Лабораториска Вежба 2

Линеарна Регресија, Lasso Регресија и Ridge Регресија

- Целта на оваа лабораториска вежба е да ги тестирате регресионите модели кои ги имаме учено на California Housing податочното множество. За решавање на истата би требало да ви помогнат примерите кои ги поминавме на часот на аудиториски кои ви препорачувам повторно да ги разгледате.

Рокот на оваа лабораториска е 2 недели после поставување на истата (**15.12.2019**). Доколку го пропуштите овој рок, повторно имате шанса да ги добиете сите поени со дополнителното барање на крај.

Што точно треба да направите?

1. **Избирање на податочно множество**.
   Ви препорачувам да работите со California Housing податочното множество кое може да го најдете на:
   https://download.mlcc.google.com/mledu-datasets/california_housing_train.csv
   a. Број на податоци (N) = 17000, број на атрибути (D) = 8. Целта е да се предвиди вредноста на median_house_value.

   Слободно може да користите и друго регресионо податочно множество доколку имате или доколку сакате да најдете на UC Irvine archive или на Kaggle.

2. **Запознавање со податочното множество и претпроцесирање**. Во овој дел повторно би требало да направите дел од визуелизациите на вашето податочно множество за да имате некоја претстава за атрибутите во однос на таргет променливата. За таа цел препорачувам да тестирате различни верзии на jointplot функцијата (https://seaborn.pydata.org/generated/seaborn.jointplot.html) од seaborn библиотеката како во O-ring Dataset примерот. Можете слободно да направите и други визуелизации кои би ги нашле на страниците на seaborn и matplotlib. Дополнително направете min-max нормализација на податоците во ранг [0,1].

3. **Регресија**. Истренирајте три регресиони модели користејќи ја библиотеката Scikit-Learn на податочното множество што сте го избрале.

    a. Линеарна регресија
    https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression
    b. Ridge регресија
    https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html#sklearn.linear_model.RidgeCV
    c. Lasso регресија
    https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html#sklearn.linear_model.LassoCV

    Како што е направено во примерот, прво поделете го оригиналното податочно множество на тренинг и тест множество и истренирајте го регресиониот модел на тренинг множеството. Притоа може да видите дека за разлика од примерите од аудиториски каде што ги користевме стандардните Lasso и Ridge класификатори од Scikit-Learn, во оваа лабораториска треба да се искористат LassoCV и RidgeCV класификаторите. Целта на овие класи (и разликата помеѓу стандардните верзии) е тоа што во нив е веќе изграден метод на cross-validation кој на самото тренинг множество го оптимизира хипер-параметарот на овие модели (alpha во scikit-learn, или lambda во книгата). Овој хипер-параметар служи за ниво на регуларизација во двата модели и поголеми вредности соодветствуваат на поголема регуларизација. Со цел сами да не го избираме овој хипер-параметар, со овие класи можеме да го најдеме тој што ќе има најсоодветна вредност во однос на податоците кои ги трениираме. Откако ќе ги истренирате регресорите, пресметајте и споредете го MSE (Mean Square Error) на тест множеството за секој од нив, приметете дали има некоја разлика во коефициентите на моделите, како и која е соодветната вредност за хипер-параметарот на регуларизација во Lasso и Ridge.

4. **Документација:**

    a. Доколку користевте jupyter, на курсот поставете .ipynb или .html фајл од истиот.
    b. Доколку користевте стандардна работна околина, на курсот поставете .zip кој го содржи .py кодот и screenshots од резултатите и коментарите.

5. ***Дополнителна работа за поминат рок од 2 недели.*** Направете Basis Function Expansion од 2 степен на вашето податочно множество како што е направено во Housing Dataset примерот и одново истренирајте ги моделите. Која е разликата во грешката со новоизградените модели?