Composition of concrete and its influence on compressive strength

Filipe P. de Farias
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
filipepfarias@fisica.ufc.br

Yvo J. M. Sales
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
yvo@gtel.ufc.br

***Abstract*—The compressive strength of concrete impacts directly on its application. The difference between the concrete for columns or beams and the concrete for pavements is mainly due the compressive strength it is able to resist. In this work, we perform an unconditional and a class-conditional mono-variate analysis as well as unconditional bi-variate and multi-variate analysis of the UCI concrete composition database.**

***Index Terms*—concrete, compresive, strength, machine, learning, pre-processing**

I. Introduction

A material formed by aggregates bonded together by a fluid material that hardens over time has been used by humans for construction since many years ago [3]. Nowadays this material is known as concrete and it's widely used in the construction field. The aggregates used in the mixture of the concrete affect directly its compressive strength which highly impacts its applications. For instance, in general, the concrete for columns or beams needs to have a greater compressive strength than the one for pavement. On the previous work, we made a statistical analysis on a dataset extracted from the UCI Machine Learning Repository (University of California, Irvine) [4] that collects information about the concentration of some aggregates used to form different types of concretes. Now we try to build a regression between those aggregates and the strength of the concrete, in order to find a relation that could indicate a mixture with better strength performance.

This work is divided as follows. A description of the data is given in Section II resulted from the previous work with the addition of the regressor (concrete compressive strength). The Section **??** brings an unconditional and a class-conditional mono-variate analysis as well as unconditional bi-variate and multi-variate analysis of the data. Finally, the conclusions and considerations are exposed in Section V.

II. Data description

The composition of each one of the N concrete samples is given by the concentrations (kg/m^3) of D components: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate and Fine Aggregate. The cement is what binds the elements of the concrete together. Indeed his technical name in the literature is *binder* [5]. The other components as blast furnace slag and fly ash, the outcomes of another industrial process reused in the concrete mixture, they have the role of increase the chemical hardness of the concrete, i.e. in a microscopic level. The water is responsible for react with the cement resulting in the cement stone. The superplasticizer gives fluid characteristics to the concrete aiming to better fill the mold and decrease the use of water. The coarse and fine aggregates give some macroscopical mechanical resistance to the concrete but can reduce its compressive strength if bad applied. Their major role is to occupy the spaces in the mold reducing the use of cement. The output is the concrete compressive strength which is measure in the stress test where a force its applied to a sample using a hydraulic press. When the sample reaches rupture the pressure, force per area of the sample, is observed.

TABLE I
Data description

Label	Component	Unit
D_1	Cement	kg/m^3
D_2	Blast Furnace Slag	kg/m^3
D_3	Fly Ash	kg/m^3
D_4	Water	kg/m^3
D_5	Superplasticizer	kg/m^3
D_6	Coarse Aggregate	kg/m^3
D_7	Fine Aggregate	kg/m^3
D_8	Age	days
$\mathbf{D_9}$	**Compressive strength**	**MPa**
Total	1030 samples	

The observations are the measured compressive strengths of each sample and, as the predictors $D_1 - D_7$, are continuous. The Age (D_8) of the concrete is extremely discrete.

III. Methods

Regression models try to find relations between the *independent variables* and the *dependent variables*, which are named, respectively, predictors and outcomes in this work. These relations can occur in different forms. The simplest one is the linear relationship, which is when the curve predictors *vs* outcomes, in the case that both are one-dimensional, forms a simple line and, in the general case, a hyperplane. In what follows, we formulate the *Linear regression*, which is

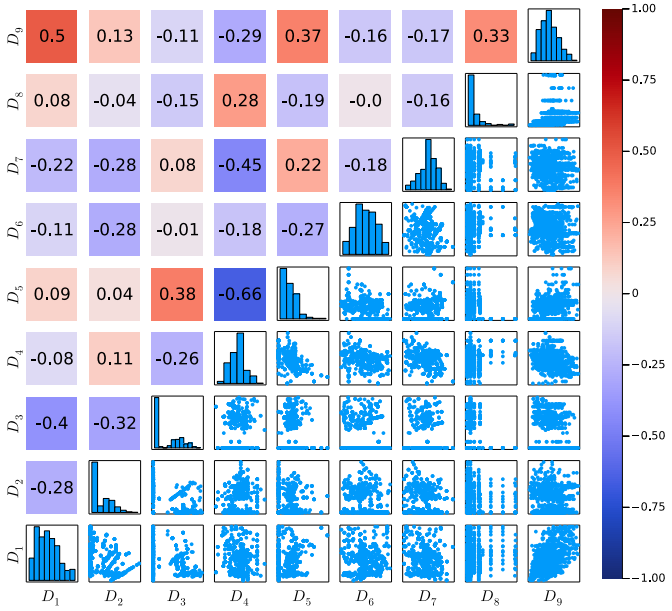

Fig. 1. Pairwise scatter, correlation and histogram plots of each concrete component.

a subclass of regressions dedicated to find a linear model to explain the relation between the predictors and the outcomes.

A. Linear regression

$$Y = \begin{bmatrix}\beta_0 & \boldsymbol{\beta}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}\\ X\end{bmatrix} + \varepsilon \tag{1}$$

This method tries to find the linear regression between the predictors and outcomes by fitting a line with linear coefficient $\boldsymbol{\beta}$ and angular coefficient $\boldsymbol{\beta}$, defined in Eq. (1), through the data. $\boldsymbol{\beta}$ is the size such as the number of predictors in X, in the way we can define the vector $\boldsymbol{\beta} = [\beta_0, ..., \beta_D]^\top$. Wrapping up $[\beta_0\ \boldsymbol{\beta}]^\top$ is a $1 \times (D+1)$, $\mathbf{x} = [\mathbf{1}\ X]^\top$ is a $(N+1) \times D$ matrix and Y is a $N \times 1$ matrix. For limitations on the implementation, it was adopted the label D_9 for the outcome, then the dimension D of the matrices must be considered without the outcome component, that is 8 components.

The fitting, i.e. finding the values for each of the linear and angular coefficients, is done by minimising a cost function that can take different forms. Each cost function yields to different optimal parameters and two of them are described in the following, the *Ordinary least squares* and *L_2-penalized least squares*. An interesting fact to observe is when the predictors correlates with the outcome, we can observe the error will be small. This is because the correlation evaluates linear variations between the variables as such as the linear regression.

The *Ordinary least squares* defines the cost function to find the optimal parameters for Eq. (1) as

$$L(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2. \tag{2}$$

This cost function represents a quadratic distance between the model, i.e. the linear combination of the coefficients and the data. We have the minimal distance, we obtain the best model. This minimisation can be done by differentiating Eq. (2) w.r.t. the β's giving the coefficients $\hat{\boldsymbol{\beta}}$ of the best model by

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top Y.$$

The *L_2-penalized least squares* modifies Eq. (2) adding a term that penalize large values of the parameters, yielding to

$$L(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{3}$$

where λ is the penalisation coefficient, a tuning parameter. Different from ordinary least squares, we find λ such the model will be the best one with *cross-validation*. The values of the coefficients for a given λ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top\mathbf{x} + \lambda)^{-1}\mathbf{x}^\top Y.$$

Partial least squares

IV. Results

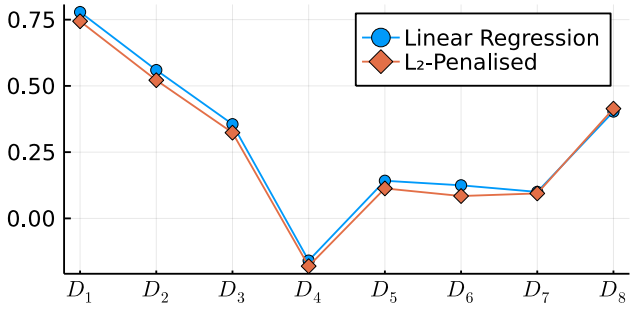

Fig. 2. Values of the weights of each predictor obtained after 30%/70% cross-validation strategy.

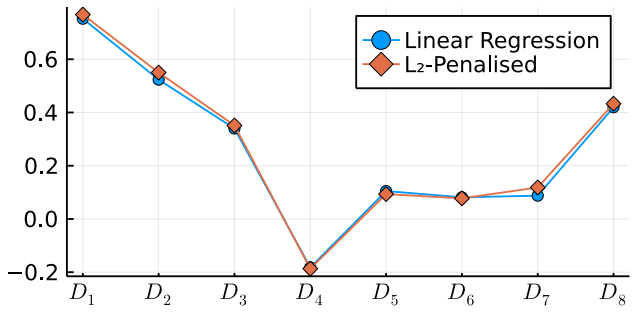

Fig. 3. Values of the weights of each predictor obtained after 5-fold cross-validation strategy.

TABLE II
ORDINARY LEAST SQUARES REGRESSION SUMMARY

CV	RMSE	R^2
70% Train / 30% Test	0.614014	0.612271
1-st fold	0.614187	0.634496
2-dn fold	0.647336	0.597003
3-rd fold	0.5952	0.599484
4-th fold	0.686377	0.530463
5-th fold	**0.582266**	**0.661407**

TABLE III
L_2-PENALIZED LEAST SQUARES REGRESSION SUMMARY

CV	RMSE	R^2
70% Train / 30% Test	0.599004	0.640309
1-st fold	0.613885	0.634857
2-dn fold	0.647344	0.596994
3-rd fold	0.594722	0.600127
4-th fold	0.68599	0.530993
5-th fold	**0.582357**	**0.661302**

V. CONCLUSION

The database of concrete is not easy to analyse if there is no previous knowledge about the problem of the components mixture. In none of the analysis the data have been shown as separable on the initially determined classes. The next step is to try to perform regression to model the compressive strength itself before trying to classify the samples.

REFERENCES

[1] ACI Manual of Concrete Practice 2000, Part 1: Materials and General Properties of Concrete. American Concrete Institute. Farmington Hills, MI.

[2] Tibshirani, Robert, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Germany, Springer New York, 2009.

[3] Mindess, S., and Young, J.F. Concrete. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.

[4] I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

[5] Khasanov, Irmuhamedova, et al, "Theoretical foundations of the structure formation of cement stone and concrete", IOP Conf. Series: Materials Science and Engineering 869 (2020)

TABLE IV
PARTIAL LEAST SQUARES REGRESSION SUMMARY

CV	RMSE	R^2
70% Train / 30% Test	0.599264	0.639996
1-st fold	0.614106	0.634593
2-dn fold	0.646593	0.597928
3-rd fold	0.594709	0.600143
4-th fold	0.6859	0.531116
5-th fold	**0.58221**	**0.661472**