Elements of Machine Learning, WS 2023/2024
Jilles Vreeken and Krikamol Muandet
Exercise Sheet #4: *Beyond Linear*

Deadline: Thursday, January 04, 2024, 15:00

Before solving the exercises, read the instructions on the course website.

- For each theoretical problem, submit a single `pdf` file that contains your answer to the respective problem. This file may be a scan of your (legible) handwriting.
- For each practical problem, submit a single `zip` file that contains
  - the completed jupyter notebook (`.ipynb`) file,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows all your results.
- For the bonus question, submit a single `zip file` that contains
  - a `pdf` file that includes your answers to the theoretical part,
  - the completed jupyter notebook (`.ipynb`) file for the practical component,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows your results.
- **Every team member** has to submit a signed Code of Conduct.
- **IMPORTANT** You must make the team on CMS *before* you upload the solutions. If you upload the solutions first and create the team after it, the solution will not show for the new team member!

Problem 1 (T, 3 Points). **Parametric or Non-parametric**

1. [$1pts$] Describe, in your own words, what the difference between a parametric and a non-parametric method is.
2. [$2pts$] For each of the following methods you have learned about in the lectures so far, decide if it is parametric or non-parametric and *explain your reasoning*:
   - LASSO
   - Smoothing Splines
   - Local Regression
   - Generalized Additive Models

Problem 2 (T, 3 Points). **Generalized Additive Models.**

In Generalized Additive Models (GAMs), we are interested in predicting our target variable $Y \in \mathbb{R}$ based on the variables $X_1, \ldots, X_p$ as follows:

$$g(Y) = \alpha + \sum_{j=1}^{p} f_j(X_j) ,$$

where we assume that $\mathbb{E}(f_j(X_j)) = 0$ for all j. For the rest of this exercise, we will assume $g = \text{id}$ to be the identity function and that the dimensionality of X is $p = 2$.

1. [$1pts$] Without proof, will iterating this algorithm produce the same result as one of the methods you have learned about in class? Explain your reasoning.
2. [$1pts$] Under which conditions will the results of the backfitting algorithm with this smoothing operator $\mathcal{S}_\lambda$ depend on the order of which $\hat{f}_j$ is updated first?

3. [$1pts$] Write down the smoothing operator based on cubic smoothing splines.

Problem 3 (T, 4 Points). **Dependencies.**
Consider the following small example with $p = 2$ predictors and $n = 2$ samples. Suppose that $x_{11} = x_{12}, x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimated intercept in a least squares model is zero, $\hat{\beta}_0 = 0$.

1. [$2pts$] What is the linear regression solution $\hat{\beta}$ in this case?
2. [$2pts$] What problem do you see?

Problem 4 (T, 10 points). **Polynomial regression & Splines.**

1. [$1pt$] In which sense is polynomial regression linear respectively non-linear?
2. [$2pt$] Given a set of basis function $\mathcal{B}_n = \{a_i x^i | a_i \in \mathbb{R}\}_{i=0}^n$ for polynomial regression of degree $n \in \mathbb{N}$, give 4 functions that can not be expressed using $\mathcal{B}_n$ and argue why they can not be expressed?
3. [$1pt$] For this exercise consider Figure 1:
   (a) Determine a spline with appropriate degree that describes the function in Figure 1.
   (b) Determine a piecewise polynomial that describes the function in Figure 1.

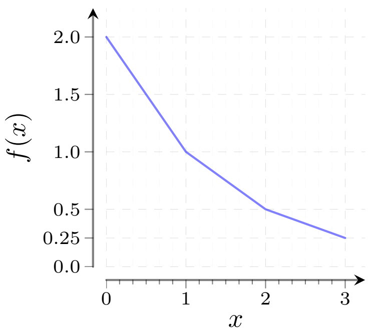


Figure 1: Function for Problem 1.3.

4. In the lecture we defined cubic splines as functions of the form:

$$f_L(x) = a + bx + cx^2 + dx^3 + e(x - \zeta)_+^3 \, .$$

Alternatively, cubic splines can also be constructed using piecewise defined polynomials of degree 3 with the condition that at the knot the polynomial is is twice continously differentiable. With this definition cubic splines take the form:

$$f_I(x) = \begin{cases} \alpha_1 + \beta_1 x + \gamma_1 x^2 + \delta_1 x^3 & x \leq \zeta \\ \alpha_2 + \beta_2 x + \gamma_2 x^2 + \delta_2 x^3 & x > \zeta \end{cases} \, .$$

Show that both functional representations of cubic splines express the same function space. For this you have to proof that both representations can be transformed in to each other. In other words, show the following two directions:

   (a) [$2pts$] Given a function of the form $f_L(x)$ constuct and equivalent $f_I(x)$.

(b) [$3pts$] Given a function of the form $f_I(x)$ constuct and equivalent $f_L(x)$.

Hint: For the second part exploit that $f_I(x)$ is twice cont. differentiable and $f_L(x)$ can be written as $f_L(x) = \alpha_1 + \beta_1 x + \gamma_1 x^2 + \delta_1 x^3 + (\delta_2 - \delta_1)(x - \zeta)^3_+$.

5. [$1pts$] Give an example function, where the continuity requirement is a problem. Use the graph of the function to argue why this is the case.

Problem 5 (P, 15 Points). **Regression Splines**
Please rename the file to include the matriculation numbers of all team members (e.g. `7010000_2567890_A4.ipynb`**).** Use `Practical_Problem_4.ipynb` file from the course website.

In this exercise, we will examine different variations of regression splines. As dataset you are provided in `spline_data.csv` with a weather dataset to fit different variations of splines to predict the temperature.

1. (7 Points) Implement a piecewise polynomial regression that divides up the data into different bins for a given set of knots and fits in each bin locally a polynomial regression model. Inside each bin, the polynomial is fitted with the least-squares method (compute closed-form solution from lecture, do not use numpy.polyfit or similar) without any continuity constraints at the knots. Your function takes as input the data, the knots and the order of the polynomial and should return the predictions of the fitted model on the data.

2. (2 Points) Evaluate your model thoroughly on the provided dataset. Compare the impact of the order of the local polynomial (linear vs cubic) on the quality of the model. In addition, you are provided with two sets of knots. What is the importance of the placement of the knots and why is one set of knots better than the other?

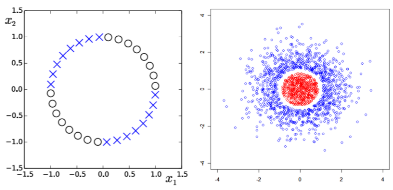
3. (3 Points) In the following we will use *smoothing splines* to fit a model on the same dataset. Use the *UnivariateSpline* function from the *scipy.interpolate* package to fit a model on the data and visualize the results. What is the main difference and advantage of smoothing splines compared to piecewise polynomial regression?

4. (3 Points) The smoothing parameter s controls the smoothness of the fit. It is a hyper-parameter that needs to be tuned. Use 5 fold cross-validation to find the optimal value for s. Use the mean-squared error to measure the quality of the model.

Problem 6 (Bonus). **Kernelization for Non-Linear Regression**
Kernels allow a way of implicitly representing nonlinear transformations of data. Kernels are functions that usually define similarities between inputs in some space $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ i.e. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The implicit transformation $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}$ from input space to some other space is used to define a kernel as follows

$$k(\mathbf{x}, \mathbf{x}') := \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}')$$

1. Give an example of a transformation $\boldsymbol{\phi}$ that acts on $\mathbf{x} \in \mathbb{R}^2$ which makes the data in the example below linearly separable (crosses in one class, circles in the other). Similarly, give an example of a transformation that makes the red and blue classes linearly separable.



   (Hint: Linear separability means that there exists a hyperplane that can separate the classes in the transformed space.)

2. Find the feature transformation $\boldsymbol{\phi}(\mathbf{x})$ corresponding to the kernel $k(x_1, x_2) = \frac{1}{1-x_1 x_2}$, with $x_1, x_2 \in (0, 1)$.

3. Given a regression task, show that for $n \in \mathbb{N}$ and $a_i \geq 0$ for $i = 0, \ldots, n$, the following kernel k

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{n} a_i (\mathbf{x}^T \mathbf{x}')^i + a_0,$$

   with $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}$ does polynomial regression. Also, evaluate the predictor corresponding to the kernel i.e. $y = \boldsymbol{\beta}^T \boldsymbol{\phi}(\mathbf{x})$ where $\boldsymbol{\beta}$ are the parameters in the space of implicit transformation.

4. Find the feature transformation $\boldsymbol{\phi}(\mathbf{x})$ corresponding to the above kernel if $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$.