DS 5220 - Fall 2018 Olga Vitek

Homework 4

Please follow the submission instructions in the Piazza post @144

Due on Blackboard before midnight on Friday November 9, 2018.
Each part of the problems 5 points

1. *[Analytical question]* Consider a problem with one predictor X and one continuous target variable (i.e., response) Y, and a dataset with 12 observations. We want to fit a cubic spline with 3 knots to these data. *[Note: you do not need to write the design matrix for the natural spline. A regular cubic spline is enough.]* 1) is it the first model or the secind model

   (a) Write out the basis functions, as well as the overall model for the data in matrix notation (i.e., write the vector Y, the design matrix X, the vector of parameters θ and the vector of errors ϵ. [Y1..Yn]? [e1…e12] Matrix with basis fn as features? x = h(x)? 12X7 [b1…b7]

   (b) Describe the process that can be used to estimate the parameters of the model. New features are the basis fn's and parameters are the Betas. Global cost function?

2. *[Implementation question]* In this problem we will focus on local linear regression. We will consider a dataset that contains yearly counts for Hepatitis A, measles, mumps, pertussis, polio, rubella, and smallpox for US states. You can access the dataset in R, via the code below:

```
install.packages("dslabs")
library("dslabs")
data(us_contagious_diseases)
```

   (a) Implement your own version of local linear regression, with a Gaussian kernel, as function of the tuning parameter λ. *[Note: you can use an existing implementation of weighted linear regression. The rest of the implementation should be your own.]*

   (b) Plot the number of occurrences of measles in California as function of years. Overlay on the plot the local linear regressions fitted to these data, based on your implementation, for several values of the tuning parameter λ. Describe the role of the tuning parameter λ.

   (c) What conclusion regarding the occurrence of the disease can you make from the local regression fit? Can you suggest an explanation for this pattern?

3. *[Implementation question]* In this problem we will use the same dataset as in Homework 3 (IPOs). Randomly partition the dataset into train/dev/evaluation subsets.

   (a) Implement Naïve Bayes classifier described, using Gaussian kernel for density estimation, as function of the tuning parameter λ. (I.e., do not use the existing implementation, but write your own code). *[Hint: refer to HTF Chapter 6.6 for details]*

(b) Use the training set to fit the Naïve Bayes classifier using your implementation above, and classify the presence/absence of venture capital funding with all the predictors. Evaluate the bandwidth parameter λ of the kernel on a grid, and select the value of the parameter that performs best on the development set. Report the predictive performance on the evaluation set.