Assignment 1

Modern Data Mining in Biostatistics

Due: Fri, Oct 21, 4pm, by email

Preliminaries

Please type your assignment and produce a PDF file which you will email me. If you cannot produce a PDF file you will have to submit by hand **by 12noon Oct21** to Marianne.

YOU ARE TO WORK ALONE ON THIS ASSIGNMENT. YOU ARE NOT TO CONSULT INTERNET FOR SPECIFIC SOLUTIONS OR TO SOLICIT SPECIFIC ADVICE FOR THIS ASSIGNMENT.

Please provide `R`code only for specific functions you wrote (like `my.ridge` or for finding degrees of freedom for HAT matrix). For "obvious" stuff I do not need to see your code.

Question 1

In this question we work with spline basis. [20pts]

Show that truncated power basis set (with a single internal knot, (ψ)):

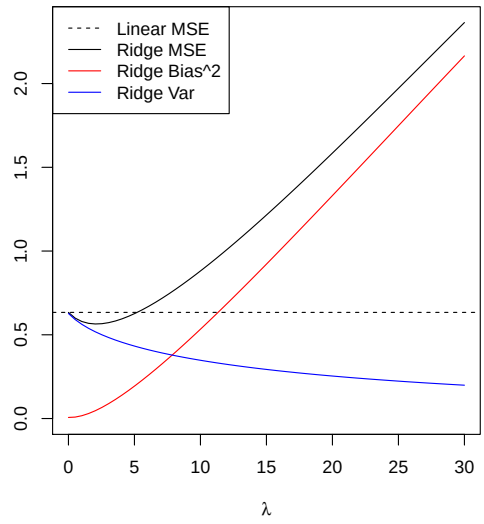
$$1,\, x,\, x^2,\, x^3,\, (x-\psi)^3_+$$

is a basis set for cubic splines. (Recall that *truncated power series* notation, $(x-a)^p_+$, which means either raise the inside to power p if $x > a$, or result is zero otherwise)

- write down the equations for two cubics in each interval and then write down the three constraints at the knot. How many independent variables (degrees of freedom) are left after the constraints?
- rewrite the two cubics using truncated power series notation (for $p = 0, 1, 2, 3$)
- Now expand the truncated power basis and collect like terms. Argue that the two basis sets span the same space
- Finally show that the truncated basis set has zero, first and second dervitives continous at the knot, but not necessarily the third

Question 2

Ridge regression can still outperform linear regression in terms of mean squared error: [35pts]



Only works for λ less than ≈ 5, otherwise it is very biased. (Why?)

13

- Use `runif()` function to generate and save 10 "large" and "small" coefficients as per slides. Use `set.seed()` function to make results reproducible!
- Write R function to generate data set. It should take vector `beta` (with first value being intercept), and parameters `n` and `sigma` for sample size and std dev of the noise and return a matrix, with last column being the response, and first column being vector of ones (intercept)
- Write function `my.ridge()` to fit a simple ridge model using matrix multiplication. It should take the data matrix with first column an assumed intercept and last column assumed the response, y, and parameter `lambda`. Make sure you deal with the intercept correctly.
  Hint: `t(X)` takes a transpose of matrix X. Matrix multiplications in R are done using `%*%` operator. `solve(A,b)` function can solve a system of linear equations as getting vector x in the following system:

$$Ax = b$$

and in the absense of b it will invert matrix A, is possible, but matrix inversion is costly so use the first form if possible. Also function `diag(p)` can be used to obtain identity matrix of dimenion p.

- To produce the figure you will need to obtain values for bias (squared) and variance, and squared prediction error. You have a choice:
  - derive expressions for bias and variance of ridge with fixed λ and data distributed normally, as in this example; or:
  - Use large number of simulations to obtain precise estimates of these statistics. In this case show using point form or pseudo-code how were these obtained
- Plot the results for lambda= a sequence of 100 values between 0 and 30. (in R: `seq(0, 30, length=100)`)

Question 3

We will look at cubic splines and smoothing splines here. [45pts]

- Load library `splines` and read help page for function `ns()`. This let's you obtain basis set for natural cubic splines for a given set of predictor values and knots.
- Read help on `smooth.spline` function which fits a smoothing spline. Pay special attention to how the penaltyy coefficient λ can be specified and what is returned.
- Generate your data:
  1. generate 1000 x-points equispaced between 0, 1 (`seq(0,1,length=1000)`)
  2. Generate the response:

     $$y = x^2 \cdot \sin(20x) + \epsilon$$

     where epsilon is iid Gaussian with sd=0.4.
- obtain a smoothing spline solution of this data with 10 degrees of freedom. What is the smoothing parameter, λ, that corresponds to that? (Hint: look and `names(fit)` where `fit` is the object returned from `smooth.spline` call)

- Generate a natural cubic spline basis on these x-points (by calling `ns()`) using 10 degrees of freedom. What is the dimensionality of the resulting evaluated basis matrix (denoted variously by $\mathbf{\Phi}$ or $\boldsymbol{B}$ in my slides)? What are the knots chosen by `ns()`?

- Modify your `my.ridge` function so that the intercept is assumed to be removed. You will call this function now with your evaluated basis matrix, $\mathbf{\Phi}$ from `ns()`, response vector, and `lambda` and obtains a ridge regression solution, as usual:

$$\widehat{y} = \mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda I\right)^{-1}\mathbf{\Phi}^T y$$

- Produce a ridge fit using you natural cubic basis and lambda=0. Make a single plot with:

  1. True function,
  2. The 1000 generated points and:
  3. the two fitted lines (smoothing spline and ridge with natural cubic spline basis): make them different colors

- What do you see? Discuss in a paragraph addressing the differents of the two fits in terms of number of knots, degrees of freedom, and penalty.

- Repeat the natural cubic spline fit with these changes:

  1. Setup basis with 20 degrees of freedom and find penalty lambda which obtains 10 degree of freedom fit by taking a trace of the Hat matrix. Report lambda.
  2. Setup basis with all internal knots (using `knots` instead of `df` parameter to `ns()` and passing all *internal* x-points as knots. Again try to find lambda which gives close to 10 degrees of freedom and report it

- In both cases produce a plot similar to above and discuss what you see and what you expected to see.

Individual work

You are to work **individually** on this assignment. While you can discuss the problems with classmates and anyone else in *general terms* you are not to attempt to obtain or provide specific solutions or answers to anyone. This also concerns internet use: you can browse for general information, but you should not be attempting to use internet to find specific answers and solutions.