MATH 462 HW 4
VERSION November 30, 2022

ADAM M. OBERMAN

Instructions. Refer to notes and references on `https://adam-oberman.github.io/Math462/`. Submit your solutions on MyCourses course page. Math exercises should be *handwritten*. You can get help from other students, but you should do the write up yourself. Coding exercises: export a PDF of the plots required.

3.1. **Classification Implementation.** Use the code provided at `https://colab.research.google.com/drive/1i9ep4yBvjAZwFOcO337w7weKhTlhfBQ6?usp=sharing`.

Hint: see the better plots on the scikitliearn page.

Exercise 3.1. *Run the classification code provided to answer the following questions.*

(a) Plot the results using values of the dataset noise levels `nmoon = ncircle` *set to 0.01, and set to 0.1.*

(b) Give the decision tree accuracy on each dataset, as a function of the `max depth`*, for values 1, 2, 3, 4, 5.*

(c) Fix the `noise level = .001`*. On which datasets does the linear classifier fail or succeed? Explain why.*

3.2. **Gradient Descent and SGD.** Define Gradient Descent by

$$w_{t+1} = w_t - \alpha \nabla_w \widehat{L}(w_t), \tag{GD}$$

Exercise 3.2 (Gradient Descent and SGD Theory)**.** *Prove Theorem 3.1.*

Theorem 3.1 (Gradient descent decreases the loss)**.** *Consider $\widehat{L} : \mathbb{R} \to \mathbb{R}$, a twice differentiable loss function, with $|\widehat{L}''(w)| \leq C_L$, for all w. Choose $0 < \alpha \leq \frac{2}{C_L}$. For any $w_t \in \mathbb{R}$, define w_{t+1} by* (GD)*. Then the loss decreases, $\widehat{L}(w_{t+1}) \leq \widehat{L}(w_t)$.*

Exercise 3.3 (Gradient Descent and SGD Implementation)**.** *You may use the code provided as a starting point, or write your own.*

https://colab.research.google.com/drive/1-YoLDf3OyH3SxLJtC5W4qG3L1zYxkyMf?usp=sharing

Consider the model problem

$$\widehat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \frac{(w - y_i)^2}{2}$$

for $w \in \mathbb{R}$, where $m = 500$ and y_i are uniformly generated over $[-1, 1]$.

(a) Run Gradient Descent (GD) *on the model problem with $\alpha = .95, .75, .5, .25, .1$ and $\alpha = .1$. Plot the loss on a log-plot (with the y-axis scaled logarithmically), so that the slope shows the rate of convergence. What is the (approximate) rate of convergence as a function of α? the x-axis should be the iteration count, and the y-axis should be the $\log$ of the error, see sample below.*

Date: November 30, 2022.

(b) *Run SGD several times, using four combinations of batch size (try: $10, 50, 100, 200$) three combinations of and constant step sizes of (try: $\alpha = .1, .02, .004$). You may need to adjust these values to get illustrative results. You should observe that the loss decreases quickly for the first few steps, then stops decreasing. Determine the approximate step number t and value of the loss in each case.* *see the figure at `https://en.wikipedia.org/wiki/Stochastic_gradient_descent` for an example plot*

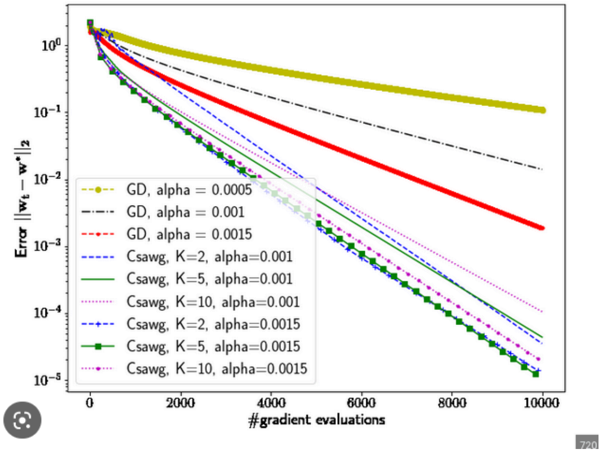

FIGURE 1. Sample convergence rate plot for gradient descen

3.3. **Exercises: Classification Theory.** Consider the empirical loss

$$\widehat{L}(h_w) = \frac{1}{m}\sum_{i=1}^{m} \ell(h_w(x_i), y_i),$$

on a dataset with with $S^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$, and $y_i \in \{0, 1\}$. Suppose that there are n_1 points with $y = 1$, and n_0 points with $y = 0$, where $n_0 + n_1 = m$. Take h_w to be the constant function $h_w(x) = w$ where $w \in [0, 1]$. Define the log loss

$$\ell_{\log}(p, y) = \begin{cases} -\log p, & y = 1 \\ -\log(1-p), & y = 0 \end{cases}$$

Exercise 3.4. *(a) Using the constant model h_w as above, simplify the sum in the empirical loss.*

(b) *Setting $\ell(w, y) = (w - y)^2$, find the value of w that minimizes the empirical loss. Express the value in terms of n_0, n_1.*

(c) *Setting $\ell(w, y) = \ell_{\log}(w, y)$, find the value of w that minimizes the empirical loss. Express the value in terms of n_0, n_1.*

(d) *Setting $\ell(w, y) = |w - y|$, what happens if we try to minimize the empirical loss over $w \in [0, 1]$? Answer in terms of the cases $n_0 > n_1, n_0 = n_1, n_0 < n_1$. Is the loss strictly convex? Does this convexity property of the loss explain anything?*

3.4. **Exercises: Features.** Consider the dataset in two dimensions given by

$$S^4 = \{(e_1, +1), (e_2, -1), (-e_1, +1), (-e_2, -1)\}, \quad e_1 = [1, 0]^\top, e_2 = [0, 1]^\top$$

Consider the polynomial features given by

$$f(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]^\top \in \mathbb{R}^6$$

along with the linear function $h_w(x) = w \cdot f(x)$, for $w \in \mathbb{R}^6$, along with the threshold classifier $c(x) = \mathrm{sign}(h_w(x))$.

Exercise 3.5. *(a) Show that S^4 can be classified with zero error (interpolated), using a function of the form $h_w(x) = w \cdot f(x)$. Provide a simple value for w that works.*

(b) Build the 4×4 feature similarity matrix $K_{ij} = f(x_i) \cdot f(x_j)$. Is K full rank?

(c) Show that the feature $f_5(x) = x_1 x_2$ is zero for every point in the dataset. When $w_5 = 1$ what does the model predict on the diagonals $x_1 = x_2$ and $x_1 = -x2$? Discuss why setting $w_5 = 0$ should generalize better than letting w_5 be non-zero.

3.5. **(PAC Learning bounds: skip this).**