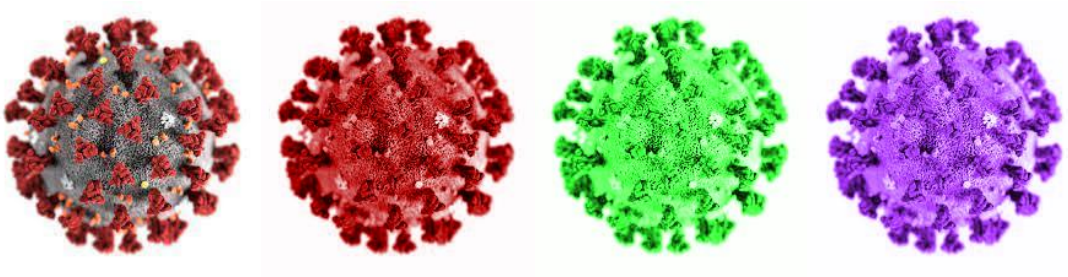
HW3: Regression

In this final assignment, we will try to predict a continuous label using our dataset.

Scientists have developed an advanced test that computes a probability that a patient is infected with the COVID virus. Unfortunately, this test is costly and cannot be performed at large scale.

Our goal is to regress the outcome of this test, i.e., the “virus score”, with the cheap features we used in the previous assignments.

Good Luck!



Instructions

- **Submission**
  - Submit by Tuesday, **28.06.2022**, 23:59. No late submissions will be accepted.
  - Submissions in pairs only in the webcourse.

- **Your code**
  - Should be clearly and briefly documented.
  - Variables/classes/functions should have meaningful names.
  - May be partially reviewed and graded.

- **Final report**
  - Should be written in a word processor (Office Word, Google docs, etc.).
    - Should not contain the code itself.
    - Do not submit jupyter notebooks as PDFs.
  - Can be in Hebrew, English, or both.
  - **You are primarily assessed based on your written report.**
  - Answer the questions in this instruction file according to their numbering.
  - Add concise explanations, figures (outputs of your code), tables, etc.
  - You are evaluated for your answers but also for clarity, readability, and aesthetics.
  - **Tables** should include feature names and suitable titles.
  - **Plots** should have suitable titles, axis labels, legends, and grid lines (when applicable).

- **Submit a zip file containing** (please use hyphens, not underscores):
  - The report PDF file with all your answers, named *id1-id2.pdf*.
  - Your code (choose the relevant options for you):
    - Working with jupyter: a notebook with your code, named *id1-id2.ipynb*.
    - Working with a "traditional" IDE: one clear main script, named *id1-id2.py*, and any additional files required for running the main script.
  - A completed *LinearRegressor.py* file with your implementation.

Preliminary: Updated Data Loading

Task: Follow the procedure below.

a. Start by **loading** the **new** raw data in *HW3_data.csv*.

   The dataset is identical to the one from the previous assignments, but we deleted the binary targets and revealed the continuous `CovidScore` and `SpreadScore` target variables which previously created the binary "`covid`" and "`spread`" target variables.

b. Make sure you did not delete the following features: `PCR_01`, `PCR_02`, `PCR_04`, `PCR_05`, `PCR_06`, `PCR_09`, `sport_activity`, `sugar_levels`, and the symptoms: `breath`, `fever`, and `sore_throat`. Note that there might be other important features for this assignment. Here we chose to specify only some of them (and some which are not). If you deleted some of those features, edit your preprocessing pipeline accordingly.

c. Make sure the data is **partitioned** correctly to train and test, according to the instructions in the previous assignments.

   The train-test partitions **must** be **identical** to the ones you used in HW1 and HW2.

d. Apply the **preprocessing** procedure from the previous assignments (including the normalization steps; you might need to normalize features that you previously deleted) on both the training and test sets (but be careful to only use the training set when computing statistics for normalization etc.).

   Note: in Lecture 08 we explain why some preprocessing steps (e.g., normalization), should be applied to the validation folds according to statistics computed on the train fold. Here, for simplicity only, you compute these statistics according to all the training samples (before splitting it to train and validation folds).

e. **Important:** do not use the test set before Section 7.

Section 1: Quick data exploration and preparation

Throughout this assignment, we mainly focus on regressing the continuous `CovidScore` variable, which was used previously to create the binary "`covid`" target variable.

(Q1) Run the following steps on the training set only.

a. Use `sns.kdeplot` to plot the distributions of the `CovidScore` conditioned on the raw `BloodType` of the patients. Since you previously removed this raw feature, temporarily return it (and remove afterwards).
   For the plots in this question, missing data (before imputation) can be ignored.

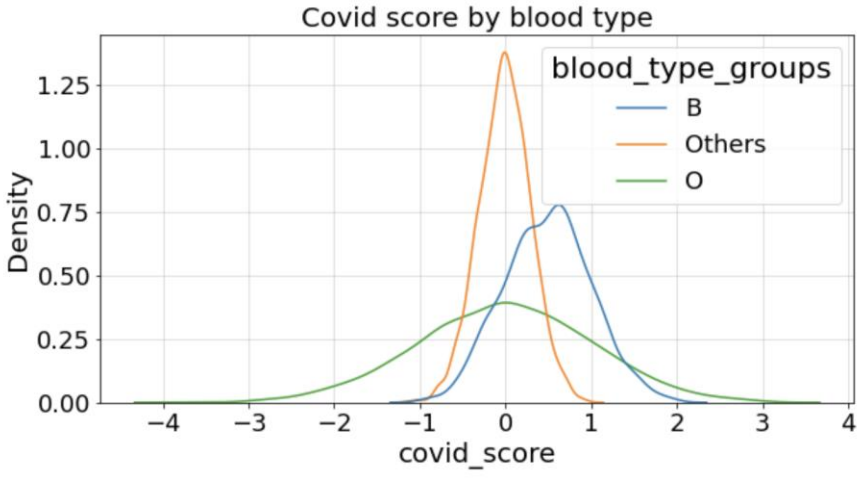
b. Plot a kde plot of the same target variable conditioned on the following groups:
   i. "O" blood types ("O+", "O-")
   ii. "B" blood types ("B+", "B-", "AB+", "AB-")
   iii. Other blood types ("A+", "A-")
   (You can do so by creating a temporary feature `Blood_Type_Groups`, having only 3 possible values.)

Important: Remember to normalize both graphs per group (check the `common_norm` keyword in the API)!

You can use the following code snippet as a starting point:

```
sns.kdeplot(data=df_train, x='covid_score',
            hue='blood_type', common_norm=False)
```

Results should roughly look like the following (the distributions here are made up).



Attach the resulting plots to your report.

Remember: include appropriate titles, axis labels, etc.

(Q2) What can we learn from the histograms above? Why bother grouping the blood types? What can we gain from such a transformation? Answer briefly (2-3 sentences).

Task: On both the train and test sets, utilize the insights we've gained on the blood type groups and create 3 suitable OHE features from these groups. Drop the `BloodType` feature and its previous transformations from HW1 (i.e., previous OHE features created from this feature). Missing `BloodType` values should be imputed with similar techniques to those you used in HW1. No need to normalize these OHE features.

Section 2: Linear regression implementation

Before we start using sklearn's modules, we want to thoroughly understand the optimization problem we use to solve linear regression – the least squares problem.

We will now implement a non-homogeneous linear regressor with the stochastic gradient descent (SGD) algorithm. The MSE loss definition for such a regressor is:

$$L(\underline{\boldsymbol{w}}, b) = \frac{1}{m}\sum_{i=1}^{m}\left(\underline{\boldsymbol{w}}^{\top}\underline{\boldsymbol{x}}_i + b - y_i\right)^2 = \frac{1}{m}\left\|\boldsymbol{X}\underline{\boldsymbol{w}} + \underline{\mathbf{1}}_m \cdot b - \underline{\boldsymbol{y}}\right\|_2^2,$$

where $\underline{\mathbf{1}}_m \in \mathbb{R}^m$ is a vector of ones and $b \in \mathbb{R}$ is a scalar. Using $\underline{\mathbf{1}}_m \cdot b$ allows the same b to be used for all rows (corresponding to all samples).

The gradient with respect to $\underline{\boldsymbol{w}} \in \mathbb{R}^d$ was presented in Tutorial 09:

$$\mathbb{R}^d \ni \nabla_{\underline{\boldsymbol{w}}} L(\underline{\boldsymbol{w}}, b) = \frac{1}{m} 2\boldsymbol{X}^{\top}(\boldsymbol{X}\underline{\boldsymbol{w}} + \underline{\mathbf{1}}_m \cdot b - \underline{\boldsymbol{y}})$$

(Q3) In your report, derive (חשבו) the analytical partial derivative $\frac{\partial}{\partial b} L(\underline{\boldsymbol{w}}, b) \in \mathbb{R}$.

To implement our regressor we will inherit from sklearn's `BaseEstimator` class (like in HW2) for compatibility with scikit-learn API. We will also inherit from `RegressorMixin`.

This is the only section where we will perform validation without using cross validation.

- For **this section only**, split your training set into a (new) training subset (80%) and a validation subset (20%).
- Copy the given `LinearRegressor` module into your notebook / project.
- Complete the following methods
  - `LinearRegressor.loss` method so that it computes the objective MSE loss $L(\underline{\boldsymbol{w}}, b)$ on a given dataset. Avoid using for loops.
  - `LinearRegressor.gradient` method so as to compute the analytic gradients $\nabla_{\underline{\boldsymbol{w}}} L(\underline{\boldsymbol{w}}, b)$ and $\frac{\partial}{\partial b} L$. Avoid using for loops.

    Tip: When possible, prefer vector operations (e.g., `np.sum`, `np.linalg.norm`).
  - `LinearRegressor.fit_with_logs` method in the module so as to compute the gradients of the current batch and update the parameters accordingly.
  - `LinearRegressor.predict` method so as to compute the model prediction.

Like in HW2, you will now verify the correctness of your implementation for the loss and its gradient by plotting the residuals $\| \underbrace{\nabla_{\underline{\boldsymbol{w}}} L(\underline{\boldsymbol{w}}, b)}_{\text{analytic}} - \underbrace{u_{\delta_w}(\underline{\boldsymbol{w}}, b)}_{\text{numeric}} \|_2$ and $| \underbrace{\frac{\partial}{\partial b} L(\underline{\boldsymbol{w}}, b)}_{\text{analytic}} - \underbrace{u_{\delta_b}(\underline{\boldsymbol{w}}, b)}_{\text{numeric}} |$ as a function of δ_w, δ_b (respectively; over many repeats).

Task: Copy the functions from the given *verify_gradients.py* into your notebook / project.
Read and understand these functions but do not edit them.

(Q4) Using your preprocessed (and normalized) dataset and `CovidScore` as our target, generate a plot that compares the numerical gradients to the analytical ones.
Do this by running the following command:

```
compare_gradients(X_train, y_train, deltas=np.logspace(-7, -2, 9))
```

Important: `X_train` should hold the features of your **normalized** training subset.
`y_train` should hold the `CovidScore` subset training labels.

Attach the plots to your report. No need to discuss these plots, but you should make sure that they make sense.

Task: Copy the function given in *test_lr.py* into your notebook / project.
Read and understand the function but do not edit it.

(Q5) We now want to evaluate the effects of different learning rates on the learning procedure. Run the following command that plots a graph of the training and validation losses as a function of the iteration number for different learning rates.

```
test_lr(X_train, y_train, X_val, y_val)
```

Important: `X_val` should hold the features of your (preprocessed) validation subset.
`y_val` should hold the `CovidScore` subset validation labels.

Note: If your model did not converge with any learning rate, you are allowed to alter the `lr_list` variable (but explain this in your report).

This part should also help you verify your implementation (loss should decay).

Attach the plots to your report, briefly discuss the results and justify the demonstrated behaviors. Which learning rate is "optimal" (briefly explain)?
For the best learning rate, does it make sense to increase the number of gradient steps (instead of the default 1500 steps)? Explain.

Now that we have experienced solving and tuning the least squares problem, we are ready for the rest of the assignment… and save the world from COVID!

Section 3: Evaluation and Baseline

Our general goal is to minimize the generalization MSE, that is $\mathbb{E}_{(x,y)\sim D}[(h(x)-y)^2]$.

As we have learned, in practice, we instead minimize the empirical error $\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2$ on a training set, and tune hyperparameters using a validation set.

In the rest of this assignment, we use the k-fold cross-validation method for better estimating the generalization error, thus improving the tuning procedure. We will use $k=5$ folds.

Similar to HW2, we use `sklearn` to perform cross-validation on the (whole) training set to evaluate the performance of models. As explained earlier, the metric we use for regression is the MSE (using `cross_validate`, set `scoring='neg_mean_squared_error'`).

Simplest baseline

We now train a simple DummyRegressor that always predicts the average `CovidScore` of the training set. We will use this regressor throughout the assignment as a baseline to which we will compare the performance of our learned regressors.

(Q6) Create a DummyRegressor. Evaluate its performance using `cross-validation`. In your report, fill in the cross-validated errors of the regressor.

Model	Section	Train MSE	Valid MSE
		Cross validated	
Dummy	2		

Task: Retrain the dummy regressor on the entire training set (= all its samples) and save it for future use (Sec 7).

Basic hyperparameter tuning

A quick reminder of the tuning process for hyperparameters:

The repeated tuning process (for a single parameter) should include:

i. Determining the tested values of the tuned hyperparameter (see numpy.logspace). You need to choose suitable values by yourself that will help you optimize the validation error.

ii. For each value, evaluating a suitable regressor using cross validation ($k = 5$).

iii. Plotting (cross-validated) train and validation errors as a function of the hyperparameter. Consider using semilogx or loglog plots.
Also plot a constant line with the validation error of the dummy regressor.

iv. Reporting the value that yields the optimal validation error and its respective error.

(Q7) Create an instance of your custom `LinearRegressor` and evaluate its performance using `cross-validation`, **remember to tune the LR!**
Report the appropriate plots and values detailed above in "the repeated tuning process". Fill in the cross-validated errors of the regressor yielded by the optimal hyperparameter.

Model	Section	Train MSE	Valid MSE
		Cross validated	
Dummy	3	filled	filled
Linear	3		

(Q8) Had we chosen not to normalize features beforehand, would the training performance of these two models have changed (assume there are no numerical errors)? Explain.

Task: Using the best performing hyperparameter, **retrain** the regressor on the entire training set (= with all its samples) and save it for future use (Sec 7).

Section 4: Ridge linear regression

In regularized linear regression, we need to tune the regularization strength (λ in our notation, `alpha` in sklearn). As mentioned earlier, k-fold cross-validation is performed with the **entire** training set (= all its samples) for the remainder of the experiments.

We will now learn to predict `CovidScore` using the sklearn.linear_model.Ridge regressor. Make sure your models are non-homogeneous (`fit_intercept=True`).

(Q9) Tune the regularization strength of the regressor. Follow the repeated tuning process described earlier. Remember to attach the required plot and specify the optimal strength with its validation error.

(Q10) Fill in the cross-validated errors of the regressor yielded by the optimal hyperparameter.

Model	Section	Train MSE	Valid MSE
		Cross validated	
Dummy	3	filled	filled
Linear	3	filled	filled
Ridge linear	4		

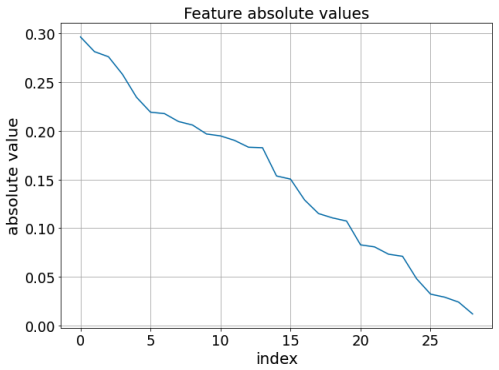
Task: Using the best performing hyperparameter, **retrain** the regressor on the entire training set (= with all its samples) and save it for future use (Sec 7).

(Q11) Specify the 5 features having the 5 largest coefficients (in absolute value) in the resulting regressor, from the largest to the smallest (among these 5).

We now wish to visualize the resulting coefficients.

(Q12) Sort and plot the absolute values of the learned coefficients. The x-axis should be the index of the parameter from largest to smallest.

The result should roughly look like the illustration to the right (the values are made up).



(Q13) Why is the magnitude of the coefficients interesting? Explain briefly.

(Q14) Had we chosen not to normalize features beforehand, would the training performance of the Ridge model have changed (assume there are no numerical errors)? Explain

Section 5: Polynomial fitting (visualization)

As a detour to better understand polynomial fitting for regression, we now focus on regressing the `SpreadScore`, using `sugar_levels` and `PCR_05` only.

Task: Create a subset of the training set, containing only those 2 features and `SpreadScore`.

Task: Visualize the data using `plot3d`, i.e., plot `SpreadScore` as a function of the `sugar_levels` and `PCR_05` features. (for now, don't use the `predictions` argument.)

(Q15) Attach the 3-d plot to your report. What can we understand from this visualization? How should this affect our choice of model to regress `SpreadScore`?

(Q16) We will once again create a baseline, this time our baseline will consist solely of (linear) Ridge regression (using the two features only), remember again to make sure your models are non-homogeneous (`fit_intercept=True`).

Tune for regularization strength as before. Remember to attach required plots, optimal strengths, and optimal validation errors (on this new target).

Task: Using the best performing hyperparameter, **retrain** the ridge regressor on the entire training set (= with all its samples).

(Q17) We will now visualize the model you just trained. Use `plot3d` to plot all the training samples and their true labels (as before). Pass your model predictions in a list to the `predictions` keyword to compare the true labels to the predicted values.

Attach the plot to your report.

We will now try to improve our linear ridge-regression model by adding quadratic features.

Task: Create a single model that: (1) transforms the original features into 2nd-degree polynomial features, (2) normalizes the transformed features, and then (3) trains the linear Ridge regressor on these normalized polynomial features. To do so, we will use a Pipeline like we did in HW2. The following snippet creates such a pipeline:

```
poly_reg = Pipeline([('feature_mapping', TODO),
                     ('normalization', TODO),
                     ('Ridge', Ridge(alpha=Lambda, fit_intercept=True)
```

Complete the first two steps in the pipeline above to make it apply a 2nd-degree PolynomialFeatures transformation, followed by a normalization of your choosing (tip: you are allowed to simply use a `StandardScaler` for all the transformed features).

(Q18) Why is re-normalization important after applying a polynomial mapping?

Hint: Your answer should revolve around the degrees of the different monomials and focus on mathematical/"optimizational" considerations rather than on statistical ones like in Lecture 08.

(Q19) Tune the regularization strength of a Ridge regressor with the polynomial mapping using cross validation ($k = 5$ as always). (Remember: in Q17 of HW2 you saw how to set hyperparameters inside a pipeline.)

Remember to attach required plots, optimal strengths, and optimal validation errors.

Task: Using the best performing hyperparameter, **retrain** the ridge regressor on the entire training set (= with all its samples).

(Q20) We will now visualize the polynomial model you just trained. Use the new `plot3d` function to plot all the training samples and their true labels (as before). Pass your model predictions in a list to the `predictions` keyword to compare the true labels to the predicted values. Attach the plot to your report.

(Q21) Discuss the results from this section. Compare the two models' capacities to fit the data at hand (use the visualizations, MSE, etc.). This should take 3-6 sentences.

Section 6: Polynomial fitting of the CovidScore

Back to fitting a regression model for `CovidScore`.

Task: Create copies of your train and test sets. Run a pipeline like the one in the former section (polynomial mapping, normalization, ridge regression model) on all features in your (preprocessed) dataset. For simplicity, you can use a MinMaxScaler in your pipeline (on all transformed features) to make sure the OHE features remain binary.

We denote the class of all possible (2nd-degree) hypotheses by $\mathcal{H}_{\text{poly}}$.

(Q22) Explain how (and why) we can expect the training and validation errors (each) to change when using such a mapping (compared to the raw data).

One student suggested using the answer to (Q1) to create a hierarchical model (i.e., a multilevel model) rather than using a polynomial mapping. In the hierarchical model, samples are first divided into multiple subsets, depending on the condition from (Q1). Then, a different linear regressor is used for each subset (multiple linear regressors).
That is, we can learn separate $\underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}}_2, \underline{\boldsymbol{w}}_3, b_1, b_2, b_3$ and build a "conditioned" hypothesis:

$$\mathcal{H}_{\text{multi}} \ni h_{\text{multi}}(\underline{\boldsymbol{x}}) = \begin{cases} \underline{\boldsymbol{w}}_1^{\top}\underline{\boldsymbol{x}} + b_1, & \underline{\boldsymbol{x}} \text{ is in the O blood type group} \\ \underline{\boldsymbol{w}}_2^{\top}\underline{\boldsymbol{x}} + b_2, & \underline{\boldsymbol{x}} \text{ is in the B blood type group} \\ \underline{\boldsymbol{w}}_3^{\top}\underline{\boldsymbol{x}} + b_3, & \underline{\boldsymbol{x}} \text{ is in the other blood type group} \end{cases}$$

The student claimed that since we added the indicator-features in (Q1), a hierarchical model (using the raw features) is a special case of the 2nd-degree polynomial mapping (in the sense that an equivalent model can be learned by fitting 2nd-degree polynomials).

(Q23) Explain why the student is correct. Show mathematically that $\mathcal{H}_{\text{multi}} \subset \mathcal{H}_{\text{poly}}$.

(Q24) Briefly specify and explain some upsides and downsides of learning with $\mathcal{H}_{\text{multi}}$ comparing to learning with $\mathcal{H}_{\text{poly}}$.

(Q25) Tune the regularization strength of a Ridge regressor with the polynomial mapping using cross validation. Use a similar pipeline to the one from **(Q19)**.
Remember to attach required plots, optimal strengths, optimal validation errors.

(Q26) Fill in the train and validation errors of the regressor yielded by the best performing hyperparameter. Remember to compute the errors using cross-validation.

Model	Section	Train MSE	Valid MSE
		Cross validated	
⋮	⋮	filled	filled
Ridge Polynomial			

Task: Train the polynomial regressor on the entire training set and save it for future use (Sec 7).

Section 7: Testing your models

Important: do not continue to this section until you have finished all previous sections.

Finally, we can let the **test set** come out and play.

At the end of the previous sections, you retrained the tuned models on the entire training set. You will now evaluate the test errors (MSE) for the models, as well as the Dummy baseline.

Remember: do not cross-validate here. Train on the entire training set (= using all its samples) and evaluate on the test set.

(Q27) Complete the entire table

Model	Section	Train MSE	Valid MSE	Test MSE
		Cross validated		Retrained
Dummy	3	filled	filled	
Linear	3	filled	filled	
Ridge linear	4	filled	filled	
Ridge Polynomial	6	filled	filled	

Which model performed best on the test set?

Briefly discuss the results in the table (from an overfitting and underfitting perspective, or any other insightful perspective).