Exercises in Modern Multivariate Statistical Learning

Version .544

2/20/2019

Contents

Section 1: Problems Concerning "The Curse of Dimensionality"

These are exercises intended to provide intuition that data in $\Re^p$ are necessarily "sparse." The realities are that $\Re^p$ is "huge" and for p at all large, "filling up" even a small part of it with data points is effectively impossible, and our intuition about distributions in $\Re^p$ is very poor.

1.1. (6HW-11) Let $F_p(t)$ and $f_p(t)$ be respectively the χ^2_p cdf and pdf. Consider the $\mathrm{MVN}_p(\mathbf{0},\mathbf{I})$ distribution and $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ iid with this distribution. With

$$M = \min\left\{\|\mathbf{Z}_i\| \,|\, i = 1,2,\ldots,N\right\}$$

write out a one-dimensional integral involving $F_p(t)$ and $f_p(t)$ giving $\mathrm{E}M$. Evaluate this mean for $N = 100$ and $p = 1, 5, 10,$ and 20 either numerically or using simulation.

1.2. (6HW-13) For each of $p = 1, 5, 10,$ and 20 generate at least 1000 realizations of pairs of points $\mathbf{x}$ and $\mathbf{z}$ as iid uniform over the p- dimensional unit ball (the set of $\mathbf{x}$ with $\|\mathbf{x}\| \le 1$). Compute (for each p) the sample average distance between $\mathbf{x}$ and $\mathbf{z}$. (For $\mathbf{Z} \sim \mathrm{MVN}_p(\mathbf{0},\mathbf{I})$ independent of $U \sim \mathrm{U}(0,1)$, $\mathbf{x} = \left(U^{1/p} / \|\mathbf{Z}\|\right)\mathbf{Z}$ is uniformly distributed in the unit ball in $\Re^p$.)

1.3. (5HW-14) For each of $p = 10, 20, 50, 100, 500, 1000$ make $n = 10{,}000$ draws of distances between pairs of independent points uniform in the cube $[0,1]^p$. Use these to make 95% confidence limits for the ratio

$$\frac{\textit{mean distance between two random points in the cube}}{\textit{maximum distance between two points in the cube}}$$

1.4. (5HW-14) For each of $p = 10, 20, 50$ make $n = 10{,}000$ random draws of $N = 100$ independent points uniform in the cube $[0,1]^p$. Find for each sample of 100 points, the distance from the first point drawn to the 5th closest point of the other 99. Use these to make 95% confidence limits for the ratio

$$\frac{\textit{mean diameter of a 5-nearest neighbor neighborhood if } N = 100}{\textit{maximum distance between two points in the cube}}$$

1.5. (5HW-14) What fraction of random draws uniform from the unit cube $[0,1]^p$ lie in the "middle part" of the cube $[\varepsilon, 1-\varepsilon]^p$, for a small positive number ε ?

The next 3 problems are based on nice ideas taken from Giraud's book.

1.6. (6HW-15) For $p = 2, 10, 100,$ and 1000 draw samples of size $n = 100$ from the uniform distributions on $[0,1]^p$. Then for every $(\mathbf{x}_i, \mathbf{x}_j)$ pair with $i < j$ in one of these samples, compute the Euclidean distance between the two points, $\|\mathbf{x}_i - \mathbf{x}_j\|_2$. Make a histogram (one p at a time) of these $\binom{100}{2}$ distances. What do these suggest about how well "local" prediction methods (that rely only on data points $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i$ "near" $\mathbf{x}$ to make predictions about y at $\mathbf{x}$) can be expected to work?

1.7. (6HW-15) Consider finding a lower bound on the number of points $\mathbf{x}_i$ (for $i = 1, 2, \ldots, n$) required to "fill up" $[0,1]^p$ in the sense that no point of $[0,1]^p$ is Euclidean distance more than ε away from some $\mathbf{x}_i$.

The p-dimensional volume of a ball of radius r in $\Re^p$ is

$$V_p(r) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} r^p$$

and Giraud notes that it can be shown that as $p \to \infty$

$$\frac{V_p(r)}{\left(\frac{2\pi e r^2}{p}\right)^{p/2} (p\pi)^{-1/2}} \to 1$$

Then, if n points can be found with ε-balls covering the unit cube in $\Re^p$, the total volume of those balls must be at least 1. That is

$$nV_p(\varepsilon) \geq 1$$

What then are approximate lower bounds on the number of points required to fill up $[0,1]^p$ to within ε for $p = 20, 50,$ and 200 and $\varepsilon = 1, .1,$ and $.01$? (Giraud notes that the $p = 200$ and $\varepsilon = 1$ lower bound is larger than the estimated number of particles in the universe.)

1.8. (6HW-15) Giraud points out that for large p, most of $\text{MVN}_p(\mathbf{0}, \mathbf{I})$ probability is "in the tails." For $f_p(\mathbf{x})$ the $\text{MVN}_p(\mathbf{0}, \mathbf{I})$ pdf and $0 < \delta < 1$ let

$$B_p(\delta) = \left\{\mathbf{x} \mid f_p(\mathbf{x}) \geq \delta f_p(\mathbf{0})\right\} = \left\{\mathbf{x} \mid \|\mathbf{x}\|^2 \leq 2\ln(\delta^{-1})\right\}$$

be the "central"/"large density" part of the multivariate standard normal distribution.

a) Using the Markov inequality, show that the probability assigned by the multivariate standard normal distribution to the region $B_p(\delta)$ is no more than $1/\delta 2^{p/2}$.

b) What then is a lower bound on the radius (call it $r(p)$) of a ball at the origin required so that the multivariate standard normal distribution places probability .5 in that ball? What is an upper bound on the ratio $f_p(\mathbf{x})/f_p(\mathbf{0})$ outside the ball with radius that lower bound? Plot these bounds as functions of p for $p \in [1,500]$.

Section 2: Problems Concerning Function Optimization and Theoretically Optimal Choices of Predictors (Based on a Complete Probability Model)

2.1. (6E1-17) Consider the 2-class classification model with the coding $y \in \{-1,1\}$ and (for sake of concreteness) $x \in \Re^1$. As is more or less standard, for $g(x)$ a generic voting function we'll consider the classifier

$$f(x) = \text{sign}(g(x))$$

Another (besides those mentioned in class) "function loss" sometimes discussed is

$$h(v) = (v-1)^2$$

a) Carefully derive the function $g^{\text{opt}}(x)$ optimizing $\text{E}h(yg(x))$ over choices of g.

b) To the extent possible, simplify a good upper bound on the 0-1 loss error rate of a classifier $f(x)$ made from your $g^{\text{opt}}(x)$ from part **a)**.

c) Suppose that in pursuit of a good classifier, one wishes to optimize an empirical version of $\text{E}h(yg(x))$, based on a training set of size N, over the class of functions of the form

$$g(x \mid \beta_0, \beta_1) = 2\Phi(\beta_0 + \beta_1 x) - 1 \quad ,$$

penalized by $\lambda\beta_1^2$ for a $\lambda \geq 0$. (Φ is the standard normal cdf.) In as simple a form as possible, give two equations to be solved simultaneously to do this fitting.

d) Suppose that as a matter of fact the two class-conditional densities operating in the model are

$$p(x \mid -1) = I[0 < x < 1] \text{ and } p(x \mid 1) = 6x(1-x)I[0 < x < 1]$$

and that ultimately what is desired is a good ordering function $\mathcal{O}(x)$, one that produces a small value of the "AUC" criterion. Do you expect the methodology of part **c)** to produce a function $g(x \mid \hat{\beta}_0, \hat{\beta}_1)$ that would be a good choice of $\mathcal{O}(x)$? Explain carefully.

2.2. (6HW-17, 5HW-18)

a) Argue carefully that for inherently non-negative response y with loss

$$L(y, \hat{y}) = \left[\ln\left(\frac{\hat{y}+1}{y+1}\right)\right]^2$$

a theoretically optimal predictor is

$$f(\mathbf{x}) = \exp\left(\mathrm{E}\left[\ln(y+1) \mid \mathbf{x}\right]\right) - 1$$

b) The Zillow Kaggle game for predicting (positive) house prices used the loss function

$$L(\hat{y}, y) = (\ln \hat{y} - \ln y)^2 = \left(\ln \frac{\hat{y}}{y}\right)^2$$

Identify the function of $\mathbf{x}$, call it $f(\mathbf{x})$, that based on a joint distribution P for $(\mathbf{x}, y)$ optimizes

$$\mathrm{E}\, L(g(\mathbf{x}), y)$$

over choices of function $g(\mathbf{x})$.

2.3. (6HW-17) Argue carefully that losses $h_1, h_2,$ and h_3 (negative Bernoulli loglikelihood term, exponential, and hinge losses) have optimizers of

$$\mathrm{E} h(y g(\mathbf{x}))$$

(functions $g^{\text{opt}}(\cdot)$) as indicated in the typed outline.

2.4. (6HW-13) Consider the loss function $L(y, \hat{y}) = (1 - y\hat{y})_+$ for y taking values in $\{-1, 1\}$ and prediction $\hat{y}$. Suppose that $P[y=1] = p$. Write out the expected (over the randomness in y) loss of prediction $\hat{y}$. Plot this as a function of $\hat{y}$ for the cases where first $p < .5$ and then $p > .5$. (These are continuous functions that are linear on the intervals $(-\infty, -1), (-1, 1),$ and $(1, \infty)$.) What is an optimal choice of $\hat{y}$ (depending upon p)?

2.5. (5HW-16) Consider a 0-1 loss $K = 2$ classification problem with $p = 1$, $\pi_0 = \pi_1 = \frac{1}{2}$, and pdfs

$$g(x \mid 0) = I[-.5 < x < .5] \text{ and } g(x \mid 1) = 12x^2 I[-.5 < x < .5]$$

a) What is the optimal classification rule in this problem? (In the notation of the slides, this is $f(x)$). What is the minimum expected loss?

b) If one were to do "feature engineering" here, adding some function of x, say $t(x)$, to make a vectors of features $(x, t(x))$ for classification purposes, hoping to eventually employ a good "linear classifier"

$$\hat{f}\left(x,t\left(x\right)\right)=I\left[a+bx+ct\left(x\right)>0\right]$$

for appropriate constants $a,b,$ and c, what (knowing the answer to **a)**) would be a good choice of $t\left(x\right)$? (Of course, one doesn't know the answer to **a)** when doing feature selection!)

2.6. (5E1-15) Consider two probability densities on the unit disk in $\Re^2$ (i.e. on $\left\{\left(x_1,x_2\right)\mid x_1^2+x_2^2\le 1\right\}$),

$$g_1\left(x_1,x_2\right)=\frac{1}{\pi} \text{ and } g_2\left(x_1,x_2\right)=\frac{3}{2\pi}\sqrt{1-\left(x_1^2+x_2^2\right)}$$

and a 2-class 0-1 loss classification problem with prior probabilities $\pi_1=\pi_2=.5$.

a) Give a formula for a best-possible single feature $T\left(x_1,x_2\right)$.

b) Give an explicit form for the theoretically optimal classifier in this problem.

c) Suppose that one uses features $x_1,x_2,x_1^2,x_2^2,$ and x_1x_2 to do 2-class classification based on a moderate number of iid training cases from this model. Would you expect better classification performance for 1) a classifier based on *logistic regression* using these features or 2) *a classification tree* using these features? Explain.

2.7. (5E1-18) Consider a $K=3$ classification model with $p=3$ class-conditional densities on $\left[0,1\right]^3$

$$g_1\left(x_1,x_2,x_3\right)=2x_1, g_2\left(x_1,x_2,x_3\right)=2x_2, \text{ and } g_3\left(x_1,x_2,x_3\right)=2x_3$$

a) Identify two real-valued features $T_1\left(\mathbf{x}\right)$ and $T_2\left(\mathbf{x}\right)$ that are complete summarizations of all information about the class label $y\in\left\{1,2,3\right\}$ provided by $\mathbf{x}=\left(x_1,x_2,x_3\right)$.

b) For the case of $\pi_1=\pi_2=\pi_3=\frac{1}{3}$ give the form of an optimal 0-1 loss classifier in terms of the values t_1 and t_2 of $T_1\left(\mathbf{x}\right)$ and $T_2\left(\mathbf{x}\right)$.

c) For the case of $\pi_1=.6,\pi_2=.4,$ and $\pi_3=0$ where $L\left(\hat{y},1\right)=10I\left[\hat{y}\ne 1\right]$ and otherwise $L\left(\hat{y},y\right)=I\left[\hat{y}\ne y\right]$, give explicitly the form of an optimal classifier in terms of the value of $\mathbf{x}=\left(x_1,x_2,x_3\right)$.

Section 3: Problems Concerning Decomposition and Control of Predictor Test Error

3.1. (6HW-11) Consider SEL prediction. Suppose that in a very simple problem with $p=1$, the distribution P for the random pair (x,y) is specified by

$$x \sim \mathrm{U}(0,1) \text{ and } y \mid x \sim \mathrm{N}\left(x^2,(1+x)\right)$$

($(1+x)$ is the conditional variance of the output). Further, consider two possible sets of functions $S=\{g\}$ for use in creating predictors of y, namely

1. $S_1=\{g \mid g(x)=a+bx \text{ for real numbers } a,b\}$, and
2. $S_2=\left\{g \mid g(x)=\sum_{j=1}^{10} a_j I\left[\frac{j-1}{10}<x\leq\frac{j}{10}\right] \text{ for real numbers } a_j\right\}$

Training data are N pairs (x_i,y_i) iid P. Suppose that the fitting of elements of these sets is done by

1. OLS (simple linear regression) in the case of S_1, and
2. according to

$$\hat{a}_j=\begin{cases}\bar{y} & \text{if no } x_i\in\left(\frac{j-1}{10},\frac{j}{10}\right] \\ \dfrac{1}{\#x_i\in\left(\frac{j-1}{10},\frac{j}{10}\right]}\displaystyle\sum_{\substack{i \text{ with} \\ x_i\in\left(\frac{j-1}{10},\frac{j}{10}\right]}} y_i & \text{otherwise}\end{cases}$$

   in the case of S_2

to produce predictors $\hat{f}_1$ and $\hat{f}_2$.

a) Find (analytically) the functions g^* for the two cases. Use them to find the two expected squared model biases $\mathrm{E}^x\left(\mathrm{E}[y\mid x]-g^*(x)\right)^2$. How do these compare for the two cases?

b) For the second case, find an analytical form for $\mathrm{E}^{\mathrm{T}}\hat{f}_2$ and then for the average squared estimation bias $\mathrm{E}^x\left(\mathrm{E}^{\mathrm{T}}\hat{f}_2(x)-g_2^*(x)\right)^2$. (Hints: What is the conditional distribution of the y_i given that no $x_i\in\left(\frac{j-1}{10},\frac{j}{10}\right]$? What is the conditional mean of y given that $x\in\left(\frac{j-1}{10},\frac{j}{10}\right]$?)

c) For the first case, simulate at least 1000 training datasets of size $N = 100$ and do OLS on each one to get corresponding $\hat{f}$'s. Average those to get an approximation for $\mathrm{E}^{\mathbf{T}} \hat{f}_1$. (If you can do this analytically, so much the better!) Use this approximation and analytical calculation to find the average squared estimation bias $\mathrm{E}^{x}\left(\mathrm{E}^{\mathbf{T}} \hat{f}_1(x) - g_1^*(x)\right)^2$ for this case.

d) How do your answers for **b)** and **c)** compare for a training set of size $N = 100$?

e) Use whatever combination of analytical calculation, numerical analysis, and simulation you need to use (at every turn preferring analytics to numerics to simulation) to find the expected prediction variances $\mathrm{E}^{x}\mathrm{Var}^{\mathbf{T}}\left(\hat{f}(x)\right)$ for the two cases for training set size $N = 100$.

f) In sum, which of the two predictors here has the best value of Err for $N = 100$?

3.2. (6HW-11) Vardeman will send out two files with respectively 100 and then 1000 pairs (x_i, y_i) generated according to P in problem **3.1**. Use 10-fold cross validation to see which of the two predictors in problem **3.1** looks most likely to be effective. (The datasets are not sorted, so you may treat successively numbered groups of $1/10$ th of the training cases as your $K = 10$ randomly created pieces of the training set.)

3.3. (5HW-14) Again consider SEL prediction. Suppose that (unknown to a statistician) a mechanism generates iid data pairs (x, y) according to the following model:

$$x \sim \mathrm{U}(-\pi, \pi)$$
$$y \mid x \sim \mathrm{N}\left(\sin(x), .25\left(|x|+1\right)^2\right)$$

(The conditional *variance* is $.25\left(|x|+1\right)^2$.)

a) What is an absolutely minimum value of Err possible regardless what training set size, N, is available and what fitting method is employed?

b) What linear function of x (which $g(x) = a + bx$) has the smallest "average squared bias" as a predictor for y? What cubic function of x (which $g(x) = a + bx + cx^2 + dx^3$) has the smallest average squared bias as a predictor for y? Is the set of cubic functions big enough to eliminate model bias in this problem?

3.4. (5HW-14) Vardeman will send out an $N=100$ dataset generated by the model of problem **3.3**. Use ten-fold cross validation (use the 1st ten points as the first test set, the 2nd 10 points as the second, etc.) based on the dataset to choose among the following methods of prediction for this scenario:

- polynomial regressions of orders 0,1,2,3,4, and 5
- regressions using sets of predictors $\{1,\sin x,\cos x\}$ and $\{1,\sin x,\cos x,\sin 2x,\cos 2x\}$
- a regression with the set of predictors $\{1,x,x^2,x^3,x^4,x^5,\sin x,\cos x,\sin 2x,\cos 2x\}$

(Use ordinary least squares fitting.) Which predictor looks best on an empirical basis? Knowing how the data were generated (an unrealistic luxury) which methods here are without model bias?

3.5. (5HW-16) As in problem **2.5**, consider a 0-1 loss $K=2$ classification problem with $p=1$, $\pi_0=\pi_1=\frac{1}{2}$, and pdfs

$$g(x\,|\,0)=I[-.5<x<.5] \text{ and } g(x\,|\,1)=12x^2 I[-.5<x<.5]$$

a) What is the "minimum expected loss" part of Err in this problem?

b) Identify the best rule of the form $g_c(x)=I[x>c]$. (In the notation of the slides, this is $g^*(x)$ for $S=\{g_c\}$. This could be thought of as the 1-d version of a "best linear classification rule" here ... where linear classification is not so smart.) What is the "modeling penalty" part of Err in this situation?

c) Suggest a way that you might try to choose a classification rule g_c based on a very large training sample of size N. Notice that a large training set would allow you to estimate cumulative conditional probabilities $G(c\,|\,y)=P[x\le c\,|\,y]$ by relative frequencies

$$\frac{\#\text{ training cases with } x_i\le c \text{ and } y_i=y}{\#\text{ training cases with } y_i=y}$$

3.6. (5E1-14) Consider a joint pdf (for $(x,y)\in(0,1)\times(0,\infty)$) of the form

$$g(x,y)=\frac{1}{x^2}\exp\left(-\frac{y}{x^2}\right) \text{ for } 0<x<1 \text{ and } 0<y$$

($x\sim \mathrm{U}(0,1)$ and conditional on x, the variable y is exponential with mean x^2.)

a) Find the linear function of x (say $\alpha+\beta x$) that minimizes $\mathrm{E}\left(y-(\alpha+\beta x)\right)^2$. (The averaging is over the joint distribution of (x,y). Find the optimizing intercept and slope.)

b) Suppose that a training set consists of N data pairs (x_i, y_i) that are independent drawn from the distribution specified above, and that least squares is used to fit a predictor $\hat{f}_N(x) = a_N + b_N x$ to the training data. Suppose that it's possible to argue (don't try to do so) that the least squares coefficients a_N and b_N converge (in a proper probabilistic sense) to your optimizers from **a)** as $N \to \infty$. Then for large N, about what value of (SEL) training error do you expect to observe under this scenario? (Give a number.)

3.7. (5E1-16) Suppose that (unknown to statistical learners) in a $p=1$ SEL prediction problem, $x \sim \mathrm{U}(0,6)$ and $y \mid x \sim \mathrm{N}\left(x-3,(x+1)^2\right)$ (the conditional *variance* is $(x+1)^2$). A statistical learner uses a class of predictors S consisting of all functions of the form $g_{a,b}(x) = a \cdot I[x<2] + b \cdot I[x \geq 2]$.

a) In this context, what are

- the minimum expected loss possible,
- the best element of S, and
- the learner's modeling penalty?
-

b) Suppose that based on a training set of size $N = N_1 + N_2$ where N_1 is the count of x_i that are less than 2 and N_2 is the count of x_i that are at least 2, the fitting procedure used is to take[1] $\hat{a} = \bar{y}_1$ and $\hat{b} = \bar{y}_2$ (with the understanding that if $N_1 = 0$ then $\hat{a} = 0$ and if $N_2 = 0$ then $\hat{b} = 0$). Write an explicit expression for the fitting penalty here. (Hint: What is the distribution of N_1? Given that an x_i is less than 2 what are the mean and variance of y? Given that an x_i is at least 2, are the mean and variance of y?)

c) Suppose that a second statistical learner uses predictors $h_{c,d}(x) = c \cdot I[x<3] + d \cdot I[x \geq 3]$. A best such predictor is in fact $h_{-1.5,1.5}(x) = -\frac{3}{2} I[x<3] + \frac{3}{2} I[x \geq 3]$. Find a linear combination of the best element of S you identified in **a)** and this best predictor available to the second learner that is better than either individual predictor.

[1] In the obvious way, $\bar{y}_1$ is the sample mean output for inputs $x_i < 2$ and $\bar{y}_2$ is the sample mean output for inputs $x_i \geq 2$.

3.8. (6HW-13) Repeat problems **3.1** and **3.2** above supposing that the distribution P for the random pair (x,y) is specified by

$$x \sim \mathrm{U}(0,1) \text{ and } y \mid x \sim \mathrm{Exp}\left(x^2\right)$$

3.9. (6HW-15) Repeat problems **3.1** and **3.2** above supposing that the distribution P for the random pair (x,y) is specified by

$$x \sim \mathrm{U}(0,1) \text{ and } y \mid x \sim \mathrm{N}\left((3x-1.5)^2, (3x-1.5)^2 + .2\right)$$

3.10. (6E1-15) Consider a SEL prediction problem where $p=1$, and the class of functions used for prediction is the set of constant functions $S = \{h \mid h(x) = c \ \forall x \text{ and some } c \in \Re\}$. Suppose that in fact

$$x \sim \mathrm{U}(0,1),\ \mathrm{E}[y \mid x] = ax + b, \text{ and } \mathrm{Var}[y \mid x] = dx^2 \text{ for some } d > 0$$

a) Under this model, what is the best element of S, say g^*, for predicting y? Use this to find the average squared model bias in this problem.

b) Suppose that based on an iid sample of N points (x_i, y_i), fitting is done by least squares (and thus the predictor $\hat{f}(x) = \bar{y}$ is employed). What is the average squared fitting bias in this case?

c) What is the average prediction error, Err, when the predictor in **b)** is employed?

3.11. (6HW-17) Consider a toy 2-class classification model for $p=1$, where $x \mid y=0$ is $\mathrm{N}(0,1)$, $x \mid y=1$ is $\mathrm{N}\left(1,(.5)^2\right)$ (the standard deviation is $.5$), and $P[y=0] = .5 = P[y=1]$.

a) Compute and plot the function $P[y=1 \mid x]$.

b) Identify the optimal 0-1 loss classifier and the best possible expected loss/error rate in this classification problem. (This is a numerical problem.)

c) Consider the set of "linear" classifiers $S = \{I[x<c] \mid c \in \Re\} \bigcup \{I[x>c] \mid c \in \Re\}$ (that make one cut in the real numbers at c and classify one way to the left of c and the other way to the right of c). Plot as functions of c the risks

$$\mathrm{E}\left(I[y=0]I[x<c] + I[y=1]I[x>c]\right)$$

for classifiers of the form $I[x<c]$ and

$$\mathrm{E}\left(I[y=0]I[x>c]+I[y=1]I[x<c]\right)$$

for classifiers of the form $I[x>c]$. What is the best element of S (say, g^*) and then what is the "modeling penalty" associated with using the class of predictor/classifiers S (the difference between the optimal error rate and the error rate for g^*)?

d) Suppose that for a training set of size $N=100$ (generated at random from the distribution described in the preamble of this problem), one will choose a cut point $\hat{c}$ half way between two consecutive sorted x_i values minimizing

$$\min\left[\#\{y_i=0\,|\,x_i<c\}+\#\{y_i=1\,|\,x_i>c\},\#\{y_i=1\,|\,x_i<c\}+\#\{y_i=0\,|\,x_i>c\}\right]$$

Then, if

$$\#\{y_i=0\,|\,x_i<\hat{c}\}+\#\{y_i=1\,|\,x_i>\hat{c}\}\le\#\{y_i=1\,|\,x_i<\hat{c}\}+\#\{y_i=0\,|\,x_i>\hat{c}\}$$

one will employ the classifier $\hat{f}(x)=I[x<\hat{c}]$ and otherwise the classifier $\hat{f}(x)=I[x>\hat{c}]$. Simulate 10,000 training samples and find corresponding classifiers $\hat{f}$. For each $\hat{f}$ compute a (conditional on the training sample) error rate (an average of two appropriate normal probabilities on half infinite intervals bounded by $\hat{c}$) and average across the training samples. What is the "fitting penalty" for this procedure? Redo this exercise, using a training set of size $N=50$. Is the fitting penalty larger than for $N=100$?

3.12. (6HW-17) Consider the model of the previous problem, but change to the "-1 and 1" coding of classes/values of y.

a) Plot the function g minimizing $\mathrm{E}\exp(-yg(x))$ over all choices of g.

Suppose then that one wishes to approximate this minimizer with a function of the form $\beta_0+\beta_1(x-\bar{x})+\beta_2(x-\bar{x})^2$ based on a training set. Vardeman will send you a training set of size $N=100$ based on the model of this problem. Use it in what follows.

b) Use a numerical optimizing routine and identify values $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ minimizing the empirical average loss

$$R(\beta_0,\beta_1,\beta_2)=\frac{1}{N}\sum_{i=1}^{N}\exp\left(-y_i\left(\beta_0+\beta_1(x_i-\bar{x})+\beta_2(x_i-\bar{x})^2\right)\right)$$

c) Now consider the penalized fitting problem where one chooses to optimize

$$R_\lambda(\beta_0,\beta_1,\beta_2)=\frac{1}{N}\sum_{i=1}^{N}\exp\left(-y_i\left(\beta_0+\beta_1(x_i-\bar{x})+\beta_2(x_i-\bar{x})^2\right)\right)+\lambda\beta_2^2$$

For several different values of $\lambda > 0$, plot on the same set of axes, the optimizer from **a)**, the function $\beta_0 + \beta_1(x-\bar{x}) + \beta_2(x-\bar{x})^2$ optimizing $R(\beta_0,\beta_1,\beta_2)$ from **b)**, and the functions optimizing $R_\lambda(\beta_0,\beta_1,\beta_2)$.

3.13. (5E2-14) At a particular input vector of interest in a SEL prediction problem, say $\mathbf{x}$, the conditional mean of $y\,|\,\mathbf{x}$ is 3. Two different predictors, $\hat{f}_1(\mathbf{x})$ and $\hat{f}_2(\mathbf{x})$ have biases (across random selection of training sets of fixed size N) at this value of $\mathbf{x}$ that are respectively $.1$ and $-.5$. The random vector of predictors at $\mathbf{x}$ (randomness coming from training set selection) has covariance matrix

$$\mathrm{Var}\left(\begin{pmatrix}\hat{f}_1(\mathbf{x})\\ \hat{f}_2(\mathbf{x})\end{pmatrix}\right)=\begin{pmatrix}1 & .25\\ .25 & 1\end{pmatrix}$$

If one uses a linear combination of the two predictors

$$\hat{f}^{\,\text{ensemble}}(\mathbf{x}) = a\hat{f}_1(\mathbf{x}) + b\hat{f}_2(\mathbf{x})$$

there are optimal values of the constants a and b in terms of minimizing the expected (across random selection of training sets) squared difference between $\hat{f}^{\,\text{ensemble}}(\mathbf{x})$ and 3 (the conditional mean of $y\,|\,\mathbf{x}$). Write out and optimize an explicit function of a and b that (in theory) could be minimized in order to find these optimal constants.

3.14. (5E1-18) Consider a $p=1$ SEL prediction problem where

$$\mathrm{E}[y\,|\,x] = x(1-x), \mathrm{Var}[y\,|\,x] = x(1-x), \text{ and } x \sim \mathrm{U}(0,1)$$

- Find the expected loss of a theoretically optimal predictor of y, $f^{\text{opt}}(x)$.

- Consider predictors of the form $f_{\mathbf{c}}(x) = c_1 I[0 \le x < .4] + c_2 I[.4 \le x < .6] + c_3 I[.6 \le x \le 1]$ for real constants $c_1, c_2,$ and c_3. Find $\mathrm{E}[y\,|\,0 \le x < .4], \mathrm{E}[y\,|\,.4 \le x < .6],$ and $\mathrm{E}[y\,|\,.6 \le x \le 1]$ and argue that these give optimal values for the constants.

- Give an explicit expression for the expected loss of the optimal predictor of the form $f_{\mathbf{c}}(x)$. Note that together with the first answer this could give the modeling penalty here.

- Give an explicit expression for the fitting penalty if based on a training set of size N, the value c_l is estimated by $\hat{c}_l = \bar{y}_l I[\text{at least one } x_i \text{ is in the interval correpsonding to } c_l]$

(where $\overline{y}_l$ is the sample mean response for training cases with x_i in the interval corresponding to c_l).

3.15. (5HW-18) Consider a SEL prediction problem where $p=1$, and the class of functions used for prediction is the set of linear functions $S=\{h\,|\,h(x)=b_0+b_1x\ \forall x \text{ and some } b_0,b_1\in\Re\}$. Suppose that in fact

$$x\sim\mathrm{U}(0,1),\ \mathrm{E}[y\,|\,x]=x+2x^2,\ \text{and}\ \mathrm{Var}[y\,|\,x]=.25x^2 \qquad (*)$$

a) Under this model, what is the best element of S, say g^*, for predicting y? Use this to find the modeling penalty/average squared model bias in this problem.

b) What is the smallest possible expected loss here (the mean squared prediction error of the theoretically best predictor, $f(x)=x+2x^2$)?

Now consider the situation where $N=50$ and simple linear regression (OLS) is used to choose an element of S based on a training set. Simulate a large number of training sets (at least 1000 of them) of this size according to model (*) using normal conditional distributions for $y\,|\,x$. For each simulated training set, find the simple linear regression slope and intercept and use these to estimate the mean vector and covariance matrix for the fitted regression coefficients (for this sample size and this model). Use the estimated mean and covariance as follows.

c) Estimate the linear function of x that is the difference between your answer to **a)** and the average linear function produced by SLR in this context. Find the expected square of this difference according to the $\mathrm{U}(0,1)$ distribution of x. (This is an estimate of the expected squared fitting bias here.)

d) Using your estimated covariance matrix, approximate the function of x that is the variance (across training sets) of the value on the least squares line at x. Find the mean of this function according to the $\mathrm{U}(0,1)$ distribution of x. (This is an estimate of the expected prediction variance.)

e) In light of **c)** and **d)** what is the (estimated by simulation) fitting penalty in this context? What then is an approximate value for Err?

Section 4: Problems Concerning Cross-Validation

4.1. (5E1-18) Consider predictors of y for $x \in [0,1]$ based on the small set of "basis functions"

$$h_1(x) = I\left[0 \le x \le \frac{1}{3}\right], h_2(x) = I\left[\frac{1}{3} < x \le \frac{2}{3}\right], \text{ and } h_3(x) = I\left[\frac{2}{3} < x \le 1\right]$$

and the very small training set

Case (i)	1	2	3	4	5	6
y_i	0	4	10	12	6	10
x_i	.1	.3	.4	.6	.7	.9

Without bothering to center y, consider using OLS to fit a predictor for y of the form $\hat{f}(x) = b_1 h_1(x) + b_2 h_2(x) + b_3 h_3(x)$ to the training set. Evaluate the LOOCV MSPE for this kind of predictor.

4.2. (5E1-18) Use the same training set as in Problem 4.1and without bothering to center y, find the 1-NN SEL predictor for y, say $\hat{f}^{\text{1-NN}}(x)$, and evaluate its LOOCV MSPE. (Specify values of the predictor for all $x \in [0,1]$ except where there are "ties." You don't need to do the arithmetic, but your LOOCV answer should evaluate to a number.)

4.3. (5HW-18) Consider the Ames House Price dataset that can be found here
https://vardeman.public.iastate.edu/stat342/stat342.html
and possible predictors of Price. In particular, consider the $p = 4$ inputs Size, Fireplace, Basementbath, and Land. There are, of course, $2^4 = 16$ possible multiple linear regression predictors to be built from these features (including the one with no covariates employed). Use repeated 8-fold cross-validation implemented through `caret` to compare these 16 predictors in terms of cross-validation root mean squared prediction errors.

4.4. (5HW-18) Consider the famous "Glass Identification" dataset of German on the UCI Machine Learning Data Repository at https://archive.ics.uci.edu/ml/datasets/Glass+Identification and kNN classification between glass Types 1 and 2.

a) Use repeated 10-fold cross-validation to find what you believe to be a best number of neighbors for this prediction task.

b) For your choice of $k =$ number of neighbors in **a)**, the variable $t(\mathbf{x}) =$ number of Type 2 cases in the k-neighborhood of $\mathbf{x}$ can take values $0,1,2,\ldots,k$. The nearest neighbor classifier classifies to class 2 if $t(\mathbf{x}) \geq k/2$. This is based on $N = 146$ training cases of which 70 are of glass Type 1 and 76 are of glass Type 2. Suppose that you want to use $\pi_1 = .7$ and $\pi_2 = .3$ and 0-1 loss. How, if at all, would you modify the ordinary kNN classifier?

Section 5: Problems Concerning Linear Theory (for Euclidean Spaces and Function Spaces), Matrix Decompositions, Principal Components, Etc.

5.1. (6HW-15) Consider the 5×4 data matrix

$$\mathbf{X}=\begin{bmatrix}2&4&7&2\\4&3&5&5\\3&4&6&1\\5&2&4&2\\1&3&4&4\end{bmatrix}$$

a) Use `R` and find the QR and singular value decompositions of $\mathbf{X}$. What are the two corresponding bases for $C(\mathbf{X})$?

b) Use the singular value decomposition of $\mathbf{X}$ to find the eigen (spectral) decompositions of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ (what are eigenvalues and eigenvectors?).

c) Find the best $rank=1$ and $rank=2$ approximations to $\mathbf{X}$.

Now center the columns of $\mathbf{X}$ to make the centered data matrix $\widetilde{\mathbf{X}}$.

d) Find the singular value decomposition of $\widetilde{\mathbf{X}}$. What are the principal component directions and principal components for the data matrix? What are the "loadings" of the first principal component?

e) Find the best $rank=1$ and $rank=2$ approximations to $\widetilde{\mathbf{X}}$.

f) Find the eigen decomposition of the sample covariance matrix $\frac{1}{5}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$. Find best 1 and 2 component approximations to this covariance matrix.

Now standardize the columns of $\mathbf{X}$ to make the matrix $\widetilde{\widetilde{\mathbf{X}}}$. Repeat parts **d)**, **e)**, and **f)** using this matrix $\widetilde{\widetilde{\mathbf{X}}}$.

5.2. (6HW-11) Repeat problem **5.1** using the matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \\ 1 & 2 & 1 \\ 2 & 2 & 1 \end{bmatrix}$$

5.3. (5HW-14) Consider the small $(7 \times 3\ )$ fake $\mathbf{X}$ matrix below.

$$\mathbf{X} = \begin{pmatrix} 10 & 10 & .1 \\ 11 & 11 & -.1 \\ 9 & 9 & 0 \\ 11 & 9 & -2.1 \\ 9 & 11 & 2.1 \\ 12 & 8 & -4.0 \\ 8 & 12 & 4.0 \end{pmatrix}$$

(Note, by the way, that $\mathbf{x}_3 \approx \mathbf{x}_2 - \mathbf{x}_1$.)

a) Find the QR and singular value decompositions of $\mathbf{X}$. Use the latter and give best $\mathit{rank} = 1$ and $\mathit{rank} = 2$ approximations to $\mathbf{X}$.

b) Subtract column means from the columns of $\mathbf{X}$ to make a centered data matrix. Find the singular value decomposition of this matrix. Is it approximately the same as that in part **a)**? Give the 3 vectors of the principal component scores. What are the principal components for case 1?

Henceforth consider only the centered data matrix of **b)**.

c) What are the singular values? How do you interpret their relative sizes in this context? What are the first two principal component directions? What are the loadings of the first two principal component directions on x_3? What is the third principal component direction? Make scatter plots of 7 points (x_1, x_2) and then 7 points with first co-ordinate the 1st principal component score and the second the 2nd principal component score. How do these compare? Do you expect them to be similar in light of the sizes of the singular values?

c') Find the matrices $\mathbf{X}\mathbf{v}_j\,\mathbf{v}_j'$ for $j = 1,2,3$ and the best $\mathit{rank} = 1$ and $\mathit{rank} = 2$ approximations to $\mathbf{X}$. How are the latter related to the former?

d) Compute the (N divisor) 3×3 sample covariance matrix for the 7 cases. Then find its singular value decomposition and its eigenvalue decomposition. Are the eigenvectors of the sample covariance matrix related to the principal component directions of the (centered) data matrix? If so, how? Are the eigenvalues/singular values of the sample covariance matrix related to the singular values of the (centered) data matrix. If so, how?

5.4. (6E1-11) As it turns out

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}+\frac{1}{\sqrt{20}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{20}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\ 0 & 0 & -\frac{4}{\sqrt{20}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{20}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{20}} \\ 0 & 0 & -\frac{4}{\sqrt{20}} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

Consider a $p=3$ linear prediction problem where the matrix of training inputs, $\mathbf{X}$, is the matrix on the left above and $\mathbf{Y}'=(4,2,2,0,2)$.

a) Find the single principal component ($M=1$) fitted coefficient vector $\hat{\boldsymbol{\beta}}^{\text{PCR}}$.

b) Find the single component ($M=1$) partial least squares vector of predictions, $\hat{\mathbf{Y}}^{\text{PLS}}$. (Provide a numerical answer.)

5.5. (5HW-16) Consider the small ($N=11$) fake $p=2$ set of predictors that can be entered into `R` using:

```
x1<-c(11,12,13,14,13,15,17,16,17,18,19)
x2<-c(18,12,14,16,6,10,14,4,6,8,2)
```

One can standardize variables in `R` using the `scale()` function.

a) Plot raw and standardized versions of 11 predictor pairs (x_1,x_2) on the same set of axes (using different plotting symbols for the two versions and a 1:1 aspect ratio for the plotting).

b) Find sample means, sample standard deviations, and the sample correlations for both versions of the predictor pairs.

c) Consider the small (11×2) fake $\mathbf{X}$ matrices corresponding to the raw and standardized versions of the data. Interpret the first principal component direction vectors for the two versions and say why (in geometric terms) they are much different.

5.6. (6HW-17) Define a symmetric version of a normalized graphical Laplacian by $\mathbf{L}^{*}=\mathbf{G}^{-1/2}\mathbf{L}\mathbf{G}^{-1/2}$. All eigenvalues of this are non-negative (and it has a 0 eigenvalue since $\mathbf{L}^{*}\left(\sqrt{g_1},\sqrt{g_2},\ldots,\sqrt{g_N}\right)'=\mathbf{0}$). For an eigenvalue λ_l^{*} and corresponding eigenvector $\mathbf{v}_l^{*}$, show that

$$\lambda_l^{*}=\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}s_{ij}\left(\frac{v_{li}^{*}}{\sqrt{g_i}}-\frac{v_{lj}^{*}}{\sqrt{g_j}}\right)^2$$

So there is reason to treat vectors $\mathbf{G}^{-1/2}\mathbf{v}_l^{*}$ (or perhaps normalized versions of them) as another set of "graphical features." Cases with similar entries in these vectors (corresponding to small λ_l) can be expected to belong to the same "connected structure" in a training set.

5.7. (6HW-15) Consider the linear space of functions on $[-\pi,\pi]$ of the form

$$f(t)=a+bt+c\sin t+d\cos t$$

Equip this space with the inner-product $\langle f,g\rangle\equiv\int_{-\pi}^{\pi}f(t)g(t)\,dt$ and norm $\|f\|=\langle f,f\rangle^{1/2}$ (to create a small Hilbert space). Use the Gram-Schmidt process to orthogonalize the set of functions $\{1,t,\sin t,\cos t\}$ and produce an orthonormal basis for the space.

5.8. (6HW-15) Consider the linear space of functions on $[0,1]$ of the form

$$f(t)=a+bt+ct^2+dt^3$$

Equip this space with the inner-product $\langle f,g\rangle\equiv\int_0^1 f(t)g(t)\,dt$ and norm $\|f\|=\langle f,f\rangle^{1/2}$ (to create a small Hilbert space). Use the Gram-Schmidt process to orthogonalize the set of functions $\{1,t,t^2,t^3\}$ and produce an orthonormal basis for the space.

5.9. (6HW-13) Consider the linear space of functions on $[0,1]^2$ of the form

$$f(t,s)=a+bt+cs+dt^2+es^2+fts$$

Equip this space with the inner-product $\langle f, g \rangle \equiv \iint\limits_{[0,1]^2} f(t,s)\, g(t,s)\, dtds$ and norm $\| f \| = \langle f, f \rangle^{1/2}$ (to create a small Hilbert space). Use the Gram-Schmidt process to orthogonalize the set of functions $\left\{1, t, s, t^2, s^2, ts\right\}$ and produce an orthonormal basis for the space.

5.10. (6E1-15) Consider the small space of functions on $[-1,1]^2$ that are linear combinations of the 4 functions $1, x_1, x_2,$ and $x_1 x_2$, with inner product defined by $\langle f, g \rangle = \iint\limits_{[-1,1]^2} f(x_1, x_2)\, g(x_1, x_2)\, dx_1 dx_2$. Find the element of this space closest to $h(x_1, x_2) = x_1^2 + x_2^2$ (in the $L_2\left([-1,1]^2\right)$ function space norm $\|h\| = \langle h, h \rangle^{1/2}$). (Note that the functions $1, x_1, x_2,$ and $x_1 x_2$ are orthogonal in this space.)

Section 6: Problems Concerning Penalized Fitting and Shrinking

6.1. (6HW-11) Consider "data augmentation" methods of penalized least squares fitting.

a) Augment a centered $\mathbf{X}$ matrix with p new rows $\sqrt{\lambda}\underset{p\times p}{\mathbf{I}}$ and $\mathbf{Y}$ by adding p new entries 0. Argue that OLS fitting with the augmented dataset returns $\hat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}}$ as a fitted coefficient vector.

b) Show how the "elastic net" fitted coefficient vector $\hat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}^{\text{elastic net}}$ could be found using lasso software and an appropriate augmented dataset.

6.2. (6E1-11) Consider the $p=3$ linear prediction problem with $N=5$ and training data

$$\mathbf{X}=\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{20}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{20}} \\ 0 & 0 & -\frac{4}{\sqrt{20}} \end{pmatrix} \text{ and } \mathbf{Y}=\begin{bmatrix} 2 \\ 3 \\ -1 \\ -1 \\ -3 \end{bmatrix}$$

In answering the following, use the notation that the jth column of $\mathbf{X}$ is $\mathbf{x}_j$.

a) Find the fitted OLS coefficient vector $\hat{\boldsymbol{\beta}}^{\text{OLS}}$.

b) For $\lambda=10$ find the fitted coefficient vector minimizing

$$\left(\mathbf{Y}-\mathbf{X}\,\mathbf{diag}(\mathbf{c})\hat{\boldsymbol{\beta}}^{\text{OLS}}\right)'\left(\mathbf{Y}-\mathbf{X}\,\mathbf{diag}(\mathbf{c})\hat{\boldsymbol{\beta}}^{\text{OLS}}\right)+\lambda\mathbf{1}'\mathbf{c}$$

over choices of $\mathbf{c}\in\Re^3$ with non-negative entries.

c) For $\lambda>0$ find the fitted ridge coefficient vector, $\boldsymbol{\beta}_{\lambda}^{\text{ridge}}$.

d) For $\lambda>0$ find a fitted coefficient vector $\hat{\boldsymbol{\beta}}_{\lambda}^{*}$ minimizing $(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})+\lambda\left(b_2^2+b_3^2\right)$ as a function of $\mathbf{b}\in\Re^3$.

e) Carefully specify the entire Least Angle Regression path of either $\hat{\mathbf{Y}}$ or $\hat{\boldsymbol{\beta}}$ values.

6.3. (6E1-13) Consider the $p=1$ prediction problem with $N=8$ and training data as below

y	8	4	4	0	2	3	6	5
x	.125	.250	.375	.500	.625	.750	.875	1.000

where use of the order $M=2$ Haar basis functions on the unit interval produces

$$\mathbf{X}=\begin{pmatrix} 1 & 1 & \sqrt{2} & 0 & 2 & 0 & 0 & 0 \\ 1 & 1 & \sqrt{2} & 0 & -2 & 0 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & -2 & 0 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & -2 & 0 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & 2 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & -2 \end{pmatrix}$$

Use the notation that the jth column of $\mathbf{X}$ is $\mathbf{x}_j$.

a) Find the fitted OLS coefficient vector $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ for a model including only $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$ as predictors.

b) Center $\mathbf{Y}$ to create $\mathbf{Y}^*$ and let $\mathbf{x}_j^*=\dfrac{1}{2\sqrt{2}}\mathbf{x}_j$ for each j. Find $\hat{\boldsymbol{\beta}}^{\text{lasso}}\in\Re^7$ optimizing

$$\sum_{i=1}^{8}\left(y_i^*-\sum_{j=2}^{8}b_j x_{ij}^*\right)^2+5\sum_{i=2}^{8}\left|b_j\right|$$

over choices of $\mathbf{b}\in\Re^7$.

c) The LAR algorithm applied to $\mathbf{Y}^*$ and the set of predictors $\mathbf{x}_j^*$ for $j=2,3,\ldots,8$ begins at $\widehat{\mathbf{Y}^*}=\mathbf{0}$ and takes a piecewise linear path through $\Re^8$ to $\widehat{\mathbf{Y}^*}^{\text{OLS}}$. Identify the first two points in $\Re^8$ at which the direction of the path changes, call them $\mathbf{W}_1$ and $\mathbf{W}_2$. (Here you may well wish to use both the connection between the LAR path and the lasso path and explicit formulas for the lasso coefficients.)

d) Find $\hat{\mathbf{Y}}^{\text{penalty}}\in\Re^8$ optimizing

$$(\mathbf{Y}-\mathbf{v})'(\mathbf{Y}-\mathbf{v})+\langle \mathbf{v},\mathbf{x}_2^*\rangle^2+2\left(\langle \mathbf{v},\mathbf{x}_3^*\rangle^2+\langle \mathbf{v},\mathbf{x}_4^*\rangle^2\right)+4\sum_{j=5}^{8}\langle \mathbf{v},\mathbf{x}_j^*\rangle^2$$

over choices of $\mathbf{v}\in\Re^8$.

6.4. (6E1-17) Below are $N=8$ training cases (x_i,y_i) for $x\in[0,1]$ and a corresponding "design matrix" holding values of the first 8 Haar basis functions (in the order $\varphi,\psi,\psi_{1,0},\psi_{1,1},\psi_{2,0},\psi_{2,1},\psi_{2,2},\psi_{2,3}$) for the x_i.

$$\underset{8\times1}{\mathbf{x}}=\begin{bmatrix}1/16\\3/16\\5/16\\7/16\\9/16\\11/16\\13/16\\15/16\end{bmatrix}\quad \underset{8\times1}{\mathbf{y}}=\begin{bmatrix}2\\-1\\3\\-2\\4\\-3\\5\\-4\end{bmatrix}\quad \underset{8\times8}{\mathbf{X}}=\begin{bmatrix}1&1&\sqrt{2}&0&2&0&0&0\\1&1&\sqrt{2}&0&-2&0&0&0\\1&1&-\sqrt{2}&0&0&2&0&0\\1&1&-\sqrt{2}&0&0&-2&0&0\\1&-1&0&\sqrt{2}&0&0&2&0\\1&-1&0&\sqrt{2}&0&0&-2&0\\1&-1&0&-\sqrt{2}&0&0&0&2\\1&-1&0&-\sqrt{2}&0&0&0&-2\end{bmatrix}$$

a) Find the OLS prediction vector $\hat{\mathbf{y}}^{\text{OLS}}$ here. (This is trivial. Note that the 8 columns of $\mathbf{X}$ are orthogonal.)

b) Find the 1-component PLS prediction vector $\hat{\mathbf{y}}^{\text{PLS}}$ here.

c) After normalizing the predictors (so that the $\Re^8$ norm of each column of the normalized $\mathbf{X}$ is 1) find the LASSO prediction vector $\hat{\mathbf{y}}^{\text{LASSO}}$ for the penalty parameter $\lambda=10$. (Center the vector of responses, remove the first column of the $\mathbf{X}$ and work with an 8×7 vector of inputs.)

d) Using the normalized version of the predictors referred to in part **c)** find a vector of coefficients $\boldsymbol{\beta}$ that minimizes

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\boldsymbol{\beta}'\mathbf{diag}(0,0,0,4,4,4,4)\boldsymbol{\beta}$$

6.5. (5HW-14) Here is a small fake dataset with $p=4$ and $N=8$.

y	x_1	x_2	x_3	x_4
3	1	1	1	1
−5	1	1	−1	1
13	1	−1	1	−1
9	1	−1	−1	−1
−3	−1	1	1	−1
−11	−1	1	−1	−1
−1	−1	−1	1	1
−5	−1	−1	−1	1

Notice that the y is centered and the x's are orthogonal (and can easily be made orthonormal by dividing by $\sqrt{8}$). Use the explicit formulas for fitted coefficients in the orthonormal features context to make plots (on a single set of axes for each fitting method, 5 plots in total) of

1. $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3,$ and $\hat{\beta}_4$ versus M for best subset (of size M) regression,
2. $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3,$ and $\hat{\beta}_4$ versus λ for ridge regression,
3. $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3,$ and $\hat{\beta}_4$ versus λ for lasso,
4. $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3,$ and $\hat{\beta}_4$ versus λ for $\alpha=.5$ in the elastic net penalty
$$\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda\left((1-\alpha)\sum_{j=1}^{p}\left|\hat{\beta}_j\right|+\alpha\sum_{j=1}^{p}\hat{\beta}_j^2\right)$$
5. $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3,$ and $\hat{\beta}_4$ versus λ for the non-negative garrote.

Also make 5 corresponding plots of the error sum of squares versus the corresponding parameter.

6.6. (6HW-11) (3.23 of HTF) Suppose that columns of $\mathbf{X}$ with rank p have been standardized, as has $\mathbf{Y}$. Suppose also that

$$\frac{1}{N}\left|\left\langle \mathbf{x}_j, \mathbf{Y}\right\rangle\right| = \lambda \ \forall j=1,\ldots,p$$

Let $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ be the usual least squares coefficient vector and $\hat{\mathbf{Y}}^{\text{OLS}}$ be the usual projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$. Define $\hat{\mathbf{Y}}(\alpha)=\alpha\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{OLS}}$ for $\alpha\in[0,1]$. Find

$$\frac{1}{N}\left|\left\langle \mathbf{x}_j, \mathbf{Y}-\hat{\mathbf{Y}}(\alpha)\right\rangle\right| \ \forall j=1,\ldots,p$$

in terms of $\alpha, \lambda,$ and $\left(\mathbf{Y}-\hat{\mathbf{Y}}^{\text{OLS}}\right)'\left(\mathbf{Y}-\hat{\mathbf{Y}}^{\text{OLS}}\right)$. Show this is decreasing in α. What is the implication of this as regards the LAR algorithm?

6.7. (5HW-14) Return to the context of problem **3.4** and the last/largest set of predictors. Center the y vector to produce (say) $\mathbf{Y}^*$, remove the column of 1's from the $\mathbf{X}$ matrix (giving a 100×9 matrix) and standardize the columns of the resulting matrix, to produce (say) $\mathbf{X}^*$.

a) If one somehow produces a coefficient vector $\boldsymbol{\beta}^*$ for the centered and standardized version of the problem, so that

$$\widehat{y^*} = \beta_1^* x_1^* + \beta_2^* x_2^* + \cdots + \beta_9^* x_9^*$$

what is the corresponding predictor for y in terms of $\left\{1, x, x^2, x^3, x^4, x^5, \sin x, \cos x, \sin 2x, \cos 2x\right\}$?

b) Do the transformations and fit the equation in **a)** by OLS. How do the fitted coefficients and error sum of squares obtained here compare to what you got in problem **3.4**?

c) Augment $\mathbf{Y}^*$ to $\mathbf{Y}^{**}$ by adding 9 values 0 at the end of the vector (to produce a 109×1 vector) and for value $\lambda = 4$ augment $\mathbf{X}^*$ to $\mathbf{X}^{**}$ (a 109×9 matrix) by adding 9 rows at the bottom of the matrix in the form of $\sqrt{\lambda}\underset{9\times 9}{\mathbf{I}}$. What quantity does OLS based on these augmented data seek to optimize? What is the relationship of this to a ridge regression objective?

d) Use trial and error and matrix calculations based on the explicit form of $\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}$ given in the slides for Module 5 to identify a value of $\tilde{\lambda}$ for which the error sum of squares for ridge regression is about 1.5 times that of OLS in this problem. Then make a series of at least 5 values from 0 to $\tilde{\lambda}$ to use as candidates for λ. Choose one of these as an "optimal" ridge parameter λ^{opt} here based on 10-fold cross-validation (as was done in problem **3.4**). Compute the corresponding predictions $\hat{y}_i^{\text{ridge}}$ and plot both them and the OLS predictions as functions of x (connect successive $(x, \hat{y})$ points with line segments). How do the "optimal" ridge predictions based on the 9 predictors compare to the OLS predictions based on the same 9 predictors?

6.8. **(5HW-14)** In light of the idea in part **c)** of problem **6.8**, if you had software capable of doing lasso fitting of a linear predictor for a penalty coefficient λ, how can you use that routine to do elastic net fitting of a linear predictor for penalty coefficients λ_1 and λ_2 in

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p}\left|\hat{\beta}_j\right| + \lambda_2 \sum_{j=1}^{p}\hat{\beta}_j^2 \ ?$$

6.9. (5HW-14) For the dataset of problem **3.4** make up a matrix of inputs based on x consisting of the values of Haar basis functions up through order $m=3$. (You will need to take the functions defined on $[0,1]$ and rescale their arguments to $[-\pi,\pi]$. For a function $g:[0,1]\rightarrow\Re$ this is the function $g^*:[-\pi,\pi]\rightarrow\Re$ defined by $g^*(x)=g\left(\frac{x}{2\pi}+.5\right)$.) 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.

6.10. (6HW-15) For an $N=100$ dataset made up according to the model of problem **3.9** 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.

6.11. (5E1-18) Consider the fitting of predictors of y for $x\in[0,1]$ using the small set of "basis functions"

$$h_1(x)=I\left[0\le x\le\frac{1}{3}\right], h_2(x)=I\left[\frac{1}{3}<x\le\frac{2}{3}\right], \text{ and } h_3(x)=I\left[\frac{2}{3}<x\le1\right]$$

Suppose further that the very small training set given in the table below is available.

Case (i)	1	2	3	4	5	6
y_i	0	4	10	12	6	10
x_i	.1	.3	.4	.6	.7	.9

Center the response (leaving the input as is) and fit a predictor (for centered response) of the form $\hat{f}(x) = b_1 h_1(x) + b_2 h_2(x) + b_3 h_3(x)$ via penalized least squares with penalty $\lambda\left(b_1^2 + b_2^2 + b_3^2\right)$ for $\lambda > 0$. (Give formulas for the 3 coefficients.)

6.12. (5HW-18) Suppose that in a two-class classification problem using coding $\{-1,1\}$ for the classes, the fake data below constitute a very small/toy training set.

y	-1	-1	1	1	1	-1	-1	1
x	1	2	3	4	5	6	7	8

Consider the production of a "voting function" of the form

$$g_{\mathbf{b}}(x) = \sum_{i=1}^{8} b_i \exp\left(-c\left|x - x_i\right|^2\right)$$

by choice of the 8 coefficients b_i (for some choice of $c > 0$) under the "function loss" $h_2(v) = \exp(-v)$ specified in Section 1.5.3 of Stat Learning Notes III. (In the parlance of machine learning, the component functions $\exp\left(-c\left|x - x_i\right|^2\right)$ are data-dependent $p = 1$ "radial basis functions.") In fact, consider "penalized" fitting.

a) One possible penalized fitting criterion is

$$\frac{1}{8}\sum_{i=1}^{8} \exp\left(-y_i g(x_i)\right) + \lambda \sum_{i=1}^{8} b_i^2$$

for some $\lambda > 0$. For choices of $c = .5$ and $c = 1$ optimize this criterion for two different values of $\lambda > 0$ and plot the four resulting voting functions on the same set of axes. Choose (by trial and error, Vardeman doesn't know which ones will work) two values of λ that produce clearly different optimizing functions. (Vardeman is presuming that `optim` or some other canned routine in `R` will be adequate to do this 8-D optimization. If this proves to be incorrect, please let him know.)

b) The function $\mathcal{K}(x,z)=\exp\left(-c\left|x-z\right|^2\right)$ is a "kernel function" in the sense of the phrase specified in Notes III Section 1.4.3. That implies that the 8×8 "Gram matrix"

$$\mathbf{K}=\left(\mathcal{K}\left(x_i,x_j\right)\right)_{\substack{i=1,2,\ldots,8\\ j=1,2,\ldots,8}}$$

is non-negative definite. Thus, with $\mathbf{b}=\left(b_1,b_2,\ldots,b_8\right)'$, $\mathbf{b}'\mathbf{K}\mathbf{b}\geq 0$ and another possible penalized fitting criterion replaces $\sum_{i=1}^{8}b_i^2$ in part **a)** with $\mathbf{b}'\mathbf{K}\mathbf{b}$. For the same values of c and λ you used in part **a)** redo the optimization using this second penalization criterion and plot the resulting voting functions. (Again, if this fails to be possible using an optimizer easily available in R let Vardeman know.) Notice, by the way, that the penalty in **a)** is a $c\rightarrow\infty$ limit of this second penalty!

c) As indicated in Stat Learning Notes III Section 1.4.3, the mapping $T(x)=\mathcal{K}(x,\cdot)$ from $\Re^1$ to functions $\Re^1\rightarrow\Re^1$ picks out $N=8$ functions that are essentially normal pdfs. Linear combinations of these form a linear subspace of this function space. Further, there is a valid inner product $\langle\cdot,\cdot\rangle$ that can be defined on this subspace, for which

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

Using this inner product

1. what is the inner product of two elements of this subspace, say $g_{\mathbf{b}^*}(x)$ and $g_{\mathbf{b}^{**}}(x)$?
2. what is the distance between $T(x)$ and $T(z)$,

   $\left\|T(x)-T(z)\right\|_{\mathcal{A}}=\left\langle T(x)-T(z),T(x)-T(z)\right\rangle_{\mathcal{A}}^{1/2}$?
3. how is the penalty in **b)** related to $\left\|g_{\mathbf{b}}\right\|_{\mathcal{A}}$ (the norm of the linear combination of functions in the function space)?

Section 7: Problems Concerning Kernels and RKHSs

7.1. (5HW-14) The functions

$$\mathcal{K}_1(\mathbf{x},\mathbf{z}) = \exp\left(-\nu\|\mathbf{x}-\mathbf{z}\|^2\right) \quad \text{and}$$

$$\mathcal{K}_2(\mathbf{x},\mathbf{z}) = \left(1+\langle\mathbf{x},\mathbf{z}\rangle\right)^d$$

are legitimate kernel functions for choice of $\nu > 0$ and positive integer d . Find the first two *kernel* principal component vectors for $\mathbf{X}$ in problem **5.3** for each of cases

1. $\mathcal{K}_1$ with two different values of ν (of your choosing), and
2. $\mathcal{K}_2$ for $d = 1, 2$.

If there is anything to interpret (and there may not be) give interpretations of the pairs of vectors for each of the 4 cases. (Be sure to use the vectors for "centered versions" of latent feature vectors.)

7.2. (6HW-17) The function of $(\mathbf{x},\mathbf{z}) \in \Re^p \times \Re^p$ defined by

$$\mathcal{K}(\mathbf{x},\mathbf{z}) = \left(1+c\langle\mathbf{x},\mathbf{z}\rangle\right)^d$$

for $c > 0$ and positive integer d is well-known to be a kernel function.

a) Argue that indeed $\mathcal{K}$ is a kernel function (is non-negative definite) using the facts (from Bishop) quoted in the typed notes.

b) For $d = 2$ consider the $c = 1$ and $c = 2$ cases of this construction for $p = 2$.

i) Describe the sets of functions mapping $\Re^2 \to \Re$ that comprise the abstract linear spaces associated with the reproducing kernels. What is the dimension of these spaces?

ii) Identify for each case a transform $T : \Re^2 \to \Re^M$ so that

$$\mathcal{K}(\mathbf{x},\mathbf{z}) = \langle T(\mathbf{x}), T(\mathbf{z})\rangle$$

(an ordinary $\Re^M$ inner product of the transformed data vectors).

iii) For $\mathbf{x},\mathbf{z}$ belonging to $\Re^2$ find the distances in the two function spaces (the RKHSs) between $T(\mathbf{x})(\cdot) = \mathcal{K}(\mathbf{x},\cdot)$ and $T(\mathbf{z})(\cdot) = \mathcal{K}(\mathbf{z},\cdot)$. (Notice that these are not the same. Metrics implied by the kernels change with the kernels.)

iv) Below is a small fake dataset. For the $c=1$ case, consider these data in the order listed and use as many of the data vectors as necessary to produce a (data-dependent) orthonormal basis for the function space. (Use the Gram-Schmidt process in the RKHS.)

x_1	x_2
1	0
0	1
−1	0
0	−1
2	2
−1	1
−2	−2
1	−1

v) Note that the fake dataset of part **iv)** is centered in $\Re^2$. Find ordinary principal component direction vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ and corresponding 8-dimensional vectors of principal component scores for the dataset. Then find the first two kernel principal component vectors corresponding to the $c=1$ case of $\mathcal{K}$.

7.3. (5E1-16) "Kernel" methods in statistical learning are built on the fact that for a legitimate kernel function $\mathcal{K}(\mathbf{x},\mathbf{z})$ there is an abstract linear space and a transform $T(\mathbf{x})$ from $\Re^p$ to that space for which the inner product of transformed elements of $\Re^p$ is

$$\langle T(\mathbf{x}),T(\mathbf{z})\rangle=\mathcal{K}(\mathbf{x},\mathbf{z})$$

(The inner product in the abstract space has all the usual linearity properties of an inner product and ξ in the abstract space has squared norm $\|\xi\|^2=\langle\xi,\xi\rangle$.)

Use the Gaussian kernel function $K(\mathbf{x},\mathbf{z})=\exp\left(-\|\mathbf{x}-\mathbf{z}\|^2\right)$ in what follows. ($\|\cdot\|$ is the usual $\Re^p$ norm.)

a) For an input vector $\mathbf{x}_i\in\Re^2$, what is the norm of $T(\mathbf{x}_i)$ in the abstract space?

b) For input vectors $\mathbf{x}_i\in\Re^2$ and $\mathbf{x}_l\in\Re^2$, how is the distance between $T(\mathbf{x}_i)$ and $T(\mathbf{x}_l)$ in the abstract space related to the distance between $\mathbf{x}_i$ and $\mathbf{x}_l$ in $\Re^p$? (Distance between ξ and ω in a linear space with inner product is $\|\xi-\omega\|$.)

7.4. (6E1-11) Suppose that $C \subset \Re^p$. This question is about "kernels" ($\mathcal{K}(\mathbf{x},\mathbf{z})$), non-negative definite functions $C \times C \to \Re$ and corresponding RKHS's.

a) Show that for $\phi : C \to \Re$, the function $\mathcal{K}(\mathbf{x},\mathbf{z}) = \phi(\mathbf{x})\phi(\mathbf{z})$ is a valid kernel. (You must show that for distinct $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$, the $M \times M$ matrix $\mathbf{K} = \left(\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)$ is non-negative definite.

b) Show that for two kernels $\mathcal{K}_1(\mathbf{x},\mathbf{z})$ and $\mathcal{K}_2(\mathbf{x},\mathbf{z})$ and two positive constants c_1 and c_2, the function $c_1\mathcal{K}_1(\mathbf{x},\mathbf{z}) + c_2\mathcal{K}_2(\mathbf{x},\mathbf{z})$ is a kernel.

c) By virtue of **a)** and **b)**, the functions $\mathcal{K}_1(x,z) = 1 + xz$ and $\mathcal{K}_2(x,z) = 1 + 2xz$ are both kernels on $[-1,1]^2$. They produce different RKHS's. Show these are different.

7.5. (6E1-11) Consider the RKHS of functions on $[-2,2]$ defined by the kernel $\mathcal{K}(x,z) = 1 + xz \cdot \exp(x+z)$ on $[-2,2]^2$. You may take as given the fact that all functions $\mathcal{K}(x,c)$ of x for a $c \in [-2,2]$ belong to this RKHS, $\mathcal{H}_{\mathcal{K}}$.

a) Show that the functions $g(x) = 1$ and $h(x) = x\exp(x)$ both belong to $\mathcal{H}_{\mathcal{K}}$.

b) Determine whether or not g and h are orthonormal. If they are not, find an orthonormal basis for the span of $\{g,h\}$ in $\mathcal{H}_{\mathcal{K}}$.

7.6. (6E2-13) Consider the function $\mathcal{K}((x,y),(u,v))$ mapping $[-1,1]^2 \times [-1,1]^2$ to $\Re$ defined by

$$\mathcal{K}((x,y),(u,v)) = (1 + xu + yv)^2 + \exp\left(-(x-u)^2 - (y-v)^2\right)$$

on its domain.

a) Argue carefully that $\mathcal{K}$ is a legitimate "kernel" function.

b) Pick any two linearly independent elements of the RKHS generated by $\mathcal{K}$ (i.e. $\mathcal{H}_{\mathcal{K}}$) and find an orthonormal basis for the 2-dimensional linear sub-space of $\mathcal{H}_{\mathcal{K}}$ they span.

7.7. (6E2-15) Murphy mentions the possibility of "kernelizing" nearest-neighbor classification. ("Kernelization" amounts to mapping $\mathbf{x} \in \Re^p$ to $\mathcal{K}(\mathbf{x},\cdot)$ in a RKHS with kernel $\mathcal{K}(\cdot,\cdot)$, $\mathcal{H}_{\mathcal{K}}$, and using inner products and corresponding distances in that space.) Using the Gaussian kernel $\mathcal{K}(\mathbf{x},\mathbf{z}) = \exp\left(-\|\mathbf{x}-\mathbf{z}\|^2\right)$, **what is** the $\mathcal{H}_{\mathcal{K}}$ distance between $\mathcal{K}(\mathbf{x},\cdot)$ and $\mathcal{K}(\mathbf{z},\cdot)$? **Describe** the set of training cases $\mathbf{x}_i \in \Re^p$ with $\mathcal{K}(\mathbf{x}_i,\cdot)$ in the $\mathcal{H}_{\mathcal{K}}$ k-nearest neighborhood of $\mathcal{K}(\mathbf{x},\cdot)$.

7.8. (6E1-17) In the class notes there is an assertion that for a finite set $\mathcal{B}$, say $\mathcal{B} = \{b_1, b_2, \ldots, b_m\}$, for $|A|$ the number of elements in $A \subset \mathcal{B}$, one kernel function on subsets of $\mathcal{B}$ is

$$\mathcal{K}(A,B) = 2^{|A \cap B|}$$

($\mathcal{B}$ could, for example, be a list of attributes that an item might or might not possess.)

a) Prove that $\mathcal{K}$ is a kernel function using the "kernel mechanics" facts in the notes. (Hint: You may find it useful to associate with each $A \subset \mathcal{B}$ an m-dimensional vector of 0s and 1s, call it $\mathbf{x}_A \in \{0,1\}^m$, with $x_{Al} = 1$ exactly when $b_l \in A$.)

b) Let $T(A)(\cdot) = \mathcal{K}(A,\cdot) = 2^{|A \cap \cdot|}$ map subsets of $\mathcal{B}$ to real-valued functions of subsets of $\mathcal{B}$. In the abstract space $\mathcal{A}$ (of real-valued functions of subsets of $\mathcal{B}$) what is the distance between $T(A)$ and $T(B)$, $\|T(A)-T(B)\|_{\mathcal{A}}$?

c) For N training "vectors" (A_i, y_i) ($A_i \subset \mathcal{B}$ and $y_i \in \Re$) consider the corresponding N points in $\mathcal{A} \times \Re$, namely $(T(A_i), y_i)$. Define a k-neighborhood $N_k(V)$ of a point (function) $V \in \mathcal{A}$ to be a set of k points (functions) $T(A_i)$ with smallest $\|T(A_i)-V\|_{\mathcal{A}}$.
Carefully describe a SEL *k*NN predictor of y, $f(V)$, mapping elements V of $\mathcal{A}$ to real numbers $\hat{y}$ in $\Re$. Then describe as completely as possible the corresponding predictor $f(T(A))$ mapping $A \subset \mathcal{B}$ to $\hat{y} \in \Re$.

d) A more direct method of producing a kind of *k*NN predictor of y is to take account of the hint for part **a)** and for subsets A and C of $\mathcal{B}$, associate m-vectors of 0s and 1s respectively $\mathbf{x}_A$ and $\mathbf{x}_C$ and define a distance between sets A and C as the Euclidean distance between $\mathbf{x}_A$ and $\mathbf{x}_C$. This typically produces a different predictor than the one in part **c)**. Argue this point by considering distances from $\mathbf{x}_A$ to $\mathbf{x}_C$ and from $\mathbf{x}_A$ to $\mathbf{x}_D$ in $\Re^m$ and from $T(A)$ to $T(C)$ and from $T(A)$ to $T(D)$ in the space $\mathcal{A}$ for cases with $|A| = 10, |C| = 4, D = |5|, |A \cap C| = 2,$ and $|A \cap D| = 3$.

7.9. (6HW-13) For a $c > 0$, consider the function $K : \Re^2 \times \Re^2 \to \Re$ defined by

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \exp\left(-c\|\mathbf{x} - \mathbf{z}\|^2\right)$$

a) Use the facts about "kernel functions" in the course outline to argue that $\mathcal{K}$ is a kernel function. (Note that $\|\mathbf{x} - \mathbf{z}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{z}, \mathbf{z} \rangle - 2\langle \mathbf{x}, \mathbf{z} \rangle$.)

b) Argue that there is a $\boldsymbol{\varphi} : \Re^2 \to \Re^\infty$ so that with (infinite-dimensional) feature vector $\boldsymbol{\varphi}(\mathbf{x})$ the kernel function is a "regular $\Re^\infty$ inner-product"

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\varphi}(\mathbf{x}), \boldsymbol{\varphi}(\mathbf{z}) \rangle_\infty = \sum_{l=1}^{\infty} \varphi_l(\mathbf{x}) \varphi_l(\mathbf{z})$$

(You will want to consider the Taylor series expansion of the exponential function about 0 and co-ordinate functions of $\boldsymbol{\varphi}$ that are multiples of all possible products of the form $x_1^p x_2^q$ for non-negative integers p and q. It is not necessary to find explicit forms for the multipliers, though that can probably be done. You do need to argue carefully though, that such a representation is possible.)

7.10. (5E1-18) The function

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \exp\left(-|x_1 - z_1| - |x_2 - z_2|\right)$$

mapping $\Re^2 \times \Re^2 \to \Re$ is a kernel function. Consider three real-valued functions (of $\mathbf{z} \in \Re^2$): $T_{(1,0)}(\mathbf{z}) = \mathcal{K}((1,0), \mathbf{z}) = \exp(-|1 - z_1| - |z_2|)$, $T_{(0,1)}(\mathbf{z}) = \mathcal{K}((0,1), \mathbf{z}) = \exp(-|z_1| - |1 - z_2|)$, and $T_{(0,0)}(\mathbf{z}) = \mathcal{K}((0,0), \mathbf{z}) = \exp(-|z_1| - |z_2|)$. Using the inner product for the linear space of functions mapping $\Re^2 \to \Re$ defined for kernel slices by $\langle T_\mathbf{x}, T_\mathbf{w} \rangle = \mathcal{K}(\mathbf{x}, \mathbf{w})$, find the projection of $T_{(0,0)}$ onto the subspace of functions spanned by the two functions $T_{(1,0)}$ and $T_{(0,1)}$ (i.e. the set of all linear combinations $c \cdot T_{(1,0)} + d \cdot T_{(0,1)}$ for constants c and d).

7.11. (6E2-15) Consider the Gaussian kernel $\mathcal{K}(x, z) = \exp\left(-(x - z)^2\right)$ for x and z in $[-2, 4]$ and a corresponding RKHS, $\mathcal{H}_\mathcal{K}$. Based on the very small (x, y) dataset below, we wish to fit a function of the form $f(x) = \alpha_0 + \alpha_1 x + h(x)$ for $h \in \mathcal{H}_\mathcal{K}$ to the dataset under the penalty

$$\sum_{i=1}^{5}\left(y_i - f\left(x_i\right)\right)^2 + 2\|h\|_{\mathcal{H}_{\mathcal{K}}}^2$$

You may use the fact that the least squares line through these data is $\hat{y} = 3.7 - .5x$. Find the optimizing $f\left(x\right)$.

y	4	4	3	3	2
x	-1	0	1	2	3

Section 8: Problems Concerning Sets of Basis Functions

8.1. (6HW-11) Find a set of basis functions for the natural (linear outside the interval (ξ_1, ξ_K)) quadratic regression splines with knots at $\xi_1 < \xi_2 < \cdots < \xi_K$.

8.2. (6HW-11) (B-Splines) For $a < \xi_1 < \xi_2 < \cdots < \xi_K < b$ consider the B-spline bases of order m, $\{B_{i,m}(x)\}$ defined recursively as follows. For $j < 1$ define $\xi_j = a$, and for $j > K$ let $\xi_j = b$. Define

$$B_{i,1}(x) = I[\xi_i \le x < \xi_{i+1}]$$

(in case $\xi_i = \xi_{i+1}$ take $B_{i,1}(x) \equiv 0$) and then

$$B_{i,m}(x) = \frac{x - \xi_i}{\xi_{i+m-1} - \xi_i} B_{i,(m-1)}(x) + \frac{\xi_{i+m} - x}{\xi_{i+m} - \xi_{i+1}} B_{i+1,(m-1)}(x)$$

(where we understand that if $B_{i,l}(x) \equiv 0$ its term drops out of the expression above). For $a = -0.1$ and $b = 1.1$ and $\xi_i = (i-1)/10$ for $i = 1, 2, \ldots, 11$, plot the non-zero $B_{i,3}(x)$. Consider all linear combinations of these functions. Argue that any such linear combination is piecewise quadratic with first derivatives at every ξ_i. If it is possible to do so, identify one or more linear constraints on the coefficients (call them c_i) that will make $q_{\mathbf{c}}(x) = \sum_i c_i B_{3,i}(x)$ linear to the left of ξ_1 (but otherwise minimally constrain the form of $q_{\mathbf{c}}(x)$).

8.3. (6HW-13) Consider the space of continuous functions on $[0,1] \times [0,1]$ that are linear (i.e. are of the form $y = a + bx_1 + cx_2$) on each of the squares

$$S_1 = [0,.5] \times [0,.5], S_2 = [0,.5] \times [.5,1], S_3 = [.5,1] \times [0,.5], \text{ and } S_4 = [.5,1] \times [.5,1]$$

a) Find a set of basis functions for the space described above.

b) Vardeman will send you a dataset generated from a model with

$$\mathrm{E}[y \mid x_1, x_2] = 2x_1x_2$$

Find the best fitting linear combination of the basis functions according to least squares.

c) Describe a set of basis functions for all continuous functions on $[0,1] \times [0,1]$ that for

$$0 = \xi_0 < \xi_1 < \xi_2 < \cdots < \xi_{K-1} < \xi_K = 1 \text{ and } 0 = \eta_0 < \eta_1 < \cdots < \eta_{M-1} < \eta_M = 1$$

are linear on each rectangle $S_{km} = [\xi_{k-1}, \xi_k] \times [\eta_{m-1}, \eta_m]$. How many such basis functions are needed to represent these functions?

8.4. (5E1-14) Suppose one desires to fit a function to N data pairs (x_i, y_i) that is linear outside the interval $[0,1]$, is quadratic in each of the intervals $[0,.5]$ and $[.5,1]$ and has a first derivative for all x (has no sharp corners). Specify 4 functions $h_1(x), h_2(x), h_3(x),$ and $h_4(x)$ and one linear constraint on coefficients $\beta_0, \beta_1, \beta_2, \beta_3,$ and β_4 so that the function

$$y = \beta_0 + \beta_1 h_1(x) + \beta_2 h_2(x) + \beta_3 h_3(x) + \beta_4 h_4(x)$$

is of the desired form.

8.5. (6HW-11) Consider radial basis functions built from kernels. In particular, consider the choice $D(t) = \phi(t)$, the standard normal pdf.

a) For $p = 1$, plot on the same set of axes the 11 functions

$$K_\lambda(x, \xi_j) = D\left(\frac{|x - \xi_j|}{\lambda}\right) \text{ for } \xi_j = \frac{j-1}{10} \quad j = 1, 2, \ldots, 11$$

first for $\lambda = .1$ and then (in a separate plot) for $\lambda = .01$. Then make plots on the a single set of axes the 11 normalized functions

$$N_{\lambda j}(x) = \frac{K_\lambda(x, \xi_j)}{\sum_{l=1}^{11} K_\lambda(x, \xi_l)}$$

first for $\lambda = .1$, then in a separate plot for $\lambda = .01$.

b) For $p = 2$, consider the 121 basis functions

$$K_\lambda(\mathbf{x}, \boldsymbol{\xi}_{ij}) = D\left(\frac{\|\mathbf{x} - \boldsymbol{\xi}_{ij}\|}{\lambda}\right) \text{ for } \boldsymbol{\xi}_{ij} = \left(\frac{i-1}{10}, \frac{j-1}{10}\right) \quad i = 1, 2, \ldots, 11 \text{ and } j = 1, 2, \ldots, 11$$

Make contour plots for $K_{.1}(\mathbf{x}, \boldsymbol{\xi}_{6,6})$ and $K_{.01}(\mathbf{x}, \boldsymbol{\xi}_{6,6})$. Then define

$$N_{\lambda ij}(\mathbf{x}) = \frac{K_\lambda(\mathbf{x}, \boldsymbol{\xi}_{ij})}{\sum_{m=1}^{11}\sum_{l=1}^{11} K_\lambda(\mathbf{x}, \boldsymbol{\xi}_{lm})}$$

Make contour plots for $N_{.1,6,6}(\mathbf{x})$ and $N_{.01,6,6}(\mathbf{x})$.

8.6. (5HW-14) For the dataset of problem **3.4** make up a 100×7 matrix $\mathbf{X_h}$ of inputs based on x consisting of the values of basis functions for natural cubic splines with knots ξ_j

$$h_1(x) = 1, h_2(x) = x, \text{ and}$$

$$h_{j+2}(x) = (x - \xi_j)_+^3 - \left(\frac{\xi_K - \xi_j}{\xi_K - \xi_{K-1}}\right)(x - \xi_{K-1})_+^3 + \left(\frac{\xi_{K-1} - \xi_j}{\xi_K - \xi_{K-1}}\right)(x - \xi_K)_+^3 \text{ for } j = 1, 2, \ldots, K-2$$

(on slide 7 of Module 9) for the 7 knot values

$$\xi_1 = -3.0, \xi_2 = -2.0 \xi_3 = -1.0, \xi_4 = 0.0, \xi_5 = 1.0, \xi_6 = 2.0, \xi_7 = 3.0$$

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

8.7. (6HW-15) Vardeman will send out a dataset of $N = 100$ pairs made up according to the model of problem **3.9** make up a 100×7 matrix $\mathbf{X_h}$ of inputs based on x consisting of the values of basis functions for natural cubic splines with knots ξ_j

$$h_1(x) = 1, h_2(x) = x, \text{ and}$$

$$h_{j+2}(x) = (x - \xi_j)_+^3 - \left(\frac{\xi_K - \xi_j}{\xi_K - \xi_{K-1}}\right)(x - \xi_{K-1})_+^3 + \left(\frac{\xi_{K-1} - \xi_j}{\xi_K - \xi_{K-1}}\right)(x - \xi_K)_+^3 \text{ for } j = 1, 2, \ldots, K-2$$

(on slide 7 of Module 9) 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.

8.8. (6HW-17) Vardeman will send you a dataset giving the maximum numbers of home runs hit by a "big league" professional baseball player in the US for each of 145 consecutive seasons. Consider these as values $y_1, y_2, \ldots, y_{145}$ and take $x_i = i$. Consider the basis functions for natural cubic splines with knots ξ_j (given on panel 7 of Module 9)

$$h_1(x) = 1, h_2(x) = x, \text{ and}$$

$$h_{j+2}(x) = (x - \xi_j)_+^3 - \left(\frac{\xi_K - \xi_j}{\xi_K - \xi_{K-1}}\right)(x - \xi_{K-1})_+^3 + \left(\frac{\xi_{K-1} - \xi_j}{\xi_K - \xi_{K-1}}\right)(x - \xi_K)_+^3 \text{ for } j = 1, 2, \ldots, K-2$$

Using knots $\xi_j = 2 + (j-1)14$ for $j = 1, 2, \ldots, 11$ fit a natural cubic regression spline to the home run data. Plot the fitted function on the same axes as the data points.

8.9. (6E1-11) Consider a toy $p=1$ SEL prediction problem with training data below.

x	-1.0	$-.75$	$-.50$	$-.25$	0	$.25$	$.50$	$.75$	1.0
y	0	2	3	5	4	4	2	2	1

Set up an $\mathbf{X}$ matrix for an ordinary multiple linear regression that could be used to fit a linear regression spline with knots at $\xi_1=-.5, \xi_2=0,$ and $\xi_3=.5$. For your set-up, what linear combination of fitted regression parameters produces the prediction at $x=0$?

Section 9: Problems Concerning Smoothing

9.1. (6HW-11) Suppose that $a < x_1 < x_2 < \cdots < x_N < b$ and $s(x)$ is a natural cubic spline with knots at the x_i interpolating the points (x_i, y_i) (i.e. $s(x_i) = y_i$).

a) Let $z(x)$ be any twice continuously differentiable function on $[a,b]$ also interpolating the points (x_i, y_i). Show that

$$\int_a^b (s''(x))^2\, dx \leq \int_a^b (z''(x))^2\, dx$$

(Hint: Consider $d(x) = z(x) - s(x)$, write

$$\int_a^b (d''(x))^2 dx = \int_a^b (z''(x))^2 dx - \int_a^b (s''(x))^2 dx - 2\int_a^b s''(x) d''(x)\, dx$$

and use integration by parts and the fact that $s'''(x)$ is piecewise constant.)

b) Use **a)** and prove that the minimizer of $\sum_{i=1}^N (y_i - h(x_i))^2 + \lambda \int_a^b (h''(x))^2 dx$ over the set of twice continuously differentiable functions on $[a,b]$ is a natural cubic spline with knots at the x_i.

9.2. (6HW-11) Let $\mathcal{H}$ be the set of absolutely continuous functions on $[0,1]$ with square integrable first derivatives (that exist except possibly at a set of measure 0). Equip $\mathcal{H}$ with an inner product

$$\langle f, g \rangle_{\mathcal{H}} = f(0) + g(0) + \int_0^1 f'(x) g'(x)\, dx$$

a) Show that

$$R(x,z) = 1 + \min(x,z)$$

is a reproducing kernel for this Hilbert space.

b) Using Heckman's development, describe as completely as possible

$$\underset{h \in \mathcal{H}}{\arg\min} \left(\sum_{i=1}^N (y_i - h(x_i))^2 + \lambda \int_0^1 (h'(x))^2\, dx \right)$$

c) Using Heckman's development, describe as completely as possible

$$\underset{h \in \mathcal{H}}{\arg\min} \left(\sum_{i=1}^N \left(y_i - \int_0^{x_i} h(t)\, dt \right)^2 + \lambda \int_0^1 (h'(x))^2\, dx \right)$$

9.3. (6HW-11) Suppose that with $p=1$,

$$y \mid x \sim \mathrm{N}\left(\frac{\sin\left(12\left(x+.2\right)\right)}{x+.2},1\right)$$

and $N=101$ training data pairs are available with $x_i=\left(i-1\right)/100$ for $i=1,2,\ldots,101$. Vardeman will send you a text file with a dataset like this in it. Use it in the following.

a) Fit all of the following using first 5 and then 9 effective degrees of freedom

- **i)** a cubic smoothing spline,
- **ii)** a locally weighted linear regression smoother based on a normal density kernel, and
- **iii)** a locally weighted linear regression smoother based on a tricube kernel.

Plot for 5 effective degrees of freedom all of y_i and the 3 sets of smoothed values against x_i. Connect the consecutive $\left(x_i,\hat{y}_i\right)$ for each fit with line segments so that they plot as "functions." Then redo the plotting for 9 effective degrees of freedom.

b) For all of the fits in **a)** plot as a function of i the coefficients c_i applied to the observed y_i in order to produce $\hat{f}\left(x\right)=\sum_{i=1}^{101}c_iy_i$ for $x=.05,.1,.2,.3$. (Make a different plot of three curves for 5 degrees of freedom and each of the values x (four in all). Then redo the plotting for 9 degrees of freedom.)

9.4. (6HW-11) Center the outputs for the dataset of problem **9.3**. Then derive sets of predictions $\hat{y}_i$ based on $\mu\left(x\right)\equiv 0$ Gaussian process priors for $f\left(x\right)$. Plot several of those as functions on the same set of axes (along with centered original data pairs) as follows:

a) Make one plot for cases with $\sigma^2=1$, $\rho\left(\Delta\right)=\exp\left(-c\Delta^2\right)$, $\tau^2=1,4$ and $c=1,4$.

b) Make one plot for cases with $\sigma^2=1,\rho\left(\Delta\right)=\exp\left(-c\left|\Delta\right|\right),\tau^2=1,4$, and $c=1,4$.

c) Make one plot for cases with $\sigma^2=.25$, but otherwise the parameters of a) are used.

9.5. (6HW-11) A $p=2$ dataset consists of $N=441$ training vectors $\left(x_{1i},x_{2i},y_i\right)$ for the distinct pairs $\left(x_{1i},x_{2i}\right)$ in the set $\left\{-1.0,-.9,\ldots,.9,1.0\right\}^2$ where the y_i were generated as

$$y_i = \frac{\sin\left(10\left(x_{1i}+x_{2i}\right)\right)}{10\left(x_{1i}+x_{2i}\right)} + \varepsilon_i$$

(with the convention that $\sin(0)/0 = 1$) for iid $\mathrm{N}\left(0,(.02)^2\right)$ variables ε_i. Vardeman will send you such a data file. Use it in the following.

a) Why should you expect MARS to be ineffective in producing a predictor in this context? (You may want to experiment with the `earth` package in `R` trying out MARS.)

b) Try 2-d locally weighted regression smoothing on this dataset using the `loess` function in `R`. Contour plot your fits for 2 different choices of smoothing parameters.

c) Try fitting a thin plate spline to these data. There is an old tutorial at

http://www.stat.wisc.edu/~xie/thin_plate_spline_tutorial.html

for using the `Tps` function in the `fields` package that might make this simple.

d) Center the outputs. Then derive a set of predictions $\hat{y}_i$ based on a $\mu(\mathbf{x}) \equiv 0$ prior for $f(\mathbf{x})$. Use $\sigma^2 = (.02)^2$, $\rho(\mathbf{x}-\mathbf{z}) = \exp\left(-2\|\mathbf{x}-\mathbf{z}\|^2\right)$, and $\tau^2 = .25$. How do these compare to the ones you made in a), b) and c)?

e) If you were going to use a structured kernel and 1-d smoothing to produce a predictor here, what form for the matrix A would work best? What would be a completely ineffective choice of a matrix A? Use the good choice of A and produce a corresponding predictor.

9.6. (6HW-13) Suppose that with $p = 1$,

$$y \mid x \sim \mathrm{N}\left(\sin\left(\frac{1.5}{x+.1}\right) + \exp(-2x), (.5)^2\right)$$

(the conditional standard deviation is $.5$) and $N = 101$ training data pairs are available with $x_i = (i-1)/100$ for $i = 1,2,\ldots,101$. Vardeman will send you a text file with a dataset like this in it. Use it in in place of the dataset described in problem **9.3** and redo all of that problem.

9.7. (6HW-13) A $p = 2$ dataset consists of $N = 81$ training vectors (x_{1i}, x_{2i}, y_i) for pairs (x_{1i}, x_{2i}) in the set $\{-2.0,-1.5,\ldots,1.5,2.0\}^2$ where the y_i were generated as

$$y_i = \exp\left(-\left(x_{1i}^2 + x_{2i}^2\right)\right)\left(1 - 2\left(x_{1i}^2 + x_{2i}^2\right)\right) + \varepsilon_i$$

for iid $\mathrm{N}\left(0,(.1)^2\right)$ variables ε_i. Vardeman will send you such a data file. Use it in the following.

a) Why should you expect MARS to be ineffective in producing a predictor in this context? (You may want to experiment with the `earth` package in `R` trying out MARS.)

b) Try 2-d locally weighted regression smoothing on this dataset using the `loess` function in `R`. "Surface plot" this for 2 different choices of smoothing parameters along with both the raw data and the mean function. (If nothing else, `JMP` will do this under its "`Graph`" menu.)

c) Try fitting a thin plate spline to these data. There is an old tutorial at

http://www.stat.wisc.edu/~xie/thin_plate_spline_tutorial.html

for using the `Tps` function in the `fields` package that might make this simple.

d) Center the outputs. Then derive a set of predictions $\hat{y}_i$ based on a $\mu(\mathbf{x}) \equiv 0$ prior for $f(\mathbf{x})$. (Use $\rho(\mathbf{x}-\mathbf{z}) = \exp\left(-2\|\mathbf{x}-\mathbf{z}\|^2\right)$ and what seem to you to be appropriate values of σ^2 and τ^2.) How do your predictions compare to the ones you made in **a)**, **b)** and **c)**?

9.8. (5HW-16) For $p=1$ suppose that N observations (x_i, y_i) have distinct x_i, and for simplicity of notation, suppose that $x_1 < x_2 < \cdots < x_N$. Consider the basis functions for natural cubic splines with K knots ξ_j given on panel 7 of Module 9:

$$h_1(x) = 1, h_2(x) = x, \text{ and}$$

$$h_{j+2}(x) = (x-\xi_j)_+^3 - \left(\frac{\xi_K - \xi_j}{\xi_K - \xi_{K-1}}\right)(x-\xi_{K-1})_+^3 + \left(\frac{\xi_{K-1} - \xi_j}{\xi_K - \xi_{K-1}}\right)(x-\xi_K)_+^3 \quad \text{for } j = 1,2,\ldots,K-2$$

Take $K = N$ and $\xi_j = x_j$ for $j = 1,2,\ldots,N$. Obviously, h_1 and h_2 have second derivative functions that are everywhere 0 and the products of these second derivatives with themselves or 2nd derivatives of other basis functions must have 0 integral from a to b.

Then for $j = 1,2,3,\ldots,N-2$

$$
\begin{aligned}
h''_{j+2}(x) &= 6\left(x-x_j\right) I\left[x_j \le x \le x_{N-1}\right] + 6\left(\left(x-x_j\right)-\left(\frac{x_N - x_j}{x_N - x_{N-1}}\right)\left(x-x_{N-1}\right)\right) I\left[x_{N-1} \le x \le x_N\right] \\
&\quad + 6\left(\left(x-x_j\right)-\left(\frac{x_N - x_j}{x_N - x_{N-1}}\right)\left(x-x_{N-1}\right)+\left(\frac{x_{N-1} - x_j}{x_N - x_{N-1}}\right)\left(x-x_N\right)\right) I\left[x_N \le x \le b\right] \\
&= 6\left(x-x_j\right) I\left[x_j \le x \le x_{N-1}\right] + 6\left(x\left(\frac{x_j - x_{N-1}}{x_N - x_{N-1}}\right)+x_{N-1}\left(\frac{x_N - x_j}{x_N - x_{N-1}}\right)-x_j\right) I\left[x_{N-1} \le x \le x_N\right] \\
&= 6\left(x-x_j\right) I\left[x_j \le x \le x_{N-1}\right] + 6\left(x-x_N\right)\left(\frac{x_j - x_{N-1}}{x_N - x_{N-1}}\right) I\left[x_{N-1} \le x \le x_N\right]
\end{aligned}
$$

Thus for $j = 1, 2, 3, \ldots, N-2$

$$
\begin{aligned}
\int_a^b \left(h''_{j+2}(x)\right)^2 dx &= 12\left(\left(x_{N-1}-x_j\right)^3 + \left(x_N - x_{N-1}\right)^3\left(\frac{x_{N-1}-x_j}{x_N - x_{N-1}}\right)^2\right) \\
&= 12\left(\left(x_{N-1}-x_j\right)^3 + \left(x_N - x_{N-1}\right)\left(x_{N-1}-x_j\right)^2\right) \\
&= 12\left(x_{N-1}-x_j\right)^2\left(x_N - x_j\right)
\end{aligned}
$$

and for positive integers $1 \le j < k \le N-2$

$$
\begin{aligned}
\int_a^b h''_{j+2}(x) h''_{k+2}(x)\, dx &= 36\left(\int_{x_k}^{x_{N-1}}\left(x-x_j\right)\left(x-x_k\right)dx + \int_{x_{N-1}}^{x_N}\left(x-x_N\right)^2 \frac{\left(x_j - x_{N-1}\right)\left(x_k - x_{N-1}\right)}{\left(x_N - x_{N-1}\right)^2} dx\right) \\
&= 36\left(\frac{\left(x_{N-1}-x_k\right)^3}{3} + \left(x_k - x_j\right)\frac{\left(x_{N-1}-x_k\right)^2}{2}\right) - 36\left(\frac{\left(x_j - x_{N-1}\right)\left(x_k - x_{N-1}\right)}{\left(x_N - x_{N-1}\right)^2}\right)\frac{\left(x_{N-1}-x_N\right)^3}{3} \\
&= 6\left(x_{N-1}-x_k\right)^2\left(2\left(x_{N-1}-x_k\right)+3\left(x_k - x_j\right)\right) - 12\left(x_j - x_{N-1}\right)\left(x_k - x_{N-1}\right)\left(x_{N-1}-x_N\right) \\
&= 6\left(x_{N-1}-x_k\right)^2\left(2x_{N-1}+x_k - 3x_j\right) + 12\left(x_{N-1}-x_k\right)\left(x_{N-1}-x_j\right)\left(x_N - x_{N-1}\right)
\end{aligned}
$$

A small smoothing example of Prof. Morris involves an $N = 11$ point data set. Here is `R` code for entering it.

```
> x <- c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1)
> y <- c(1.0030100, 0.8069872, 0.6690364, 0.6281389,
0.5542417, 0.5105527, 0.5306341, 0.5023222, 0.6103748,
0.7008915, 0.9422990
```

Do the smoothing spline computations "from scratch" using the above representations of the entries of the matrix $\mathbf{\Omega}$. That is,

a) Compute the 11×11 matrix $\mathbf{\Omega}$.

b) For $\lambda = 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, and 0 compute the smoother matrices $\mathbf{S}_\lambda$ and the effective degrees of freedom. Compare your degrees of freedom to what Prof. Morris found and compare your $\mathbf{S}_{.001}$ to his.

c) Find the penalty matrix $\mathbf{K}$ and its eigen decomposition. Plot as functions of x_i (or just i assuming that you have ordered the values of x) the entries of the eigenvectors of this matrix (connect successive points with line segments so that you can see how these change in character as the corresponding eigenvalue of $\mathbf{K}$ increases—the corresponding eigenvalue of $\mathbf{S}_\lambda$ decreases). Which $\Re^{11}$ components of the observed $\mathbf{Y}$ are most suppressed in the smoothing operation? Can you describe them in qualitative terms?

Section 10: Problems Concerning Miscellaneous "Standard" SEL Predictors

10.1. (6HW-11) Beginning in Section 5.6, Izenman's book uses an example where PET yarn density is to be predicted from its NIR spectrum. This is a problem where $N = 21$ data vectors $\mathbf{x}_j$ of length $p = 268$ are used to predict the corresponding outputs y_i. Izenman points out that the `yarn` data are to be found in the `pls` package in `R`. (The package actually has $N = 28$ cases. Use all of them in the following.) Get those data and make sure that all inputs are standardized and the output is centered. (Use the N divisor for the sample variance.)

a) Using the `pls` package, find the $1, 2, 3,$ and 4 component PCR and PLS $\hat{\boldsymbol{\beta}}$ vectors.

b) Find the singular values for the matrix $\mathbf{X}$ and use them to plot the function $\text{df}(\lambda)$ for ridge regression. Identify values of λ corresponding to effective degrees of freedom $1, 2, 3,$ and 4. Find corresponding ridge $\hat{\boldsymbol{\beta}}$ vector.

c) Plot on the same set of axes $\hat{\beta}_j$ versus j for the PCR, PLS and ridge vectors for number of components/degrees of freedom 1. (Plot them as "functions," connecting consecutive plotted $(j, \hat{\beta}_j)$ points with line segments.) Then do the same for $2, 3,$ and 4 numbers of components/degrees of freedom.

d) It is (barely) possible to find that the best (in terms of R^2) subsets of $M = 1, 2, 3,$ and 4 predictors are respectively, $\{x_{40}\}, \{x_{212}, x_{246}\}, \{x_{25}, x_{160}, x_{215}\}$ and $\{x_{160}, x_{169}, x_{231}, x_{243}\}$. Find their corresponding coefficient vectors. Use the `lars` package in `R` and find the lasso coefficient vectors $\hat{\boldsymbol{\beta}}$ with exactly $M = 1, 2, 3,$ and 4 non-zero entries with the largest possible $\sum_{j=1}^{268} \left|\hat{\beta}_j^{\text{lasso}}\right|$ (for the counts of non-zero entries).

e) If necessary, re-order/sort the cases by their values of y_i (from smallest to largest) to get a new indexing. Then plot on the same set of axes y_i versus i and $\hat{y}_i$ versus i for ridge, PCR, PLS, best subset, and lasso regressions for number of components/degrees of freedom/number of nonzero coefficients equal to 1. (Plot them as "functions," connecting consecutive plotted (i, y_i) or $(i, \hat{y}_i)$ points with line segments.) Then do the same for 2,3, and 4 numbers of components/degrees of freedom/counts of non-zero coefficients.

f) Section 6.6 of JWHT uses the `glmnet` package for `R` to do ridge regression and the lasso. Use that package in the following. Find the value of λ for which your lasso coefficient vector in **d)** for $M = 2$ optimizes the quantity

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{268}\left|\hat{\beta}_j\right|$$

(by matching the error sums of squares). Then, by using the trick of problem 8, employ the package to find coefficient vectors $\boldsymbol{\beta}$ optimizing

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\left((1-\alpha)\sum_{j=1}^{268}\left|\hat{\beta}_j\right| + \alpha\sum_{j=1}^{268}\hat{\beta}_j^2\right)$$

for $\alpha = 0,.1,.2,\ldots,1.0$. What effective degrees of freedom are associated with the $\alpha = 1$ version of this? How many of the coefficients β_j are non-zero for each of the values of α? Compare error sums of squares for the raw elastic net predictors to the linear predictors using coefficients modified elastic net coefficients

$$(1+\lambda\alpha)\,\hat{\beta}_{\lambda,\alpha}^{\text{elastic net}}$$

Section 11: Problems Concerning Neural Networks and Other Flexible High Dimensional SEL Predictors

11.1. (6HW-11) The "point and click" `JMP` software will fit neural nets (with logistic sigmoidal function, $\sigma(\cdot)$) and random forests. It has pretty good built in help files, from which you should be able to figure out how to use the software.

a) Use `JMP` to produce a neural net fit for the dataset in problem **9.3** with an error sum of squares about like those for the 9 degrees of freedom fits. Provide appropriate JMP reports/summaries. You'll be allowed to vary the number of hidden nodes for a single-hidden-layer architecture and to vary a weight for a penalty made from a sum of squares of coefficients. Each run of the routine makes several random starts of an optimization algorithm. Extract the coefficients from the `JMP` run and use them to plot the fitted function of x that you settle on. How does this compare to the plotted fits produced in 8.3?

b) Use `JMP` to produce a random forest fit for the dataset in problem 8.3 with an error sum of squares about like those for the 9 degrees of freedom fits. Plot fitted values against x and compare to the other fits you've made to this dataset.

c) Try to reproduce what you got from `JMP` in a) and b) using the `R` packages `neuralnet` and `randomForest` (or any others you find that work better).

11.2. (6HW-13) Redo problem **10.1** with the data of problem **9.6**.

11.3. (6HW-13) Consider again the data of problem **9.5**.

a) Use the neural network and random forest routines in `JMP` to fit the data to get error sums of squares like you got in problem **9.5**. How complicated does the network architecture have to be in order to do a good job fitting these data? Contour or surface plot your fits.

b) Fit radial basis function networks based on the standard normal pdf ϕ,

$$f_\lambda(\mathbf{x}) = \beta_0 + \sum_{i=1}^{81} \beta_i K_\lambda(\mathbf{x}, \mathbf{x}_i) \quad \text{for} \quad K_\lambda(\mathbf{x}, \mathbf{x}_i) = \phi\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{\lambda}\right)$$

to these data for two different values of λ. Then define normalized versions of the radial basis functions as

$$N_{\lambda i}(\mathbf{x}) = \frac{K_\lambda(\mathbf{x}, \mathbf{x}_i)}{\sum_{m=1}^{81} K_\lambda(\mathbf{x}, \mathbf{x}_m)}$$

and redo the problem using the normalized versions of the basis functions.

11.4. (6HW-13) Redo problem **11.3** with the data of problem **9.7**.

11.5. (5HW-14) Again use the dataset of problem **3.4**.

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.

b) Produce a single hidden layer neural net fit with an error sum of squares about like those for the 9 degrees of freedom fits using `nnet()`. You may need to vary the number of hidden nodes for a single-hidden-layer architecture and vary the weight for a penalty made from a sum of squares of coefficients in order to achieve this. For the function that you ultimately fit, extract the coefficients and plot the fitted mean function. How does it compare to the plots made in a)?

c) Each run of an `nnet()` begins from a different random start and can produce a different fitted function. Make 5 runs using the architecture and penalty parameter (the "decay" parameter) you settle on for part **b)** and save the 100 predicted values for the 10 runs into 10 vectors. Make a scatterplot matrix of pairs of these sets of predicted values. How big are the correlations between the different runs?

d) Use the `avNNet()` function from the `caret` package to average 20 neural nets with your parameters from part **b)** .

e) Fit radial basis function networks based on the standard normal pdf ϕ,

$$f_\lambda(\mathbf{x}) = \beta_0 + \sum_{i=1}^{81} \beta_i K_\lambda(\mathbf{x}, \mathbf{x}_i) \text{ for } K_\lambda(\mathbf{x}, \mathbf{x}_i) = \phi\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{\lambda}\right)$$

to these data for two different values of λ. Define normalized versions of the radial basis functions as

$$N_{\lambda i}\left(\mathbf{x}\right)=\frac{K_{\lambda}\left(\mathbf{x},\mathbf{x}_{i}\right)}{\sum_{m=1}^{81}K_{\lambda}\left(\mathbf{x},\mathbf{x}_{m}\right)}$$

and redo the problem using the normalized versions of the basis functions.

11.6. (5HW-14) Use all of MARS, thin plate splines, local kernel-weighted linear regression, and neural nets to fit predictors to both the noiseless and the noisy Mexican hat data in `R`. For those methods for which it's easy to make contour or surface plots, do so. Which methods seem most effective on this particular dataset/function?

Section 12: Problems Concerning Rectangle-Based Predictors (Trees and PRIM)

12.1. (5E1-18) Use the training set below and without bothering to center y, carefully build a binary regression tree with 6 final nodes (employing 5 splits, each at one of the values .2, .35, .5, .65, and .8). For each split, give the associated SSE provided by the split. Make a tree diagram for representing your development. If SSE is penalized by $\lambda = 6$ times the number of tree nodes, which of the trees met in your construction is most attractive?

Case (i)	1	2	3	4	5	6
y_i	0	4	10	12	6	10
x_i	.1	.3	.4	.6	.7	.9

Section 13: Problems Concerning Predictors Based on Bootstrapping (Bagging)

13.1. (5E1-18) Use the same training set below and without bothering to center y, consider bagging a SEL predictor for y of the form

$$\hat{f}(x) = b_1 I[x < .5] + b_2 I[x \geq .5]$$

fit by OLS. Below, $B = 10$ bootstrap samples are represented in terms of case indices and the corresponding values of b_1 and b_2 are provided. Find an OOB MSPE for a bagged predictor $\hat{f}_{\text{bag}}^{10}$.

Cases (i)	b_1	b_2
2, 3, 5, 5, 5, 6	7.000	7.000
2, 2, 3, 4, 5, 6	6.000	9.333
1, 1, 1, 1, 2, 6	0.800	10.000
1, 1, 3, 4, 5, 6	3.333	9.333
1, 1, 2, 3, 5, 6	3.500	8.000
1, 1, 2, 4, 4, 5	1.333	10.000
2, 2, 2, 3, 5, 5	4.000	6.000
2, 3, 3, 4, 4, 5	8.000	10.000
1, 2, 2, 2, 3, 4	3.200	12.000
2, 3, 4, 5, 6, 6	7.000	8.666

Section 14: Problems Concerning Ensemble SEL Prediction- Bayes Model Averaging, Boosting, and Stacking

14.1. (5E1-18) Use the same training set below and without bothering to center y, consider using boosting to create a SEL predictor for it. As your set of "basis functions for successive corrections" (the "$h(x,\gamma)$" of the short course slides) adopt the 10 indicator functions

$$l_1(x)=I[x<.2], l_2(x)=I[x<.35], l_3(x)=I[x<.5], l_4(x)=I[x<.65], l_5(x)=I[x<.8]$$
$$u_1(x)=I[x\geq .2], u_2(x)=I[x\geq .35], u_3(x)=I[x\geq .5], u_4(x)=I[x\geq .65], u_5(x)=I[x\geq .8]$$

Take $\hat{f}_0(x)=\bar{y}$ and using a "learning rate" of .5, find $\hat{f}_1(x)$, the first boosted iterate. (This will be $\hat{f}_0(x)$ plus a multiple of one of the indicator functions.)

Case (i)	1	2	3	4	5	6
y_i	0	4	10	12	6	10
x_i	.1	.3	.4	.6	.7	.9

Section 15: Problems Concerning Application of SEL Predictors

15.1. (6HW2-15) This question concerns analysis of a set of sale price data obtained from the Ames City Assessor's Office. There is an `Ames Home Price Data` file on Vardeman's Stat 602 Web page containing information on sales May 2002 through June 2003 of $1\frac{1}{2}$ and 2 story homes built 1945 and before, with (above grade) size of 2500 sq ft or less and lot size 20,000 sq ft or less, located in Low- and Medium-Density Residential zoning areas. $n = 88$ different homes fitting this description were sold in Ames during this period. (2 were actually sold twice, but only the second sales prices of these were included in our data set.) (The rows of the file have been shuffled randomly, so that you may use 8 successive sets of 11 rows as folds for cross-validation purposes if you end up programming you own cross-validations.)

For each home, the value of the response variable

$Price$ = recorded sale price of the home

and the values of 14 potential explanatory variables were obtained. These variables are

Size the floor area of the home above grade in sq ft,

Land the area of the lot the home occupies in sq ft,

Bedrooms a count of the number of bedrooms in the home

Central Air a **dummy** variable that is 1 if the home has central air conditioning and is 0 if it does not,

Fireplace a count of the number of fireplaces in the home,

Full Bath a count of the number of full bathrooms above grade,

Half Bath a count of the number of half bathrooms above grade,

Basement the floor area of the home's basement (including both finished and unfinished parts) in sq ft,

Finished Bsmnt the area of any finished part of the home's basement in sq ft,

Bsmnt Bath a **dummy** variable that is 1 if there is a bathroom of any sort (full or half) in the home's basement and is 0 otherwise,

Garage a **dummy** variable that is 1 if the home has a garage of any sort and is 0 otherwise,

Multiple Car a **dummy** variable that is 1 if the home has a garage that holds more than one vehicle and is 0 otherwise,

Style (*2 Story*) a **dummy** variable that is 1 if the home is a 2 story (or a $2\frac{1}{2}$ story) home and is 0 otherwise, and

Zone (*Town Center*) a **dummy** variable that is 1 if the home is in an area zoned as "Urban Core Medium Density" and 0 otherwise.

a) In preparation for analysis, standardize all explanatory variables that are not dummy variables (those we'll leave in raw form), and center the price variable making a data frame with 15 columns. Say clearly how one goes from a particular new set of home characteristics to a corresponding set of predictors. Then say clearly how a prediction for the centered price to a prediction for the actual dollar price.

b) Find linear predictors for centered price of all the following forms:

- OLS
- Lasso (choose λ by 8-fold cross-validation)
- Ridge (choose λ by 8-fold cross-validation)
- Elastic Net with $\alpha = .5$ (choose λ by 8-foldcross-validation)
- PCR (choose the number of components by 8-fold cross-validation)
- PLS (choose the number of components by 8-fold cross-validation)

For each predictor that you have a way to do so, evaluate the effective degrees of freedom.

c) Plot on the same set of axes as a function of index (1 through 15) the values of co-ordinates $\hat{\beta}_j$ of $\hat{\boldsymbol{\beta}}$ for each predictor in **b)**. (Connect successive coordinates of a given $\hat{\boldsymbol{\beta}}$ with line segments so that you can track the different methods across the plot. Use different plotting symbols and colors for the 6 different methods.) If you see anything interesting in the plot comment on it.

d) Plot on the same set of axes as a function of index (1 through 88) the values of co-ordinates $\hat{y}_i$ of $\hat{\mathbf{Y}}$ for each predictor in **b)**. (Connect successive coordinates of a given $\hat{\mathbf{Y}}$ with line segments so that you can track the different methods across the plot. Use different plotting symbols and colors for the 6 different methods.) If you see anything interesting in the plot comment on it.