CMPE 482 - Assignment 3

Spring 2021

Due 05.05.2021, 20.00

1. **(2 points)** Download the Sacramento house sales data provided in the Python companion for our book [*IALA*]. As in the book, we will use this data set to attempt to predict house prices based on the features we are provided. Conduct the feature engineering steps described in our book's Chapter 13.1.2 - you can use the relevant code in the Python companion for this. As in the book, you should end up with a total of 8 basis functions. Take the *first* 600 observations as your training set, and the *last* 174 observations as the test set (If you are unsure about concepts such as *training set, test set, generalization* etc. read Section 13.2 of your book before beginning this assignment).
    - **a.** Use your training set to find and print the least squares estimate $\hat{\theta}$.
    - **b.** Using $\hat{\theta}$, find and print the RMS on the training set and the test set.
    - **c.** Examine the usefulness of the basis function f_3 suggested by the authors in terms of generalization. Create a second version of your data, this time dropping the column that correspond to the basis function f_3 described in your book. Find a new $\hat{\theta}'$ using least squares method with this version of the data. Again compute the training set RMS and test set RMS of your predictions, and compare these with the results of your previous model that produced $\hat{\theta}$. Based on your analysis, is the inclusion of basis function f_3 useful? What did you base your decision on?
2. **(2 points)** Continue using your data from Q1 (the version that includes f_3). We will now convert this problem to a multi-objective least-squares problem by including a regularization term, as described in Section 15.3.1 in your book under the title *Tikhonov regularized inversion*.
    - **a.** Find and print $\hat{\theta}_1$ and $\hat{\theta}_2$ using regularized least squares, by setting λ equal to $.00001$ and 100000, respectively. Compare $\hat{\theta}_1$ and $\hat{\theta}_2$ to the original $\hat{\theta}$ you found in Q1. Which one is closer to $\hat{\theta}$? What does the other one look like? Why?
    - **b.** Find the value of λ that is best for *generalization* among the candidates $\{0, .001, .01, .1, 1.\}$. Print the training RMS and test RMS for all values of λ, and clearly state your chosen λ as the best one.
3. **(2 points)** Continue using your data from Q1 (the version that includes f_3). We will now add polynomial and interaction terms to our basis functions as further feature engineering.
    - **a.** Temporarily ignoring the training/test set distinction, create polynomial and interaction terms of the original basis functions by using their multiplications with each other as new bases. You should end up with 774 observations with 36 basis functions such that (ignoring x's for brevity): $f_1, f_2, \ldots, f_1^2, f_2^2, \ldots, f_1 f_2, f_1 f_3, \ldots, f_7 f_8$. (Note: There are practical reasons specific to this data set that makes this a bad idea, but we will ignore them for the sake of the example.)
    - **b.** Using the training/test set distinction in Q1 (first 600 - last 174), and with $\lambda = .1$, decide whether this new set of features are an improvement on the original set in terms of generalization.

c. Imagine we only had access to 30 observations as our training set, and we still wanted to use our 36 basis features. Examine the equation at the end of Section 15.3.1. Can we still find a solution with $\lambda > 0$ using this equation? Can we find a solution with $\lambda = 0$ using this equation? If we cannot, what alternative do we have for $\lambda = 0$ (you may assume that the basis columns are independent for this last question)?

4. (4 points) ***Piecewise polynomial interpolation***. For this question, you are expected to fit a piecewise polynomial $y \approx \hat{f}(x)$ to a set of points $\{(x_i, y_i)\}_{i=1}^{N}$. Let $p_1(x)$, $p_2(x)$ and $p_3(x)$ be polynomials of degree $D = 4$, and let $\hat{f}(x)$ be defined as follows:

$$\hat{f}(x) = \begin{cases} p_1(x) = \sum_{j=0}^{D} c_j x^j, & -1.5 \le x < -0.5 \\ p_2(x) = \sum_{j=0}^{D} c_{D+1+j} x^j, & -0.5 \le x < 0.5 \\ p_3(x) = \sum_{j=0}^{D} c_{2D+2+j} x^j, & 0.5 \le x < 1.5 \end{cases}$$

Our goal for this question will be to find the coefficient vector $\mathbf{c}$ which minimizes the squared error $E = \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2$. While doing so, we require $\hat{f}(x)$ to be ***continuous***, and also to have continuous ***first*** and ***second*** derivatives.

a. For $N = 150$, sample $x_1, x_2, \ldots, x_N$ from the uniform distribution $\mathcal{U}[-1.5, 1.5]$, and sort them so that $x_1 \le x_2 \le \cdots \le x_N$. For each x_i, compute $y_i = x_i(e^{-x_i^2} + \cos 6x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. Create a scatter plot of $\mathbf{x} = [x_1, \ldots, x_N]^T$ vs $\mathbf{y} = [y_1, \ldots, y_N]^T$ for the points you generated.

b. Suppose $x_{M_1} < -0.5 \le x_{M_1+1}$ and $x_{M_2} < 0.5 \le x_{M_2+1}$ for some integers M_1 and M_2. Derive the necessary conditions for satisfying continuity requirements, and express the piecewise polynomial fitting problem as an optimization problem in the form of (see page 341 of *IALA*):

$$\begin{aligned} &\text{minimize } \|A\mathbf{c} - \mathbf{y}\|^2 \\ &\text{subject to } C\mathbf{c} = \mathbf{d} \end{aligned}$$

c. Form A, C, and $\mathbf{d}$, and then plot the heatmaps of A^T and C.

d. Solve the optimization problem for $\mathbf{c}$ using *constrained least squares via QR factorization* algorithm (see Algorithm 16.2 on IALA). Output the vector $\mathbf{c}$ in the form of a stem-plot.

e. Plot the resulting $\hat{f}(x)$ along with the scatter plot of $\mathbf{x}$ vs $\mathbf{y}$.

f. Verify that the resulting $\mathbf{c}$ satisfies the *KKT conditions*, i.e. there exists a vector $\mathbf{z}$ for $\mathbf{c}$ such that

$$\begin{bmatrix} 2A^T A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} 2A^T \mathbf{y} \\ \mathbf{d} \end{bmatrix}$$

Output the vector $\mathbf{z}$ in the form of a stem-plot, and show that the equality above is satisfied.

g. For which values of D would we still have *feasible* solutions? Explain your answer briefly for the cases where D is equal to 0, 1, 2 and 3.