whole dataset on which perform a regression, being able of detecting those features that contains the most of the information could be very useful. In this specific case it was avoided the use of any technique that perform linear combination in the feature space in order to maintain those features "pure" and allowing the possibility of a *a posteriori* interpretation of the results.

Prior Penalized regression The first step in this pipeline is the extraction and the *regularization* of the numerical features. Three different methods were used for the regularization purpose:

- Standard: Standardizing the features by removing the mean and scaling to unit variance.
- MinMax: Transforming the features by scaling each feature in the range $[0, 1]$.
- Robust: Standardizing the features to unite variance giving less importance to the outliers, so in a more *robust* way.

In combination with each one of these method three different penalized *regression* algorithm were used:

- **Ridge** regression, which implements a penalization on the L^2 norm of the coefficients vector.

- **Lasso** regression, which implements a penalization on the L^1 norm of the coefficients vector.

- **Elastic net** regression, which implements a penalization halfway between the former two.

After performing the standardization step and the regression step for every possible combination, 9 sets of values are obtained. Those arrays contains the regression *coefficients* of each features in every partial pipeline. Those values are useful to detect the most influential quantity for the regression purpose: those features with the highest absolute values are seen as more important for each distinguished regression algorithm. Just a single set of coefficients for every regression algorithm is shown below in Figure 1, after a sorting on the absolute values. The influence of the standardizing method has resulted negligible in most of the case, though.

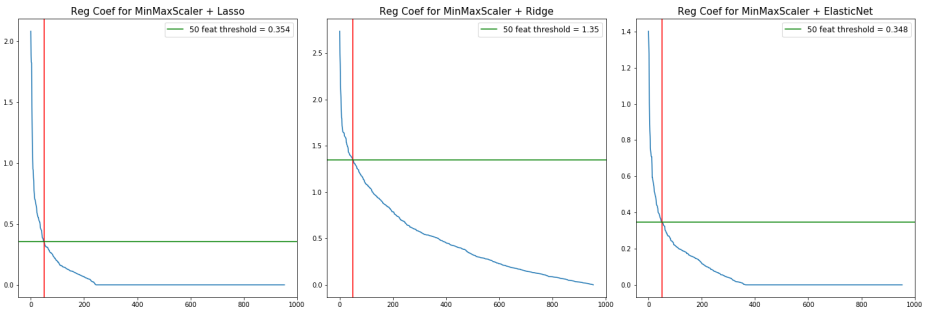

Figure 1: The three plots above show the three coefficients sets obtained through the three regression method on the whole dataset after the MinMax regularization. The values have been sorted by absolute value. The red vertical line shows the cut over the first 50 features. The green horizontal line detect the value of the 50^{th} coefficient. In all of the three curves it's evident the elbow *behave of the curve and that the* TOP 50 *features cut falls around the elbow point.*

Dimensionality Reduction From the silhouette of the curves in Figure 1 it's clear the presence of an ***elbow***: a point where the curve starts decreasing more gently. That point was taken as *reference* for the number of features to select from the whole set. The idea under this choice is that considering features beyond the elbow point pay less and less back respect to those feature before this point.

The coefficients form different regressors live in different value range, so I devised a *non parametric* comparison method to score the features. The method is not based on the value of a feature's coefficient but on its *position* in the sorted (by absolute value) array. As a score for each feature it was made a count of how

frequent every feature appears in the top ranked in the different charts given by the different regression methods.

Example - *Brain Challenge* The *Brain Challenge*'s dataset contains around 1000 features (i.e. 954 numerical values). After the prior regressions the nine sets of coefficients were procured (three of which are shown in Figure 1). The precise detection of the *elbow* point it's not a trivial task, so in a first phase of the project it was detected by eye to be around the 50^{th} sorted features for each one of the values sets. This way it was possible to give a point to each features when it fell in the *TOP 50* sorted set. The same exact process was repeated for smaller sets of features, in this case for the *TOP 25* and the *TOP 10*. A brief snapshot of the classification obtained follows below (a more detailed list in Appendix)

Feature Name	TOP50 scores	TOP25 scores	TOP10 scores
Left.Putamen	9	9	9
Left.Thalamus.Proper	9	9	9
lh_superiorfrontal_thickness	9	9	6
X3rd.Ventricle	9	9	6
Left.Cerebellum.Cortex	9	6	6
...	...	...	...

Regression on the Reduced Dataset

In order to *test* the quality of the dimensionality reduction a further regression should be performed both on the whole dataset and on the reduced ones. This time the restriction to linear models it's not necessary, indeed suport vector regressors and gaussian process were widely used too.

Those new regressors have been trained on the majority of the available dataset (around the 90%) to predict the well known target value. Once trained, every regressor processes the remaining part of the dataset and make a prediction on the target values. Those predicted values are then put in comparison with the true ones with a straightforward linear model fitting. The R^2 score of that last fitting serves as *quality score* of the regression on the reduction features space. This very quality score was used to quantify the change in performance of the same regressor on different reductions of the same feature space.

The detailed description of two different applications of this pipeline follows: both the faced dataset were made of biomedical data and allow the prediction of the age of the surveyed.

1^{st} Application - *Brain Challenge*

The Dataset The dataset available for this analysis is a collection of 954 numerical features collected from NMR images of the brain of different patients

in different sites, and 5 further categorical features that provide personal information of the surveyed like age, gender and the site of medical analysis. The aim of the analysis was to put in relation those numerical features with the age of the patients. In Figure 2 is shown the distribution among gender and age in the 16 different sites of analysis.

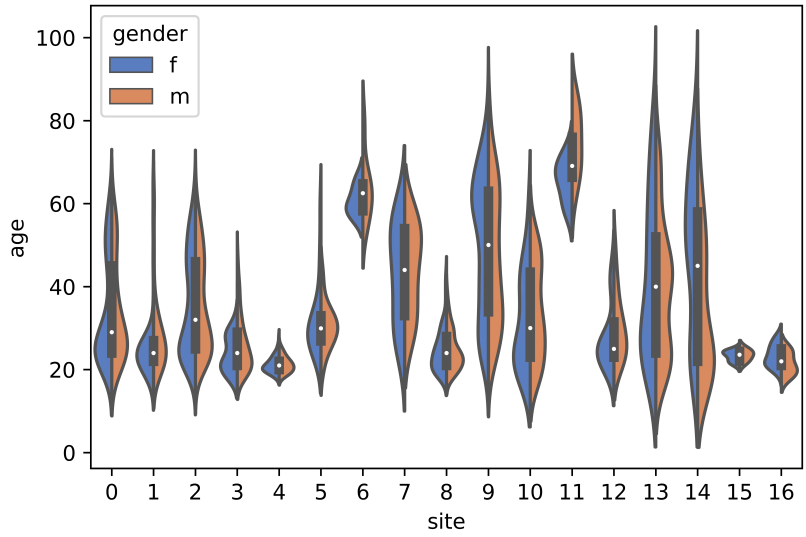

Figure 2: Series of violin graphs that show the distribution in gender and age among all the site of analysis. It turned out a clear majority of young surveyed of age around 20 years respect to the more grown, and only 3 or 4 sites cover homogeneously a wide range of age. The distribution between male and female instead appears to be quite symmetrical for each site.

Dimensionality Reduction Has previously stated, as can be seen in Figure 1, the elbow point for these 9 sets of coefficients fell around the 50^{th} features. So three different selection were made; counting those features that appeared in the *TOP 50*, in the *TOP 25* and in the *TOP 10* features of every set of regression coefficients. The complete list of features with the three kinds of scores is attached in Appendix .

Regression on Reduced Dataset In order to test the dimensionality reduction efficacy I performed as described above a further regression, using a Support Vector Regression (SVR) algorithm. The three reduction that were employed were the dataset containing all and only the 98 features that appeared at least once in a coefficient *TOP 50*, the 49 appeared in the *TOP 25* and the 18 appeared in the *TOP 10*. The SVR was trained on these dataset through a cross

validated grid agorithm, that allows to scearch directly for those parameter that better fit the samples. The three different age predictions on the three dataset were put in relation with the true test values with a simple linear fitting, and a R^2 score were computed. In Figure 3 those three fitting are shown.

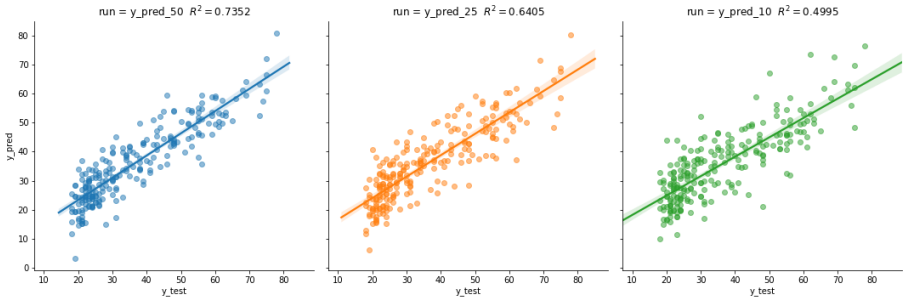

Figure 3: These are the three R^2 score obtained training the SVR on three different reductions of the feature space: selecting only those features that fell at least once in the TOP 50*,* TOP 25 *and* TOP 10*. It's easy to see that reducing the dimension of the space the R^2 value drops. It's interesting seeing how a reduction in dimension from 954 to 98 (an order of magnitude) could anyway provide such a good R^2 score, around 0.74.*

In the plots in Figure 3 those R^2 score are reported. It's easy to see that reducing the dimension of the space the R^2 value drops. The first linear fit it's the one relative to the reduction in dimension to the most important 98 features. The R^2 score of that prediction is quite good (0.74). The successive two plot instead represent two more strict reduction to 49 and 18 features, in these cases the R^2 score drops coherently with the reduction.

It's interesting seeing how a reduction in dimension from 954 to 98 (a solid order of magnitude) could anyway provide such a good R^2 score, around 0.74. These behaviors have been considered as good hints of the correct work of this pipeline for the dimensionality reduction.

2^{nd} Application - *Cardiological Data*

The Dataset The dataset available for this analysis is a collection of around 2122 samples and 83 features. Only 72 of the features are numerical, the other are made of categorical value and more complex ones. Also in this case the aim of the analysis was the age prediction starting from the data.

Dimensionality Reduction Almost all the consideration made previously for the Brain Challenge application hold here: the sorted coefficients presented all an elbow point around the 15^{th} feature. Three cuts were made here as well detecting the *TOP 20*, *TOP 10* and *TOP 5* in every set of coefficient. The list of the most important features according this method is shown in Appendix .

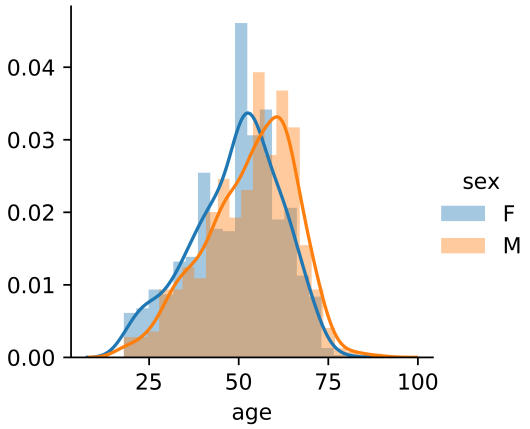

Figure 4: In this dataseta the distribution among ages between gender is quite similar. It generally seems an homogeneous survey.

Regression on Reduced Dataset The *a posteriori* regression in this case was made with a Gaussian Process Regressor (GPR). This regressor was trained four independent times with the whole dataset and with the three reduced dataset according to the *TOP 20* (23 features), *TOP 10* (12 features) and *TOP 5* (5 features). Also this time the algorithm used for the training acted in a cross validated way searching for the best parameter. The four sets of predicted ages were put in comparison with the true values and the R^2 score computed (Figure 5).

Figure 5: These are the three R^2 score obtained training the GPR on four different reductions of the feature space: selecting the whole dataset and those features that fell at least once in the TOP 20, TOP 10 *and* TOP 5. *It's easy to see that reducing the dimension of the space the R^2 value drops. It's interesting to see how the reduction in dimension from the whole dataset (72 features) to the 23 features in the* TOP 20 *slightly reduces the score from 0.35 to 0.31. A further reduction instead sharply reduces the score.*

In the plots shown in Figure 5 it's possible to see the same behaviour of the

R^2 score as in the 1^{st} application: the more the feature space is reduced the more the score drops. It's interesting to see how the reduction in dimension from the whole dataset (72 features) to the 23 features in the *TOP 20* slightly reduces the score from 0.35 to 0.31, a further reduction instead sharply reduces the score. This particular behavior strengthen the belief in the the efficacy of the algorithm for the dimensionality reduction.

Used Material

In order to read and analyze the data for this project it was used exclusively `Python` code. In particular for reading and manipulating dataframe the library `Pandas`, for visualizing the results the library `Seaborn` and for the computing parts mainly the library `SciKitLeaern`. In particular:

- for the regularization: `MinMaxScaler, StandardScaler, RobustScaler`
- for the penalized regressions: `ElasticNetCV, LassoCV, RidgeCV`
- for the further regressions: `GaussianProcessRegressor, SVR`
- for searching the best parameters: `GridSearchCV`

and a lot of other tools from that library.

Appendix 1 - Brain Challenge Features Charts

Feature Name	TOP50 scores	TOP25 scores	TOP10 scores
Left.Putamen	9	9	9
Left.Thalamus.Proper	9	9	9
lh_superiorfrontal_thickness	9	9	6
X3rd.Ventricle	9	9	6
Left.Cerebellum.Cortex	9	6	6
Right.Lateral.Ventricle	6	6	6
lh_G_front_sup_thickness	6	6	6
lh_S_circular_insula_inf_thicknessstd	6	6	6
Brain.Stem	6	6	6
Right.Amygdala	6	6	6
rh_G_front_sup_thickness	6	6	3
lh_Pole_occipital_thickness	9	6	0
lh_WhiteSurfArea_area	9	6	0
lh_S_precentral.sup.part_thickness	6	6	0
X4th.Ventricle	6	6	0
rh.fimbria	6	6	0
BrainSegVol.to.eTIV	6	6	0
Right.VentralDC	6	6	0
lh_S_circular_insula_sup_thicknessstd	6	6	0
rh_parsopercularis_thicknessstd	6	6	0
rh_S_subparietal_area	6	6	0
lh_posteriorcingulate_area	6	6	0

Appendix 2 - Cardiological Data Features Charts

Feature Name	TOP20 scores	TOP10 scores	TOP5 scores
ac_slope	9	9	9
ad_slope	9	9	9
sdsd	9	9	9
smoke	9	9	9
rmssd	9	9	9
t_ad	9	9	0
bc_slope	9	9	0
ibi	9	9	0
t_bd	9	6	0
t_ac	9	6	0
afib	9	3	0
tpr	9	3	0
dt_var	9	0	0
bd_slope	9	0	0
sdnn	9	0	0
b	9	0	0
c	9	0	0
AGI	9	0	0
ae_slope	6	0	0
pnn50	6	0	0
b - (d/a)	3	0	0
pnn20	3	0	0