CS 6140: Machine Learning — Fall 2021— Paul Hand

HW 6
Due: Wednesday November 10, 2021 at 11:59 PM Eastern time via Gradescope.

Names: [Put Your Name(s) Here]

You can submit this homework either by yourself or in a group of 2. You may consult any and all resources. You may submit your answers to this homework by directly editing this tex file (available on the course website) or by submitting a PDF of a Jupyter or Colab notebook. When you upload your solutions to Gradescope, make sure to tag each problem with the correct page.

Question 1. *Model Validation - KNN*

In this problem, you will build a k-nearest neighbor classifier based on the Iris dataset using 5-fold cross validation. Report the test error on each fold, plot the cross-validation estimate of the test accuracy as a function of k, and present an estimate of the test error of your final predictor. For this problem, code your own solver for KNN.

Response:

Question 2. *Stochastic Gradient Descent*

In this problem, you will build your own solver of Stochastic Gradient Descent. Do not use built-in solvers from any software packages. In this problem, you will use stochastic gradient descent to solve

$$\min_y \frac{1}{n} \sum_{i=1}^{n} (y - x_i)^2, \tag{1}$$

where $x_i \in \mathbb{R}$ for $i = 1 \dots n$.

(a) Using calculus, derive a closed-form expression for the minimizer y^*.

Response:

(b) Generate points $x_i \sim \text{Uniform}[0, 1]$ for $i = 1 \dots 100$. Use Stochastic Gradient Descent with a constant learning rate to solve (1). Use $G(y) = \frac{d}{dy}(y - x_i)^2$ for a randomly chosen $i \in \{1 \dots n\}$. Create a plot of error (relative to y^*) versus iteration number for two different learning rates. Make sure your plot clearly shows that SGD with the larger learning rate leads to faster initial convergence and a larger terminal error than SGD with the smaller learning rate. Use a log plot in order to show the initially exponential decay of error.

Response: