# CS480/680: Introduction to Machine Learning

Homework 2

Due: 11:59 pm, May 25, 2020, submit on LEARN.

Include your name and student number!

Submit your writeup in pdf and all source code in a zip file (with proper documentation). Write a script for each programming exercise so that the TAs can easily run and verify your results. Make sure your code runs!
[Text in square brackets are hints that can be ignored.]

**Exercise 1: Ridge Regression Implementation (5 pts)**

Recall that ridge regression refers to

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \overbrace{\underbrace{\tfrac{1}{2n}\|X\mathbf{w}+b\mathbf{1}-\mathbf{y}\|_2^2}_{\text{error}}+\lambda\|\mathbf{w}\|_2^2}^{\text{loss}}, \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$ are the given dataset and $\lambda > 0$ is the regularization hyperparameter.

1. (1 pt) Show that the derivatives are

$$\frac{\partial}{\partial \mathbf{w}} = \tfrac{1}{n}X^\top(X\mathbf{w}+b\mathbf{1}-\mathbf{y}) + 2\lambda\mathbf{w} \tag{2}$$

$$\frac{\partial}{\partial b} = \tfrac{1}{n}\mathbf{1}^\top(X\mathbf{w}+b\mathbf{1}-\mathbf{y}). \tag{3}$$

2. (2 pts) Implement the gradient descent algorithm for solving ridge regression. The following incomplete pseudo-code may of help.

   Test your implementation on the Boston housing dataset (to predict the median house price, i.e., $y$). Train and test splits are provided on course website. Try $\lambda \in \{0, 10\}$ and report your training error, training loss and test error. [Your training loss should monotonically decrease during iteration; if not try to tune your step size $\eta$, e.g. make it smaller.]

**Algorithm 1:** Gradient descent for ridge regression.

**Input:** $X \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{w}_0 = \mathbf{0}_d$, $b_0 = 0$, $\mathsf{max\_pass} \in \mathbb{N}$, $\eta > 0$, $\mathsf{tol} > 0$

**Output:** $\mathbf{w}, b$

1 **for** $t = 1, 2, \ldots, \mathsf{max\_pass}$ **do**

2 $\quad\mathbf{w}_t \leftarrow$

3 $\quad b_t \leftarrow$

4 $\quad$**if** $\|\mathbf{w}_t - \mathbf{w}_{t-1}\| \le \mathsf{tol}$ **then** `// can use other stopping criteria`

5 $\qquad$**break**

6 $\mathbf{w} \leftarrow \mathbf{w}_t,\ b \leftarrow b_t$

3. (1 pt) We note that given $\mathbf{w}$, we can actually solve $b$ by setting the derivative (3) to 0. Re-implement Line 3 in Algorithm 1 with this closed-form solution. Does the modification converge to the same solution?

4. (1 pt) If we center our data beforehand, i.e., by subtracting their mean we get $X^\top\mathbf{1} = \mathbf{0}$ and $\mathbf{1}^\top\mathbf{y} = 0$. What is the optimal value of $b$ in this case? [You may verify your result by running your code above.]

**Exercise 2: Margin (5 pts)**

1. (2 pts) Recall that a hyperplane is parameterized by its normal vector $\mathbf{w} \in \mathbb{R}^d$ and offset $b \in \mathbb{R}$:

$$\partial H_{\mathbf{w},b} := \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x}\rangle + b = 0\} \tag{4}$$

Compute the distance from a given point $\mathbf{z} \in \mathbb{R}^d$ to the hyperplane $\partial H_{\mathbf{w},b}$:

$$\min_{\mathbf{x} \in \partial H_{\mathbf{w},b}} \ \tfrac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2. \tag{5}$$

[Hint: (a) $\mathbf{z} = \mathbf{z}^\perp + \mathbf{z}^\parallel$ (orthogonal and parallel to $\mathbf{w}$); or (b) derive and solve the Lagrangian dual.]

Ans:

2. (1 pt) Compute the distance from $\mathbf{z}$ to the halfspace:

$$H_{\mathbf{w},b} := \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x} \rangle + b \leq 0\}. \tag{6}$$

Ans:

3. (2 pts) Consider a binary dataset consisting of two data points $(\mathbf{x}_1, y_1 = 1)$ and $(\mathbf{x}_2, y_2 = -1)$. Compute its margin (i.e. among all *separating* hyperplanes, the largest minimum distance to all data points). You need to justify your solution.

Ans: