Homework 8, Kernels

36-462/662, Spring 2022

Due at 6 pm on Thursday, 24 March 2022

Agenda: Practice in theory with going back and forth between the "primal" view (coefficients on features) and the "dual" view (weights on training points); real-data practice with kernel ridge regression.

General hint for theory problems: *The Matrix Cookbook* (Petersen and Pedersen 2012) is an invaluable reference for all the bits of matrix algebra and calculus one may have momentarily forgotten.

1. (10) **Online reading questions** are on Canvas, and look ahead to what we will do next week. (But studying those readings will also help with this assignment.)

2. **Bounded coefficient vectors versus bound weights on points** We saw in class, and in the notes, how kernel machines take the form
$$s(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) \tag{1}$$
and this is equivalent to explicitly using the features,
$$s(x) = \sum_{j=1}^{d} \beta_j \phi_j(x) \tag{2}$$
In what follows, we show how $\|\beta\|$ and $\|\alpha\|$ are related, even though those are vectors of different dimensions (d and n, respectively). This is convenient when, for example, we want to apply the methods which penalized or constraint coefficients to kernel machines.

   a. (6) Show that
$$\beta_j = \sum_{i=1}^{n} \alpha_i \phi_j(x_i) \tag{3}$$

   b. (6) Show that
$$\|\beta\|^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} \sum_{j=1}^{d} \phi_j(x_i)\phi_j(x_{i'}) \tag{4}$$

   c. (6) Show that
$$\|\beta\|^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} K(x_i, x_{i'}) \tag{5}$$

   d. (6) Show that
$$\|\beta\|^2 = \alpha \cdot \mathbf{K}\alpha \tag{6}$$

3. **Kernel least squares and kernel ridge regression** Our data consist of pairs $(t_1, y_1), (t_2, y_2), \ldots (t_n, y_n)$, and we have chosen a kernel function K. Define $\mathbf{y}$ as the $[n \times 1]$ matrix of the y_is, and $\mathbf{K}$ as the $[n \times n]$ matrix where $K_{ij} = K(t_i, t_j)$. α will be our $[n \times 1]$ matrix of weights for the training points.

a. (5) Explain why the unregularized kernel least-squares regression problem is

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n}(\mathbf{y} - \mathbf{K}\alpha)^T(\mathbf{y} - \mathbf{K}\alpha) \tag{7}$$

Hint: HW4 Q2b.

b. (4) Show that the solution to that optimization problem is

$$\hat{\alpha} = \mathbf{K}^{-1}\mathbf{y} \tag{8}$$

c. (5) Explain why the kernel ridge regression (KRR) problem is

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n}(\mathbf{y} - \mathbf{K}\alpha)^T(\mathbf{y} - \mathbf{K}\alpha) + \lambda \alpha^T \mathbf{K}\alpha \tag{9}$$

Hint: Conclusion of Q1.

d. (5) Show that the gradient of the KRR objective function is

$$-\frac{2}{n}\mathbf{K}\mathbf{y} + \frac{2}{n}\mathbf{K}^2\alpha + 2\lambda\mathbf{K}\alpha \tag{10}$$

Hint: $\mathbf{K}$ is symmetric (why?).

e. (5) Show that one solution to the first order condition for the KRR problem is

$$\hat{\alpha} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{y} \tag{11}$$

What are the dimensions of $\mathbf{I}$ here?

f. (5) Show that the fitted values on the training set are given by

$$\hat{m} = \mathbf{K}(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{y} \tag{12}$$

What does this tell you about the fit when $\lambda = 0$?

g. (1) Are there any solutions to the first-order condition *other than* the one you found in Q3e?

4. **Fitting an anomaly** The data file `anomaly.csv`, on the class website, contains a predictor variable (`t`) and a response (`anomaly`). Use the odd-numbered rows as training points, and the even-numbered rows as testing points. (Ordinarily we'd divide randomly but I have my reasons, which will become clear later.) *Hint*: if you're using the `kernlab` package, you will find the example of kernel ridge regression in the notes helpful. (You do not have to use `kernlab`, but make sure any other package you do use can do everything called for.)

   a. (1) Plot the data, using different colors or symbols to distinguish between the training and testing points.

   b. (2) Using a Gaussian kernel with $\sigma = 0.01$, build the kernel matrix $\mathbf{K}$, and include a picture of the matrix, with rows on one axis, columns on the other axis, and either color or a third axis showing the values of the matrix. Describe the shape of this image in words. What does this picture tell us about which training points are seen as similar by the kernel?

   c. (5) Fix $\lambda = 1$ and add both fitted values (for the training set) and predicted values (for the testing set) to your plot from Q4a. For full credit, connect them into a single continuous line or curve.

   d. (5) For 11 different λ values between 10^{-6} and 100, find the fitted and predicted values, as in Q4c, and add them to that plot. You can present your results either as an array of 12 (small) plots, all on one page, or use color (or some other device) to have all 12 curves in a single plot, if you can make that legible. Describe, in words, how the curves change as λ gets bigger or smaller. *Hint*: You will probably *not* want to have these all evenly spaced.

e. (5) Use leave-one-out cross-validation to estimate the MSE for the different λs you considered in Q4d. Make a plot showing RMSE (not MSE) as a function of λ, estimating RMSE both with the testing set and LOOCV. How well do the two methods agree about where to put the best value of λ? Are either (or both) of those estimates reasonable-looking, considering the figure(s) in Q4d? Hint: The short-cut formula of lecture 10, and Q3f.

f. (4) For 11 different values of σ between 10^{-6} and 100, calculate the kernel matrices and plot the fitted and predicted values, using the λ you selected in Q4e. (If you skipped Q4e, use $\lambda = 1$.) As in Q4d, present the results either as an array of 12 small plots on a single page, or one plot with 12 distinct lines. Describe, in words, what happens to the curves as σ gets larger or smaller.

g. (3) For each σ you considered in Q4f, find the best λ, among those you considered in Q4d, using the same technique as in Q4e. Report the best combination of σ and λ, and the plot of the data, the fitted values and the predicted values at that combination.

5. (1) **Timing** How long, roughly, did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All plots and tables are generated by code included in the R Markdown file. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. **Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.**

Extra credit (5): In Q4g, you jointly optimized σ and λ for kernel ridge regression, by evaluating all possible combinations in a grid of values ("grid search"). Write a function, using `optim()` (or similar) which will jointly optimize σ and λ, using leave-one-out cross-validation. What's the optimal combination of σ and λ? Show the resulting curve, together with the data. For this problem, show your code.

References

Petersen, Kaare Brandt, and Michael Syskind Pedersen. 2012. "The Matrix Cookbook." Technical University of Denmark, Intelligent Signal Processing Group. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274.