Housing Prices

Anna Olena Zhab'yak, Michele Maione

**Summary** – This paper implements the ridge regression and the --- to the dataset of housing prices in order to predict the median value of the house.

# Introduction and description of the problem

Okay, so, let's talk about this problem of predicting the value of a house. So, this is a very important problem, at least in the United States. It's estimated that household wealth is nearly 50% invested in real estate. So, this is clearly important. Both to consumers, individuals, as well as policy makers.

Besides that, the constructed house price model can improve the growth of the real estate market.

Still the problem of estimating the value of a house is that the value of the neighbourhood also is influencing the va,ue of my house[[1]](#footnote-1)

The most common problem with large dataset is the multicollinearity of the features which tend to overfit when it comes to implement a certain algorithms in order to predict the value. This is one of these case where the value of a house is given by a relationship of more features that sometimes can be treated as one and not in group.

The ridge regression is a machine learning method that helps to overcome the problem of multicollinearity by introducing panalities on the …. And the …

The importance of having a good algorithm that can predict the median house value given the features.

The importance of the bias.variance trade off and the analysis of the risk minimization given different algorithms will be provided

This paper is trying to answer to the question, wthat is the best model to predict the median house price given the house characteristics aka features. For choosing the regularization parameter in practice, cross-validation (CV) is widely used.

# Most important related works

<https://openreview.net/pdf?id=HklRwaEKwB>

Our work is connected to prior works on ridge regression in high-dimensional statistics (Serdobol-skii, 2007) and wireless communications (Tulino & Verd ́u, 2004; Couillet & Debbah, 2011). Amongother related works, El Karoui & K ̈osters (2011) discuss the implications of the geometric sensitivityof random matrix theory for ridge regression, without considering our problems. El Karoui (2018)and Dicker (2016) study ridge regression estimators, but focus only on the risk for identity covari-ance. Hastie et al. (2019) study “ridgeless” regression, where the regularization parameter tends tozero.

# Notation and relevant definitions

**Loss function**

which measures how different the prediction y ^ {\displaystyle {\hat {y}}} of a hypothesis is from the true outcome y . {\displaystyle y.} The [risk](https://en.wikipedia.org/wiki/Risk_(statistics)) associated with hypothesis h ( x ) {\displaystyle h(x)} is then defined as the [expectation](https://en.wikipedia.org/wiki/Expected_value) of the loss function:

The **loss function** maps an [event](https://en.wikipedia.org/wiki/Event_(probability_theory)) or values of one or more variables onto a [real number](https://en.wikipedia.org/wiki/Real_number) intuitively representing some "cost" associated with the event.

it is used for [parameter estimation](https://en.wikipedia.org/wiki/Parameter_estimation), and the event in question is some function of the difference between estimated and true values for an instance of data.

We use a nonnegative loss function to measure the discrepancy between the predicted label and the correct label .

In the regression task we define the quadratic loss

That is the squared distance between y and y hat.

When then ; If and *c* is large, then also will be large.

The risk function , or the expected loss, becomes the mean squared error of the estimate with respect to the probability distribution

L2 Loss [Mean Square Error (MSE)](https://medium.freecodecamp.org/machine-learning-mean-squared-error-regression-line-c7dde9a26b93) is the most commonly used regression loss function.

in regression we refer to the estimation of a probability density as the loss function as the eucleadian distance in a function space , ideed:

And the risk function becomes the squared distances between our target variable and predicted values

## Empirical Risk Minimization

It’s a principle that defines a family of algorithms of learning and it’s used to give theoretical limits to their performance. The risk is how well an algorithm will work in practice, because we don’t know the distribution of the data that it will work on, but we can measure the performance on a training test, that is the empirical risk.

So it is the learning algorithm that outputs some predictor in F minimizing the training error.

where is a set of predictors and is the loss function.

*empirical risk minimizer is a function*

We also assume that we are given a non-negative real-valued [loss function](https://en.wikipedia.org/wiki/Loss_function) L ( y ^ , y ) {\displaystyle L({\hat {y}},y)}

R ( h ) = E [ L ( h ( x ) , y ) ] = ∫ L ( h ( x ) , y ) d P ( x , y ) . {\displaystyle R(h)=\mathbf {E} [L(h(x),y)]=\int L(h(x),y)\,dP(x,y).}

The empirical risk minimization principle states that the learning algorithm should choose a hypothesis h ^ {\displaystyle {\hat {h}}} which minimizes the empirical risk:[[2]](#footnote-2)

*TEST Error*

,

training error

Where high complexity models could have very low bias, but high variance.

Whereas low complexity models have high bias, but low variance.

## Ridge Regression

We start from the classic linear regression model

Let be the data domain and

The linear predictor is a linear function of the data points , with

and where

wich can be rewritten as

Regression with squareloss , we want the linear predictor to be as good as possible, for istance we can look at the risk that would be the base optimal

ERM for is

vector of predictions

vector of real labels

, so we can define

=

Let be the design matrix matrix with features and observatios , what we get from data points, that are rows

,

So now

The function is a convex function

The minimization priblem becomes a problem to

minimize the eucleadian distance. Using the gradiend descend approach

So we get the gradient of w.

And the solution is, closed form

Least squares solution, we minimize squared error

has to be non singular matrix (Usually it is invertible

) if the data points span r

And second is condition for **general position**

Closed form for erm

Hight bias if the function is non linear or can be nearly linear then my error will be small

If we change exaples, the fit can change significally, it is instable and it leads to variance error. High bias and high variance.

If the inverse in unstable so the matrix non singular

To sability and optimization

This is a linear fit, could find a good solution on the training data, but give a bad estimate on different data. The reason why it occurs it is because of multicollinearity of the prediction vectors (aka not orthogonal [[3]](#footnote-4))

and more in general

with large or small, the risk that the model can overfit[[4]](#footnote-5) the data is high.

The new estimated parameter becomes

,

α is a regularizer parameter

There could be two different situations, for so we are in the prevoious regression

for ,

If alfa goes up, if we minimize over we do not care aboute the forst term couse the regularizaton will dominate, we want w alfa be small

Shrink the linear regression solution towords to zero, rapresenting by w alfa.

W alfa hat has more bias but also less variance

wanted to trade off between bias and

variance to get to that sweet spot of having good predictive performance.

biased estimators of the regression coefficients in the regression model

regularization by introducing some bias ( bias = how well the fit correspond to the true value )

What ridge regression essentially does is tunning the bias variance tradeoff and overcome the problem of overfitting the data.

The solution, to do the fit, for a suitable value of w is given by the optimization problem where the loss function is the objective function and performing the gradient descend it is possible to tune the parmeter alfa to minimize the squared loss.

*In multiple regression it is shown that parameter estimates based on minimum residual sum of squares*

*have a high probability of being unsatisfactory, if not incorrect, if the prediction vectors are not*

*More in generale where D is a large nimber, so we have lots of features, the regression model has a lot of fleibility to explain the data and there is the risk of overfitting*

*Model can rapidly overfit the data, if n small*

*We now introduce, > 0, in the ERM functional:*

**K-fold cross-validation**

K-fold cross-validation is a non-parametric method for evaluating the accuracy of a predictive rule.[[5]](#footnote-6)

for model selection

Hyperparameters whose value must be determined before the training phase can start

The computed error on the testing part of each fold is

is the dataset divided in subsets , where is the testing part of the

We want to estimate the for a fixed hyperparameter

And the CV is the avarage of the errors

given the choice of two predictors, it repeatedly picksthe more accurate of the two. Fix a training\test set

So for tuning the hyperparameter

we look on

where

Estimate risk on all S

# Proof of a technical result

Before performing the regression pre-procession of data is necessary.

In particular our database has 10 features; one of this is a categorical value. The capegorical feature (proximity to the ocean) was turned into a dummy variable and the missing values were filled with the mean of the entire column.

Data was standardized.

# Some critical considerations

Outliers problem

Even if ridge regressioni s modelled in a way to deal with multicollinearity and outliers, it still being sensitive to outliers. In our work there were some outliers that were pushing the fit in a wrong direction.

<https://iopscience.iop.org/article/10.1088/1742-6596/890/1/012150>

<https://iopscience.iop.org/article/10.1088/1742-6596/890/1/012150/pdf>

# Copyright

We declare that this material, which We now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.

# Bibliographical references

* Ramsay J. & Huber M. (1987) - The Techniques of Modern Structural Geology. Volume 2, Folds and Fractures. Academic Press, London, 391 pp.
* Ramsay J. & Huber M. (1987) - The Techniques of Modern Structural Geology. Volume 2, Folds and Fractures. Academic Press, London, 391 pp.

1. Coursera [↑](#footnote-ref-1)
2. V. Vapnik (1992). [<http://papers.nips.cc/paper/506-principles-of-risk-minimization-for-learning-theory.pdf> *Principles of Risk Minimization* for Learning Theory.*]* [↑](#footnote-ref-2)
3. https://www.math.arizona.edu/~hzhang/math574m/Read/RidgeRegressionBiasedEstimationForNonorthogonalProblems.pdf [↑](#footnote-ref-4)
4. Overfitting : ….. [↑](#footnote-ref-5)
5. <https://arxiv.org/pdf/1909.11696.pdf> [↑](#footnote-ref-6)