Lab assignment evaluation: Regression

Bachelor in Mobile and Space Communications Engineering
Bachelor in Telematics Engineering
Bachelor in Sound and Image Engineering
Bachelor in Telecommunication Technologies Engineering

Name and surname: _____

1. Let `model_trained` be a regression model which has been previously fitted. What is the appropiate way to compute the value of the coefficient R^2 over the test set?
   Please note that `X_test_s` represents the (standardized) set of test samples and `y_test` denotes the correct (or desired) values?

   √ `r2_score=model_trained.score(X_test_s,y_test)`

   ◯ `r2_score=model_trained.fit(X_test_s,y_test)`

   ◯ `r2_score=model_trained.predict(X_test_s,y_test)`

   ◯ `r2_score=model_trained.score(X_test_s)`

2. Indicate the appropiate way to divide a dataset (`X`, `y`) into 3 subsets: a training subset (`X_train`, `y_train`) that contains 50 % of the total data, a validation subset (`X_val`, `y_val`) that contains 25 % of the total data, and finally, a test subset (`X_test`, `y_test`) including the remaining 25 % of the total data.

   ◯
   ```
   X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.25)
   X_val, X_test, y_val, y_test = train_test_split(X,y, test_size = 0.25)
   ```

   ◯
   ```
   X_train X_aux, y_train, y_aux = train_test_split(X,y, test_size = 0.5)
   X_val, X_test, y_val, y_test = train_test_split(X_train,y_train, test_size = 0.5)
   ```

   ◯
   ```
   X_train, X_aux, y_train, y_aux = train_test_split(X,y, train_size = 0.5)
   X_val, X_test, y_val, y_test = train_test_split(X_aux,y_aux, test_size = 0.25)
   ```

   √
   ```
   X_train, X_aux, y_train, y_aux = train_test_split(X,y, train_size = 0.5)
   X_val, X_test, y_val, y_test = train_test_split(X_aux,y_aux, test_size = 0.5)
   ```

3. Which of the following statements is incorrect?

   ◯ In a linear regression model, it is not necessary to carry out hyperparameter optimization.

◯ In a Ridge regression model, it is possible to optimize the value of the hyperparameter `alpha`.

◯ In a Lasso regression model, it is possible to optimize the value of the hyperparameter `alpha`.

√ **In a KernelRidge regression model, it is possible to carry out the optimization of both `alpha` and `beta` hyperparameters.**

4. When designing a KernelRidge regression model with a polinomical kernel, Student A decides to optimize the values of the hyperparameters `alpha` and `degree` according to the following ranges of values:

```
alpha = [0.001, 0.01, 0.1,  1, 10, 100]
degree = [2, 4, 6, 8, 10, 12]
```

However, Student B decides to explore the following values:

```
alpha = [0.001, 0.01, 0.1,  1, 10, 100, 200]
degree = [2, 4, 6, 8, 10]
```

Which student will fit more regression models in the validation stage?

◯ Both of them will fit the same number of models

√ **Student A**

◯ Student B

◯ None of the above answers is correct

5. What indicates that there is a perfect fit in a regression model?

◯ The value $R^2 > 0$, which corresponds to RMSE $= 1$.

◯ The value $R^2 = 0$, which corresponds to RMSE $= 1$.

◯ The value $R^2 < 1$, which corresponds to RMSE $= 0$.

√ **The value $R^2 = 1$, which corresponds to RMSE $= 0$.**

6. There are five basic steps when implementing a regression model:

(a) Check the results of the fitted model to know whether the model is satisfactory.

(b) Eventually do appropriate transformations to data to work with.

(c) Apply the model for predictions (over the test set) and calculate the R^2 coefficient.

(d) Import the packages and classes that you need.

(e) Create the regression model and fit it with existing data.

However, those steps are currently listed in the wrong order. What's the correct order?

◯ e, d, b, a, and c.

√ **d, b, e, a and c.**

◯ e, c, a, b and d.

◯ d, e, c, b and a.

7. To test linear relationship of the true (or correct) values and estimated values, which of the following plots has been used in the assignment?

   ◯ Barchart

   √ **Scatter plot**

   ◯ Histogram

   ◯ None of the above answers is correct

8. It is created a linear regression model as follows:

   ```
   from sklearn.linear_model import LinearRegression
   lr = LinearRegression()
   ```

   This model is fitted by using the (standardized) training subset:

   ```
   lr.fit(X_train_s, y_train_val)
   ```

   If it is printed the value of the attribute `.coef_` of the model (`lr.coef_`), what numeric valued would be printed out?

   ◯ 10

   ◯ 1

   √ **9**

   ◯ None of the above answers is correct

9. A linear regression model has been fitted and the values of its attribute `.coef_` have been shown running the following lines of code:

   ```
   from sklearn.linear_model import LinearRegression
   lr = LinearRegression().fit(X_train_s, y_train_val)
   print(lr.coef_)
   ```

   The result (displayed on the screen) is:

   ```
   [3.65 -11 56.9 -40 -88 20]
   ```

   How would you select the 2 most relevant features aiming at obtaining a reduced dataset?

   ◯
   ```
   columns_selected = np.array([2, 5])
   X_train_reduced = X_train_s [:,columns_selected]
   ```

   ◯
   ```
   columns_selected = np.array([3, 4])
   X_train_reduced = X_train_s [:,columns_selected]
   ```

√
```
columns_selected = np.array([2, 4])
X_train_reduced = X_train_s [:,columns_selected]
```

◯
```
columns_selected = np.array([3, 5])
X_train_reduced = X_train_s [:,columns_selected]
```

10. How would you obtain the number of samples of the training subset (`[X_train_s, y_train_val]`) in the assignment?

√ `n_samples = X_train_s.shape[0]`

◯ `n_samples = X_train_s.shape[1]`

◯ `n_samples = y_train_val.shape[1]`

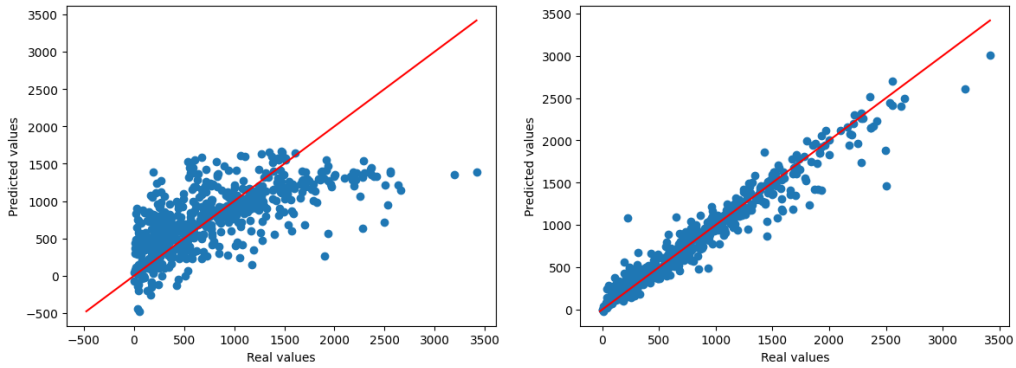
◯ `n_samples = len(X_train_s.shape[0])`

11. What have you used the object `StandardScaler` in the assignment for? When applying the `.fit_transform()` method in the dataset `X`, what is the result?

12. When fitting a linear regression model, how could you obtain the most relevant features in the dataset?

13. Which regression model has shown the best performance in the assignment? Why do you think that this model works better than the others?

14. It is shown below two scatter plots, which relate the actual (or correct) values to be estimated and the predicted values obtained when using two different regression models. On the left side, it is displayed the results obtained when using a linear regression model, while on the right side, it is depicted the results obtained when using a KernelRidge regressor.



Having a look at the two plots, which regression model would achieve a better R^2 coefficient and why? Justify your answer.

15. As it is known, any regression model in `sklearn` includes, among others, the `.fit()`, `.predict()` and `.score()` methods.

    For the linear regression model, indicate which subset has been used for both `.fit()` and `.predict()` methods, and what each of these methods performes or calculates with that subset.