Machine Learning for Bioinformatics

Exercise Sheet

Freie Universität Berlin, SoS 2024

Week 7 · Assignment on 29.05.2024. Submit until 07.06.2024 11:00 p.m.

Please note that the jupyter notebook must be submitted
along with the exercise sheet![1]

Name: Evita Vardhani Matriculation no.:
Name: Matriculation no.:

Linear and kernel regression basics

In this exercise, we work with a simple toy data set to first explore polynomial regression. Afterwards, we convert the regression problem into a kernel regression model and compare performances between the two methods. You will be implementing both methods in the accompanied Jupyter notebook.

Assignment 1. *From polynomial regression to kernel regression*

1. Complete the implementation for computing polynomial features. Points: 0
2. Complete the implementation of the *PolynomialRegression* class. Points: 0
3. Please specify the results of the grid search with LOO-CV. Points: 1
   Optimal degree: 10 Minimum MSE: 0.4169
4. Complete the implementation of the *KernelPolynomialRegression* class. Points: 0
5. Please specify the results of sweeping the degree with LOO-CV for the kernel regression method. Points: 1
   Optimal degree: 1 Minimum MSE: 0.3407
6. Please specify the results of sweeping α with LOO-CV. Points: 1
   Minimum MSE for polynomial regression: 0.4123
   Minimum MSE for kernel regression : 0.3838
7. Given a matrix $A \in \mathbb{R}^{p \times p}$. When do we need to use the Moore-Penrose inverse X^+ instead of X^{-1}? Points: 1
   $\operatorname{rank} A < p$: ✔ $\operatorname{rank} A > p$: $\operatorname{rank} A = p$:

[1] No points are awarded if an answer is only partially correct. Answers must be supported by results from the jupyter notebook.

8. Let $A = X^\top X$ and $X \in \mathbb{R}^{n \times p}$ with $\operatorname{rank} X = \min(n, p)$. What is the rank of A if $n < p$? Points: 1

p: ✔ n:

9. Assume that $n < p$. Which of the following satements is correct? Points: 1

   For linear regression, the Moore-Penrose inverse can be used to compute the minimum ℓ_2-norm solution.

   Correct: ✔

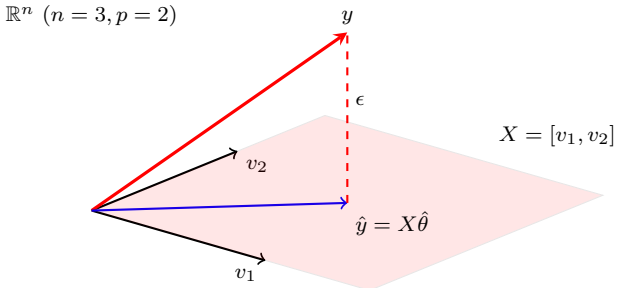
   Kernel regression is computationally more efficient than linear regression.

   Correct: ✔

10. Maximizing the likelihood of linear regression is a convex (or concave) optimization problem. Points: 1

    Correct: Incorrect: ✔

11. Recall the following picture from our lecture, which illustrates the OLS projection



   Assume the special case where $v_1 = (1, 0, 0)^\top$ and $v_2 = (0, 1, 0)^\top$. Derive $\left\|y - X\hat{\theta}\right\|_2$ for $y = (y_1, y_2, y_3)^\top$, which is the length of ϵ at the OLS estimate $\hat{\theta}$. The solution a simple expression, which can be easily derived with geometric arguments! Points: 1

$$\left\|y - X\hat{\theta}\right\|_2 = \text{sqrt(v_1**2+v_2**2) (tan 45)}$$

Applied kernel regression

Every gene is flanked by a promoter sequence where transcription factors initiate transciption. Promoters typically contain multiple so called transcription factor binding sites, which consist of short DNA subsequences that are recognized by

transcription factors as binding sites. Each type of transcription factor recognizes a different DNA subsequence as binding site. Each cell type has a different set of transcription factors present in the nucleus, which hence activate a different set of genes. This means that from the DNA sequence of promoters we might be able to predict the expression of genes. For many transcription factors, the preferred binding sites are known. Our data (X, y) with $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ consists of the predictors X and the expression values y for n genes. The matrix X contains scores for each gene (rows) and each transcription factor (columns). The score x_{ij} is high if the promoter of gene i contains a binding site for transcription factor j.

Assignment 2. *Prediction of gene expression from promoter sequences*

1. Specify the performance of the linear regression method. Points: 1

   Mean R^2: O(p**3)

2. Specify the optimal parameter for the kernel regressor with RBF kernel. Points: 1

   Optimal α: grid_search.

3. The mean of the data is a better predictor than a linear regression model. Points: 1

   Correct: ✔ Incorrect:

4. Assume that you want to summarize your results for a scientific report or paper. Which of the following claims can you report? Points: 1

   Gene expression is fully controlled by promoters: ✔

   Promoters contain almost no information about gene expression:

   The results are inconclusive:

Maximum number of points: 12