Assignment 2

Machine Learning COMS 4771

Spring 2014, Itsik Pe'er

Assigned: Feb 3rd (Revised: Feb 5th) Due: Class time, Feb 12th

Submission: Your submission folder on Courseworks

1. Use `SimHousingPrices`[1] to simulate data that is a polynomial with normally-distributed noise. The function `SimPoly` should receive as input:

   `RealThetas`: A real vector θ of D+1 coefficients for a D-degree polynomial $P(x)$

   `StdDev`: A non-negative scalar σ that denotes the scale of fluctuation of the output around the polynomial value

   `x`: A real vector of input datapoints

   The function should provide as output:

   `y`: The outputs. Each output y_i is $P(x_i)+e_i$ where e_i is a simulated value of a normally-distributed random variable, with mean zero and variance σ^2 .

   The function should be in a submitted folder called `Assignment02.Problem01`

   [20 points]

2. Define a cubic polynomial with θ based on the digits in your UNI (mine would be $2x^3+x^2+6x+9$ as my uni is ip2169). Use `SimPoly` to simulate outputs with this polynomial and σ=0.1. Simulate outputs for N training inputs and M testing inputs that are uniformly distributed in [-1,1]. Perform polynomial curve fitting of degrees 0 to 8 by defining the relevant pseudoinverse for the relevant matrices, and compare empirical risks on training and testing data by plotting them along the degree axis. Do all this three times: run #1 with N=10, M=10; run #2 with N=100, M=10; run #3 with N=10, M=100.

   Your code should save files with the following information (as columns of numbers):

   `x.train.[R].txt` - for `R` =1,2,3: 3 training inputs for the corresponding run

   `x.test.[R].txt` - for `R` =1,2,3: 3 testing inputs for the corresponding run

   `y.train.[R].txt` - for `R` =1,2,3: 3 training outputs for the corresponding run

   `y.test.[R].txt` - for `R` =1,2,3: 3 testing outputs for the corresponding run

   `ThetaStar.[R].[D].txt` - for `R` =1,2,3, and `D` =0, … ,8 : 3x11 files, each with the fit coefficients for for the corresponding run and corresponding degree polynomial.

   `Risk.train.[R].txt` - for `R` =1,2,3: 3 training empirical risk values for the corresponding run

   `Risk.test.[R].txt` - for `R` =1,2,3: 3 testing outputs for the corresponding run

   The function to do all of this should be called should be called `FitCubic()` in a submitted folder called `Assignment02.Problem02`

   [60 points]

3. Simulate data for logistic regression. Use `SimHousingPrices`[1] to simulate classification data that is drawn with probability that is logistically dependent on a linear combination of inputs, plus normally-distributed noise. The function `SimLogistic` should receive as input:

   `RealThetas`: A real vector θ of D+1 linear coefficients

   `x`: A real matrix with D columns of input datapoints (row vectors)

   The function should provide as output:

   `y`: The binary outputs. Each output y_i is randomly chosen with probability $Pr(y_i = 1) = 1/(1+exp(-z_i))$ where z_i is a θ-defined linear combination of the coordinates of the i-th input vector

   The function should be in a submitted folder called `Assignment02.Problem03`
   [20 points]

   Good luck![1]

[1] You are encouraged to use the function in the posted solution for Assignment #1. Using your own function is allowed, but it is at your own risk.