MTH 4920/5050 Homework 1

Linear Regression and Linear Basis Expansions (LBF)

Deadline: Wednesday, January 27

Instructions

Submit **one** Python notesbook or PDF file for grading.

- Include text explanations of your process, well-commented code, and written or typed work for the optimization problem.
- If you submit a notebook file, **run the code before submitting** so outputs are visible.
- If you submit a PDF, **embed screenshots of your code and the outputs** (or copy/paste) so outputs are visible and submit code files separately.
- Any language I can easily run is acceptable, but I highly recommend Python due to the built-in functions and compatibility with my code (so you can reuse/modify it).

Problems

1. This problem uses the daily temperatures dataset for Melbourne, AUS. You will try to predict the temperatures based on the day with regression models.

   (a) Fit the least squares line to training data to predict the temperatures from the date and evaluate its fit on a testing data. Plot the line with the data.

   (b) Construct a appropriate LBF expansion ($M \geq 5$) for use with this problem. Fit the least squares LBF model to training data and evaluate its fit on a testing set. Plot the curve with the data.
   [*Hint.* Use properties in the data to choose a reasonable basis.] [15 points]

2. This problem uses the Combined Cycle Power Plant Data Set containing data collected from a power plant and its net hourly electrical energy output.

   (a) Fit the least squares hyperplane to the training data to predict the energy output (EP) of the plant using the other data columns and evaluate its fit on the test set.

   (b) Fit the least squares quadratic polynomial to the training set, and evaluate its fit on the testing set.
   [*Hint.* Use a linear basis for quadratic polynomials in X_1, X_2, X_3, and X_4.] [10 points]

3. Suppose we use a linear basis expansion h_0, ..., h_M to create a linear model with parameters $\theta = (\theta_0, ..., \theta_M)$

$$f(x_i) = \sum_{m=0}^{M} \theta_m h_m(x_i)$$

   Find an exact matrix expression for θ minimizing the ridge regression loss function,

$$\min_{\theta} L(\theta) = \min_{\theta} \left(\sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \sum_{m=0}^{M} \theta_m^2 \right),$$

   where $\lambda \in \mathbb{R}$, $X \in \mathbb{R}^{n \times d}$, and $y \in \mathbb{R}^n$ are constant.
   [*Hint.* Write L in matrix notation first.] [10 points]