HW #1 Stat 602 Sp 19

Do problems 1.6, 1.7, 2.2, 2.3 (and 2.4), 2.5, 3.3, 3.4, 3.5, 3.11, 3.12, and 5.7 of the *Exercises for Statistical Learning* document.

HW #2 Stat 602 Sp 19

1. Problems 4.3, 4.4, 5.1 of the *Exercises for Statistical Learning* document.

2. For the dataset of Problem 3.4 (of the *Exercises for Statistical Learning* document) make up a matrix of inputs based on x consisting of the values of Haar basis functions up through order $m=3$. This will produce a 100×16 matrix $\mathbf{X_h}$.

a) Find $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ and plot the corresponding $\hat{y}$ as a function of x with the data also plotted in scatterplot form.

b) Center y and standardize the columns of $\mathbf{X_h}$. Find the lasso coefficient vectors $\hat{\boldsymbol{\beta}}$ with exactly $M=2,4,$ and 8 non-zero entries with the largest possible $\sum_{j=1}^{16}\left|\hat{\beta}_j^{\text{lasso}}\right|$ (for the counts of non-zero entries). Plot the corresponding $\hat{y}$s as functions of x on the same set of axes, with the data also plotted in scatterplot form.

3. For the dataset of Problem 3.4 (of the *Exercises for Statistical Learning* document) make up a 100×7 matrix $\mathbf{X_h}$ of inputs based on x consisting of the values of functions on slide 7 of Section 4.2 for the 7 knot values

$$\xi_1=0,\xi_2=.1,\xi_3=.3,\xi_4=.5,\xi_5=.7,\xi_6=.9,\xi_7=1.0$$

Find $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ and plot the corresponding natural cubic regression spline, with the data also plotted in scatterplot form.

4. Consider the function optimization problem solved by the cubic smoothing splines over the interval $[0,1]$. Suppose that $N=11$ training data pairs (x_i,y_i) have $x_1=0,x_2=.1,x_3=.2,\ldots,x_{11}=1.0$ and that the basis functions on slide 7 of Section 4.2 are employed.

a) Find the matrices $\mathbf{H},\boldsymbol{\Omega},$ and $\mathbf{K}$ associated with the smoothing spline development of Section 5.1 for a problem where $\mathbf{Y}'=(0,1.5,2,.5,0,-.5,0,1.5,3.5,4.5,3.5)$. Using the result of Problem 9.3 of the *Exercises for Statistical Learning* document, you can compute $\boldsymbol{\Omega}$ in closed form. Do so and give the numerical versions of all 3 of these matrices.

b) Do an eigen analysis of the matrix $\mathbf{K}$ and then plot as functions of row index i the coordinates of the eigenvectors of $\mathbf{K}$. Put these plots on the same set of axes, or stack plots of them one above the other in descending order of the size of eigenvalue. (Connect consecutive plotted points for a given eigenvector with line segments and use different plotting symbols, colors, and/or line weights and types so that you can make qualitative comparisons of the nature of these.) If possible, compare the "shapes" of the plots for the two largest and two smallest corresponding eigenvalues. Qualitatively speaking, what kinds of components of a vector of observed values, **Y** get "most suppressed" in the smoothing?

c) Plot effective degrees of freedom for the spline smoother as a function of the parameter λ. Do simple numerical searches to identify values of λ corresponding to degrees of freedom 2.5,3,4, and 5.

5. In a $p=1$ smoothing context like that in Problem 4 above, where $N=11$ training data pairs (x_i, y_i) have $x_1=0, x_2=.1, x_3=.2,\ldots,x_{11}=1.0$, consider locally weighted regression as on slides 5-7 of Section 5.1 based on a Gaussian kernel.

a) Compute and plot effective degrees of freedom as a function of the bandwidth, λ. (It may be most effective to make the plot with λ on a log scale or some such.) Do simple numerical searches to identify values of λ corresponding to effective degrees of freedom 2.5,3,4, and 5.

b) Compare the smoothing matrix "$\mathbf{S}_\lambda$" for Problem 3 producing 4 effective degrees of freedom to the matrix "$\mathbf{L}_\lambda$" in the present context also producing 4 effective degrees of freedom. What is the 11×11 matrix difference? Plot, as a function of column index, the values in the 1st, 3rd, and 5th rows of the two matrices, connecting with line segments successive values from a given row. (Connect consecutive plotted points for a given row of a given matrix and use different plotting symbols, colors, and/or line weights and types so that you can make qualitative comparisons of the nature of these.)

6. Again use the data set of Problem **4**.

a) Fit with approximately 5 and then 9 effective degrees of freedom

i) a cubic smoothing spline (using smooth.spline()) , and

ii) a locally weighted linear regression smoother based on a tricube kernel (using loess(…,span= ,degree=1)).

Plot for approximately 5 effective degrees of freedom all of y_i and the 2 sets of smoothed values against x_i. Connect the consecutive $(x_i, \hat{y}_i)$ for each fit with line segments so that they plot as "functions." Then redo the plotting for 9 effective degrees of freedom.

7. Consider the White Wines Dataset[1] from the UCI Machine Learning Data Repository (http://archive.ics.uci.edu/ml/datasets/Wine+Quality) on Canvas. Consider SEL prediction of

[1] The White Wines Dataset is not absolutely ideal as an example in that the response variable can take only integer values 1 through 10 and is probably not really an interval-level variable in the

what can be learned about wine "quality" from the 11 input variables. There are roughly 5000 cases in this dataset and it is about at the (size) limit of what is conveniently handled using R and an ordinary laptop. (Other faster software like Python or Matlab and/or implementation on a server or cluster may be required for bigger datasets with many machine learning applications.)

Using caret find sets of "best" (according to repeated 10-fold cross-validation) predictions for the quality ratings for

i) kNN prediction
ii) elastic net prediction
iii) PCR prediction
iv) PLS prediction
v) MARS prediction (implemented in earth)

Make a scatterplot matrix for all these plus the y values and OLS predictions. Compute a correlation matrix for these 7 sets of values and display this rounded to 2 decimal places.

8. Use again the data of Problem **4.** Center the y values and standardize x. (We will abuse notation and use x and z to stand for standardized versions of input values.)

This question will make use of the kernel function $\mathcal{K}(x,z)=\exp\left(-\frac{(x-z)^2}{2}\right)$ and the mapping $T(x)(\cdot)=\mathcal{K}(x,\cdot)$ that associates with input value $x\in\Re$ the function $\mathcal{K}(x,\cdot):\Re\to\Re$ (an abstract "feature").

In the (very high-dimensional) space of functions mapping $\Re\to\Re$, the $N=11$ training set generates an 11-D subspace of functions consisting of all linear combinations of the $T(x_i)$. Two possible inner products in that subspace are the "L_2" inner product

$$\langle g,h\rangle_{L_2}=\int g(x)h(x)dx$$

and the inner product defined for functions in the range of $T(\cdot)$ by

$$\langle T(x)(\cdot),T(z)(\cdot)\rangle_{\mathcal{A}}=\mathcal{K}(x,z)$$

a) Apply the first 3 steps of the Gram-Schmidt process to the abstract features of the training data (considered in the order given in the data table) to identify 3 orthonormal functions $\Re^2\to\Re$ that are linear combinations of $T(x_1),T(x_2),T(x_3)$. Do this first using the L_2 inner

first place (being more ordinal in nature). For purposes of exercise we will ignore these matters, and treat the quality rating as a measured numerical response with SEL.

product, and then using the kernel-based inner product. Are the two sets of 3 functions the same?

Note that both the functions $Z(x) \equiv 0$ and $M = \frac{1}{11}\sum_{i=1}^{11} T(x_i)$ are linear combinations of (in the subspace of functions generated by) the $T(x_i)$. It probably makes sense to "center" the abstract features generated by the training set, replacing each $T(x_i)$ with

$$S(x_i) = T(x_i) - M$$

b) Compute the "centered Gram matrix" for "kernel PCA"

$$\mathbf{G} = \left(\left\langle S(x_i), S(x_j) \right\rangle_{\mathcal{A}} \right)_{\substack{i=1,2,\ldots,11 \\ j=1,2,\ldots,11}}$$

c) Do an eigen analysis for the matrix $\mathbf{G}$. (For Euclidean features, this matrix would be a multiple of a sample covariance matrix.) The eigenvectors of this matrix can be thought of as giving coefficients defining 11 linear combinations of the functions $S(x_i)$ that are orthogonal and span the same space as the centered abstract features and are of decreasing importance as index increases. Plot these 11 functions. If there are any kind of trends obvious in the nature of these plots with decreasing eigenvalue, describe them.

d) Find the projection of a function $S(.65)$ onto the span of $\{T(x_i)(\cdot)\}_{i=1,2,\ldots,11}$ in $\mathcal{A}$ and plot this function and its projection on the same axes.