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

Problem 1 (For Tutorials on 18.12 and 19.12). **Splines.**
In the lectures, you have learned that the space of cubic splines with K knots has dimension $K + 4$. But how did we arrive at this number?

1. Assume that we have $K = 1$ knot, and let ζ be this knot. Further, let the spline be written as

$$f(x) = \begin{cases} a_3x^3 + a_2x^2 + a_1x + a_0, & x \le \zeta \\ b_3(x-\zeta)^3 + b_2(x-\zeta)^2 + b_1(x-\zeta) + b_0, & x > \zeta. \end{cases}$$

   Let $a_0, \ldots, a_3$ be given. Show that b_0, b_1, b_2 are fully determined by the constraint imposed by f being twice differentiable at $x = \zeta$. What does this mean for the degrees of freedom of the model?

2. Write down a similar presentation for a *quadratic* spline with $K = 1$ knot at ζ. How many parameters does this model have under the requirement that the spline be differentiable *once*?

3. How large is the difference in free parameters between quadratic and cubic splines? Is this difference bigger than you would intuitively expect? Is it smaller?

4. Explain why cubic splines would be more suitable than quadratic splines for the data in Fig. 1. The $K = 2$ knots ζ_1, ζ_2 are located as shown.

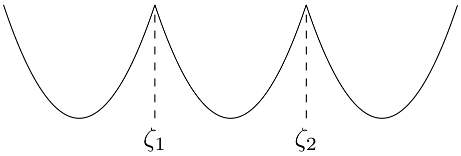

Figure 1: Three Parabolas side by side.

Solution.

1. f being twice differentiable at $x = \zeta$ means that the zeroth, first and second derivatives of both cases in the definition of f match at this point. The constraints are therefore

$$\begin{aligned} a_3\zeta^3 + a_2\zeta^2 + a_1\zeta + a_0 &= b_0 && f \\ 3a_3\zeta^2 + 2a_2\zeta + a_1 &= b_1 && f' \\ 6a_3\zeta + 2a_2 &= 2b_2 && f'' \end{aligned}$$

so that given $a_0, \ldots, a_3$ the only free parameter is b_3. We therefore have only $K = 1$ additional degree of freedom in choosing our spline.

2. For the quadratic spline with one knot, we can write

$$f(x) = \begin{cases} a_2x^2 + a_1x + a_0, & x \leq \zeta \\ b_2(x-\zeta)^2 + b_1(x-\zeta) + b_0, & x > \zeta. \end{cases}$$

Then the zeroth and first-order constraints are

$$\begin{aligned} a_2\zeta^2 + a_1\zeta + a_0 &= b_0 \\ 2a_2\zeta + a_1 &= b_1 \end{aligned}$$

so again, the only free parameter given $a_0, \ldots, a_2$ would be b_2. Therefore we have $K + 3 = 4$ free parameters a_0, a_1, a_2, b_2.

3. The result that cubic splines have $K+4$ free parameters and quadratic splines have $K+3$ parameters is rather unintuitive. After all, for large K we have “almost” the same number of parameters—relative to how many parameters we have in total. That is, for each of the $K+1$ intervals separated by the K knots we have on average

$$\begin{aligned} \frac{K+4}{K+1} &= 1 + \frac{3}{K+1} && \text{(cubic)} \\ \frac{K+3}{K+1} &= 1 + \frac{2}{K+1} && \text{(quadratic)} \end{aligned}$$

free parameters, which for large K is not very much of a difference.
And yet, in every single interval the function is a cubic polynomial, which we would expect to be much more expressive than a quadratic polynomial.

4. While the displayed data is piecewise quadratic, a quadratic spline is not suitable for modeling it. The issue is the following: a quadratic spline can't have an inflection point in its interior. That is, once the left-most parabola is oriented, the orientation of the second parabola is fixed as shown in the Fig. 2 on the left. Therefore, when fitting a quadratic spline to the data, the best we can do (in terms of orientation, not scale), is to direct the first and third parabolas correctly, and hope the second isn't too bad.

   In contrast, with cubic splines we do not have this issue, as seen in Fig 2 on the right. Since cubic splines can have an inflection point on the inside, they can turn around and capture more of the parabola in this way.

   Of course, at the end of the day, when constraints (e.g. natural splines) or regularization (e.g. smoothing splines) are included, the difference between the two models becomes smaller.

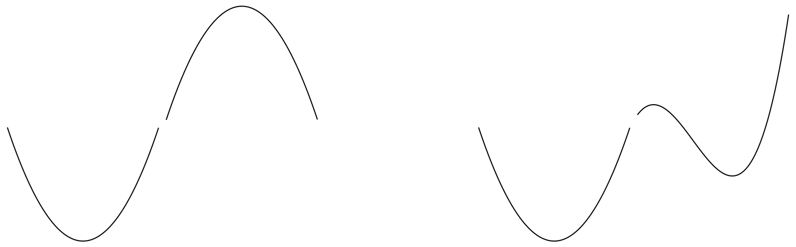

Figure 2: A representation of why quadratic splines do not do well at capturing three parabolas with the same direction, while cubic splines have a better ability to capture them. Left: The first parabola completely determines the orientation of the second parabola. Right: With well-chosen parameters, the left-most part going in the wrong direction can be made very small, while the correctly-oriented part can be made to capture the second parabola arbitrarily well.

Problem 2 (For Tutorials on 18.12 and 19.12). **Generalized Additive Models.**
In Generalized Additive Models (GAMs), we are interested in predicting our target variable $Y \in \mathbb{R}$ based on the variables $X_1, \dots, X_p$ 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 $g = \mathrm{id}$ to be the identity function and that the dimensionality of X is $p = 2$.

1. In the lecture we have seen the backfitting algorithm. Let the smoothing operators $\mathcal{S}_j = \mathcal{S}_\lambda$ for $\lambda \geq 0$ take the following form

$$\hat{\beta}_j = \arg\min_{\beta} \sum_i \left(y_i - \alpha - \sum_{k:k \neq j} \hat{f}_k(x_{ki}) - \beta x_{ji} \right)^2 + \lambda \beta^2$$
$$\hat{f}_j(X_j) = \hat{\beta}_j X_j \; .$$

Write out the first iteration of the backfitting algorithm. That is, compute the parameters $\hat{\beta}_j$ for both $\hat{f}_1$ and $\hat{f}_2$ after the first iteration of the algorithm.

Solution.

1. We know that in the first step, $\hat{f}_2 = 0$ at the start, so that we can write

$$\begin{aligned}\hat{\beta}_1 &= \arg\min_{\beta} \sum_i (y_i - \alpha - \beta x_{i1})^2 + \lambda\beta^2 \\ &= \frac{x_{:,1}^\top y}{x_{:,1}^\top x_{:,1} + \lambda} \ ,\end{aligned}$$

where $x_{:,1}$ denotes the vector containing all observations of the first feature. Consequently, the residual is $y' = y - \hat{\beta}_1 x_{:,1}$. plugging this into the update for $\hat{\beta}_2$ we obtain

$$\begin{aligned}\hat{\beta}_2 &= \frac{x_{:,2}^\top y'}{x_{:,2}^\top x_{:,2} + \lambda} \\ &= \frac{x_{:,2}^\top y}{x_{:,2}^\top x_{:,2} + \lambda} - \hat{\beta}_1 \frac{x_{:,2}^\top x_{:,1}}{x_{:,2}^\top x_{:,2} + \lambda} \ ,\end{aligned}$$

which is an adjusted version of the estimate $\frac{x_{:,2}^\top y}{x_{:,2}^\top x_{:,2}+\lambda}$ obtained by “correcting for” the correlation between the variables X_1 and X_2.