Assignment 4

Machine Learning: Algorithms and Theory

Prof. Ulrike von Luxburg / Diego Fioravanti / Moritz Haas / Tobias Frangen

Summer term 2018 — due to **May 15th**

Exercise 1 (K-fold cross validation, 1+3+2+2+1 points)

In this exercise we choose parameters for ridge regression via cross validation. We will implement the cross validation algorithm given in the lecture in the chapter starting at slide number 258. We will use the implementation of ridge regression provided by scikit-learn, which we import with `from sklearn.linear_model import Ridge`.

All scikit-learn algorithms have a standard API. Using Ridge as an example:

`c = Ridge(`λ`)`

creates a Ridge regression with parameter λ,

`c.fit(X, Y)`

trains the classifier, and

`Y_pred = c.predict(X_test)`

applies the classifier. We recommend the use of this API in this exercise.

(a) Implement a function

`def MSE(Y_pred , Y)`

that computes the MEAN SQUARE ERROR between two vectors Y and Y' of length n. The mean square error is defined as

$$\mathrm{MSE}(Y, Y') = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2$$

(b) Here we implement the inner loop of the k-fold. Implement a function called

`def k_fold_evaluation_MSE(classifier, X, Y, K=10):`

which, given a classifier with an standard scikit-learn API, executes K folds on X and Y and as error between Y and the Y_{pred} uses the mean square error. Here X is an $n \times d$ vector and Y is an $n \times 1$ vector.

(c) Here we implement the outer loop of the k-fold. Implement a function called

`k_fold_cv_ridge(X, Y, lambdas, K=10):`

Where X is a $n \times d$ vector, Y is a $n \times 1$ vector and `lambdas` is the list of regularization parameters that we want to use. The function executes a Ridge regression of part a) and it returns the list of average MSE for the λ in `lambdas` (where the average is taken over the different folds). Plot the MSE on a log scale.

(d) For each $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ plot on the same plot both the cross validation error returned by b) and the MSE computed on the test set. Do you notice a pattern?

(e) We will now try to see the limitations of cross validation. In principle we can apply cross validation to the k-NN algorithm in order to choose k. What is the problem with this approach?

In the notebook you will find the code that will print the results of running cross validation on knn. Does it align with your previous explanation? Justify your answer.

Exercise 2 (Comparing different types of regressions, 1+2+2+1 points)
In this exercise we will compare different types of regressions: linear, ridge and lasso. In the variables `X_train`, `Y_train`, `X_test`, `Y_test` we load and split the data that we will use.

(a) Apply linear regression and compute the MSE between Y_{test} and the predictions.

(b) For $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$, using 10 folds compute the cross validation errors and test errors for ridge regression. Plot them on the same plot. Does the cross validation select the best possible λ?

(c) For $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$, using 10 folds compute the cross validation errors and test errors for Lasso regression. Plot them on the same plot. Does the cross validation select the best possible λ?

(d) Discuss what you see in the graphs, which regressions works best? Does changing λ do what you expect?

Exercise 3 (Design your own exam questions, 3 points (1 point per question)) In this exercise, everybody is supposed to come up with suggestions for three exam questions. This is a good way to recap/understand the concepts discussed so far.

Put yourself in our place! We do not want to ask stupid questions. We would like to ask "nice questions". In general, written exams contain three types of questions:

- Questions that are just about **reproducing** knowledge. We won't ask these type of questions because you will be allowed to bring notes to the exam, in this context they are pointless.
- Questions for testing whether the person **understands** the concepts and can apply them to simple situations.
- Questions that require to **transfer** knowledge to new situations.

Your task is now to design exam questions along with their solutions of the two last-mentioned types, one of the type understanding and one of the type transferring, for each of the following topics of the lecture:

Linear Regression, Lasso, Ridge Regression, Empirical risk minimization and Regularized risk minimization framework.

Enter your questions in the LaTex file `my_exam_questions.tex` that we provided and send it to your tutors.

After the class we will put all your questions online. At the end of the course, these questions can help everybody to prepare for the exam! So take your time to invent good questions! You are also welcome to suggest more questions than required. The more questions you come up with, the better!