Homework 3

AMATH 582/482, Winter 2022

Assigned Feb 11, 2022. Due on Feb 25, 2022 at midnight.

Directions, Reminders and Policies

Read these instructions carefully:

- **You are required to upload a PDF report to Canvas along with a zip of your code. Note the PDF and the zip should be uploaded separately.**
- The report should be a maximum of 6 pages long with references included. Minimum font size 10pts and margins of at least 1inch on A4 or standard letter size paper.
- Do not include your code in the report. Simply create a zip file of your main scripts and functions, without figures or data sets included, and upload the zip file to Canvas.
- Your report should be formatted as follows:
  - Title/author/abstract: Title, author/address lines, and short (100 words or less) abstract. This is not meant to be a separate title page.
  - Sec. 1. Introduction and Overview
  - Sec. 2. Theoretical Background
  - Sec. 3. Algorithm Implementation and Development
  - Sec. 4. Computational Results
  - Sec. 5. Summary and Conclusions
  - Acknowledgments (no more than four or five lines, also see the point below on collaborations)
  - References
- I suggest you use LaTeX(Overleaf is a great option) to prepare your reports. A template is provided on Canvas under the Syllabus tab. You are also welcome to use Microsoft Word or any other software that properly typesets mathematical equations.
- I encourage collaborations, however, everything that is handed in (both your report and your code) should be your work. You are welcome to discuss your assignments with your peers and seek their advice but these should be clearly stated in the acknowledgments section of your reports. This also includes any significant help or suggestions from the TAs or any other faculty in the university. You don't need to give all the details of the help you received, just a sentence or two.
- Your homework will be graded based on how completely you solved it as well as neatness and little things like: did you label your graphs and include figure captions. **The homework is worth 20 points. 10 points will be given for the overall layout, correctness and neatness of the report, and 10 additional points will be for specific technical things that the TAs will look for in the report itself.**
- **Late submissions will not be accepted on Canvas, send them to bamdadh@uw.edu directly. Late reports are subject to a 2 points/day penalty up to five days. They are no longer accepted afterwards. For example, if your report is three days late and you managed to get $16/20$, your final grade will be $16 - 6 = 10$.** Basically, you will lose 2% of your overall course grade for each day the report is late. So be careful.

Problem Description: Qualifying Red Wine

You have been hired as a data scientist by a winery in Portugal. Your task is to develop an algorithm that predicts the quality of wine from a series of chemical measurements. Your algorithm will be used by the marketing team to price a new batch of products.

You have access to a training data set `wine_training.csv` consisting of 1115 instances (different types of wine that were measured in the lab) and a test data set `wine_test.csv` with 479 instances. Each instance of the data has 11 attributes (features) that are outlined in the description file `wine_description.txt`. The corresponding output to each set of features is the quality of the wine on a scale of 0 to 10 provided by experts. Finally, you are given the lab measurements for a batch of five new wines in `wine_new_batch.csv` for which you are required to predict the qualities.

Some comments and hints

Here are some pointers.

1. Using numpy's `loadtxt` or MATLAB's `csvread` is convenient for importing the data.
2. Your first step after reading in the data should be to normalize and center your input features (the x_j's) as well your outputs (the y_j's) so that they have mean 0 and standard deviation 1. This has a notable impact on the performance of kernel methods.
3. Kernel regression is very sensitive to the choice of regularization parameter λ and length scale of the kernel σ. This might require a lot of trial and error to fine tune the parameters in Task 3, which in turn can take a lot of computational time. First prototype your code on a subset of the data and find a ball-park value for the parameters before running full simulations.
4. Task 3 can be very computationally demanding if you choose to do 2D cross-validation (CV) as in Lectures 16 and 17. This will limit the set of values of σ and λ that you can consider. A good approach is to start with a few values over a wide range and refine your search successively.

Tasks

Below is a list of tasks to complete in this homework and discuss in your report.

1. Use linear regression (least squares) to fit a linear model to the training set.
2. Use kernel ridge regression to fit a nonlinear model to the training set using the Gaussian (RBF) kernel as well as the Laplacian kernel

$$k_{\text{rbf}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\right), \quad k_{\text{lap}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma}\right), \quad \text{for } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^{11}.$$

3. Use 10-fold CV to tune the length scale σ and the regularization parameter λ for each of the above kernels. Report your choices of the optimal values of σ, λ and provide a clear explanation of why and how you picked those values. Keep in mind that you won't be able to report the "true" optimal values here. We are looking for an informed/good choice given your computational budget.
4. Provide a table reporting the training and test mean squared errors (MSEs) of all three models: linear regression, and Gaussian and Laplacian kernels with the optimal hyperparameters found via CV. Discuss your findings.
5. Use your three models to predict the quality of the new batch of wines and report the output of each model on the 0-10 scale.