NANS Projekat

Multivarijabilna regresija nad Housing Dataset-om

Student:
Dušan Svilarković RA196-2015

Okruženje i jezik koje se koristi

- Jupyter Notebook
- Python 3.4

Biblioteke koje će biti korišćene

- Pandas (dataframes, read_csv)
- Numpy (data manipulacije)
- Sklearn (metrike, raspored uzoraka na test train,cv i test dataset, elastic net, gradient boosting i SVR)
- Copy (za deepcopy korišćenje, zbog karakteristika samog python-a kao jezika)
- Matplotlib (pyplot za ispis korelacionih matrica)

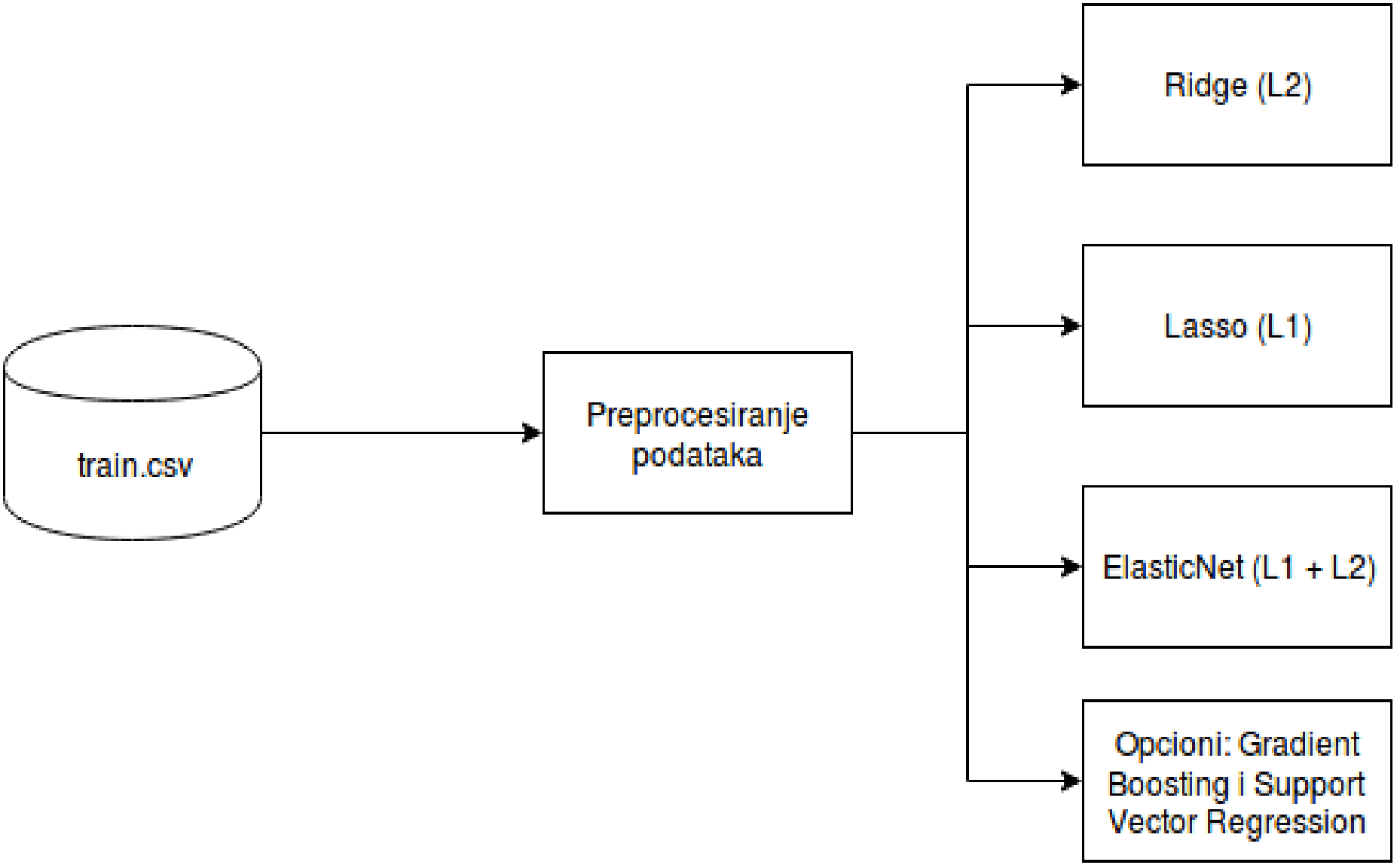
Uvod

- Skup podataka koji će se koristiti:

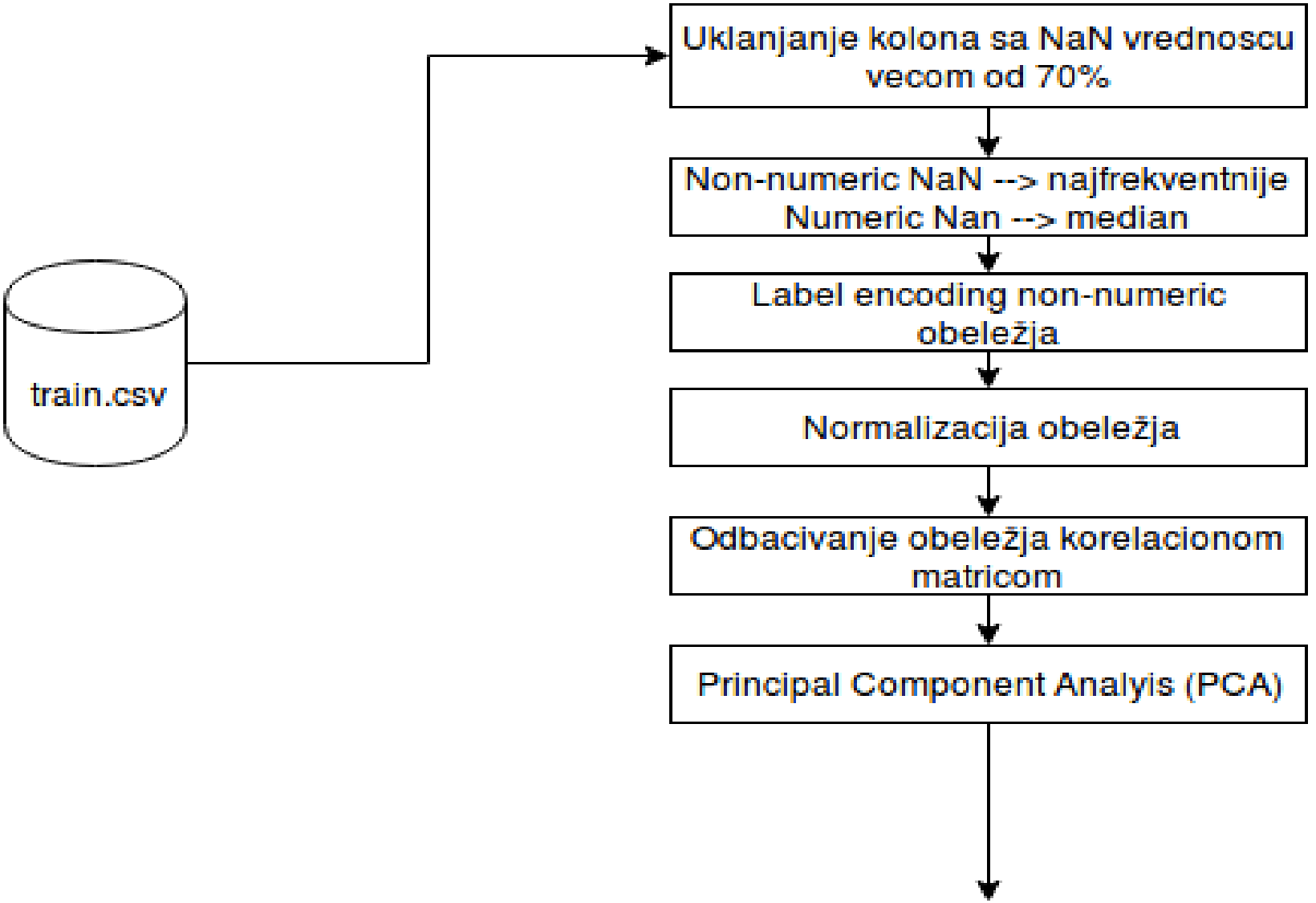
  Nekretnine u Ames, Iowa, USA
- 81 obeležje, 1460 uzoraka

```
Index([u'Id', u'MSSubClass', u'MSZoning', u'LotFrontage', u'LotArea',
       u'Street', u'Alley', u'LotShape', u'LandContour', u'Utilities',
       u'LotConfig', u'LandSlope', u'Neighborhood', u'Condition1',
       u'Condition2', u'BldgType', u'HouseStyle', u'OverallQual',
       u'OverallCond', u'YearBuilt', u'YearRemodAdd', u'RoofStyle',
       u'RoofMatl', u'Exterior1st', u'Exterior2nd', u'MasVnrType',
       u'MasVnrArea', u'ExterQual', u'ExterCond', u'Foundation', u'BsmtQual',
       u'BsmtCond', u'BsmtExposure', u'BsmtFinType1', u'BsmtFinSF1',
       u'BsmtFinType2', u'BsmtFinSF2', u'BsmtUnfSF', u'TotalBsmtSF',
       u'Heating', u'HeatingQC', u'CentralAir', u'Electrical', u'1stFlrSF',
       u'2ndFlrSF', u'LowQualFinSF', u'GrLivArea', u'BsmtFullBath',
       u'BsmtHalfBath', u'FullBath', u'HalfBath', u'BedroomAbvGr',
       u'KitchenAbvGr', u'KitchenQual', u'TotRmsAbvGrd', u'Functional',
       u'Fireplaces', u'FireplaceQu', u'GarageType', u'GarageYrBlt',
       u'GarageFinish', u'GarageCars', u'GarageArea', u'GarageQual',
       u'GarageCond', u'PavedDrive', u'WoodDeckSF', u'OpenPorchSF',
       u'EnclosedPorch', u'3SsnPorch', u'ScreenPorch', u'PoolArea', u'PoolQC',
       u'Fence', u'MiscFeature', u'MiscVal', u'MoSold', u'YrSold', u'SaleType',
       u'SaleCondition', u'SalePrice'],
      dtype='object')
```

,,Pipeline” za korišćenje



Preprocesiranje podataka



Izbačene zbog viška NaN vrednosti

Izbačene zbog korelacije

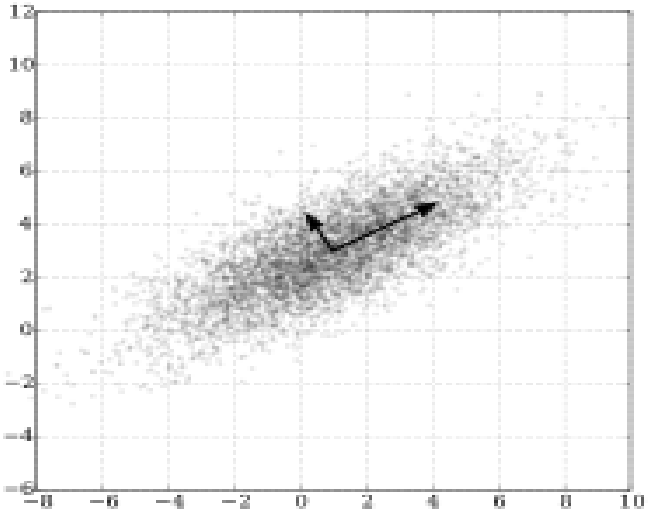
PCA-postupak

1.Izračunati matricu korelacije

2.Preko Singular Value Decomposition dobijamo sve karakteristične vektore

3.Uzimamo prvih k najboljih karakterističnih vektora

4.Transformišemo n-dimenzionalne podatke u k-dimenzionalne pomoću karakterističnih vektora



- Problem: over-fitting,preterano poklapanje podataka
- Rešenje: regularizacija, namerno dodavanje penala na kriterijum optimalnosti radi dobijanja rešenja koje bolje generalizuje

Regularizacija-postupak

- Podela uzorka na **train**(60%), **cross-validation**(20%) i **test**(20%) set.
- Uzimanje penalizacionog parametra iz skupa [0,1] sa korakom od 0.5 i treniranje modela uz pomoć **train** seta.
- Izbor najoptimalnijeg penalizacionog parametra na osnovu kriterijuma optimalnosti iz **train** i **cross-validation** seta.
- Konačno testiranje r2 skora nad **test** setom za najoptimalniji penalizacioni parametar.

Postupci za treniranje parametara

- Lasso - cirkularno koordinatno spuštanje
- Ridge – gradijentno spuštanje
- ElasticNet – bibliotečka implementacija , iterative soft thresholding

Kriterijumi optimalnosti za prva tri algoritma

- Ridge $\min_{w \in \mathbb{R}^n} \left\{ \frac{1}{N} ||y - Xw||_2^2 + \lambda ||w||_2^2 \right\}$
- Lasso $\min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2N} ||y - Xw||_2^2 + \lambda ||w||_1 \right\}$
- Elastic Net $\min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2N} ||y - Xw||_2^2 + \lambda_1 ||w||_1 + \lambda_2 ||w||_2^2 \right\}$

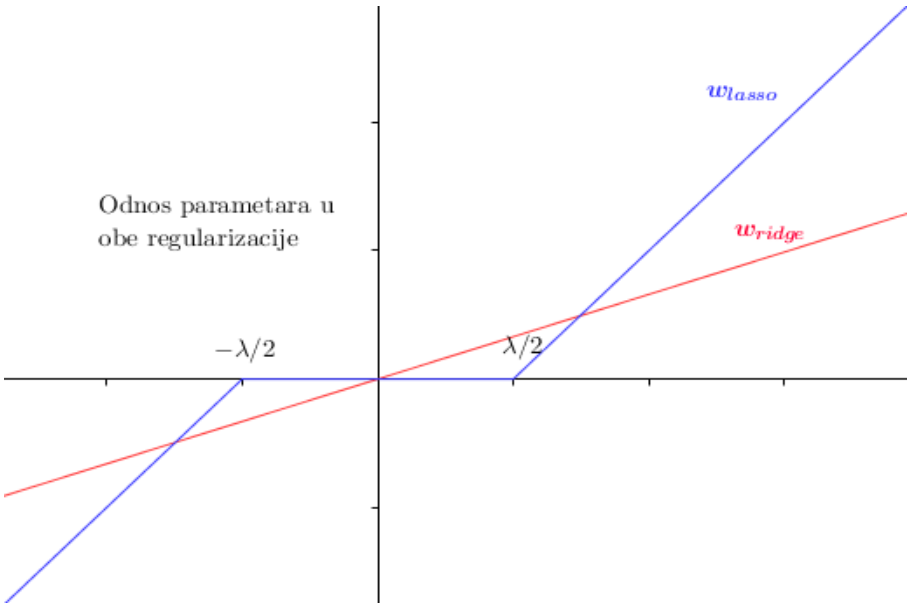
Razlike u odabiru parametara?

Razlike između ridge i lasso regresije

- Ridge će imati glatku krivu gradijenta, dok se kod lasso mora koristiti aproksimacija kad se gradijent približava nuli, što je upravo korišćeni thresholding u algoritmu.

  soft_thresholding(ro, lam)

- Dok ridge čuva sva obeležja ali bolje ,,pazi" pri spuštanju do optimalnih parametar, lasso čuva samo ona najznačajnija, a ostala anulira



Metrika

- R2-score

  r2_score_dusan(y_pred,y_orig)

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} = 1 - \frac{SSE}{SST}$$

Algoritam za ridge

ridge_regresija_gradijenta(X,y,lam,w,alpha,num_iter)

Za svako w_i iz vektora parametara w

$$w_0 = w_0 - \left(\frac{1}{N} \alpha \sum_{j=1}^{N} x_0^{(j)} (\sum_{k=1}^{m} x_k^{(j)} w_k - y_j) \right)$$

$$w_i = w_i - \left(\frac{1}{N} \alpha \sum_{j=1}^{N} x_i^{(j)} (\sum_{k=1}^{m} x_k^{(j)} w_k - y_j) + \frac{\lambda}{N} w_i \right), i \geq 1$$

Ponavljati do konvergencije ili unapred odredjenog broja koraka

Algoritam za lasso

lasso_regresija(X,y,lam,w,num_iter)

Za svako w_j iz skupa promenljivih w

Izraunati meru za normalizaciju parametara

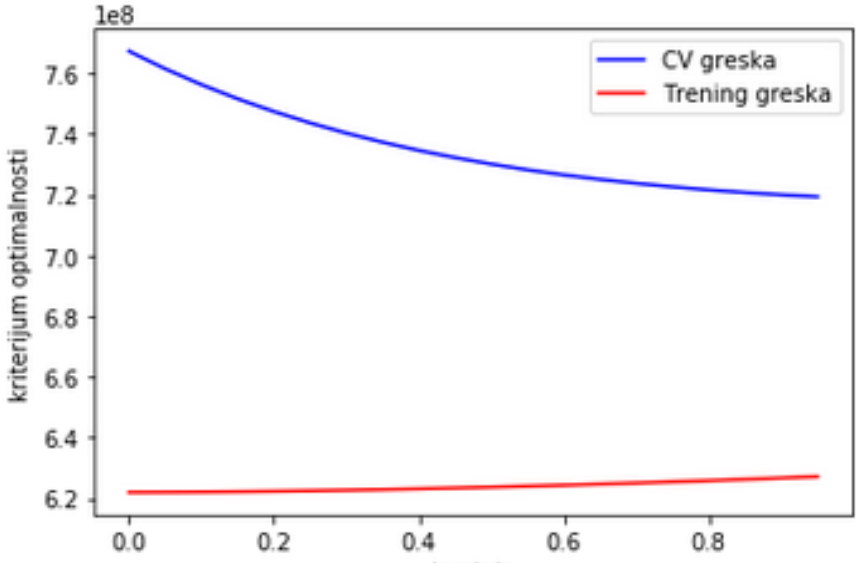
$$z_j = \sum_{i=1}^{N} (x_j^{(i)})^2$$

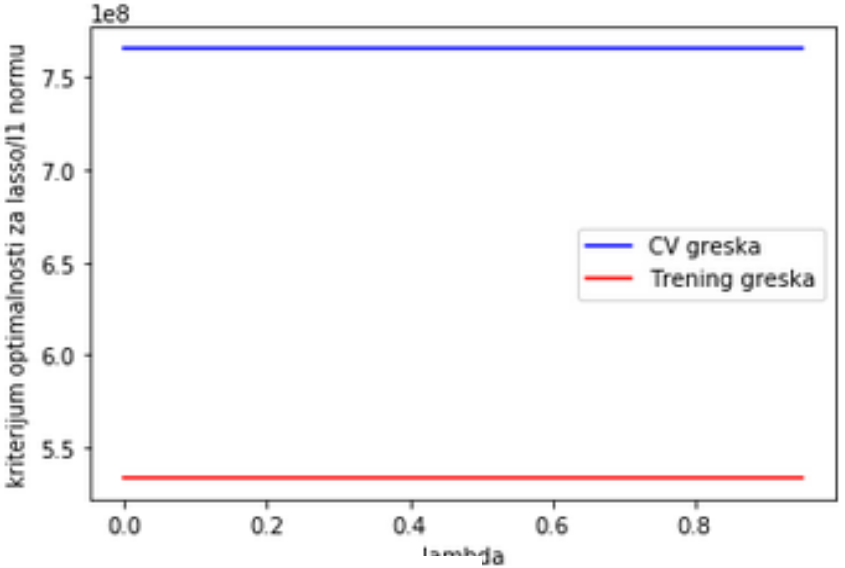
Izracunati za podgradijent $\rho_j = \sum_{i=1}^{N} x_j \left(y_i - \sum_{k=1 \wedge k \neq j}^{m} x_k^{(i)} w_k \right)$

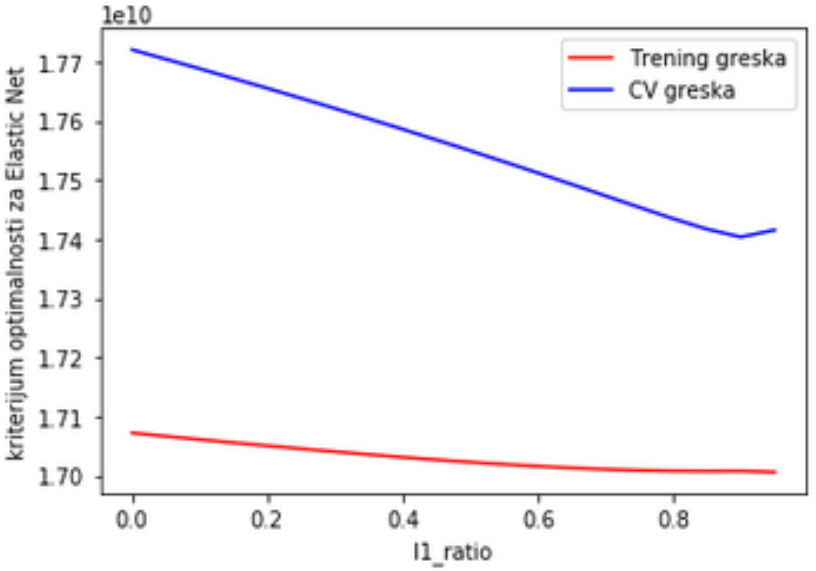
Izracunati parametar $w_j = \begin{cases} (\rho_j + \lambda/2)/z_j, & \rho_j < -\lambda/2 \\ 0, & \rho_j \in [-\lambda/2, \lambda/2] \\ (\rho_j - \lambda/2)/z_j, & \rho_j > \lambda/2 \end{cases}$

Ponavljati do kriterijuma zaustavljanja ili broja iteracija

Dobijeni rezultati regularizacije za lasso,ridge i elastic net







Reference

- Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project - D De Cock
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/kernels
  06.02.2018
- Statistical Learning with Sparsity: the Lasso and Generalizations -Hastie, Tibshirani, Wainwright
- Machine Learning by Andrew Ng, Coursera
- Machine Learning: Regression, Univerisity of Washington, Coursera
- https://math.meta.stackexchange.com/questions/5020/mathjax-basic-tutorial-and-quick-reference

Dodatno

- Support Vector Regression (neuspešno)
- Gradient Boosting