CS480/680: Introduction to Machine Learning

Homework 1

Due: 11:59 pm, January 28, 2021, submit on Crowdmark (yet to be set up, stay tuned).

Last Updated: January 17, 2021

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: Perceptron Implementation (5 pts)

Convention: All algebraic operations, when applied to a vector or matrix, are understood to be element-wise (unless otherwise stated).

Algorithm 1: The perceptron algorithm.

```
Input: $X \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \{-1, 1\}^n$, $\mathbf{w} = \mathbf{0}_d$, $b = 0$, $\mathsf{max\_pass} \in \mathbb{N}$
Output: $\mathbf{w}, b, mistake$
for $t = 1, 2, \ldots, \mathsf{max\_pass}$ do
    $mistake(t) \leftarrow 0$
    for $i = 1, 2, \ldots, n$ do
        if $y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \le 0$ then
            $\mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$                // $\mathbf{x}_i$ is the $i$-th row of $X$
            $b \leftarrow b + y_i$
            $mistake(t) \leftarrow mistake(t) + 1$
```

Implement the perceptron in Algorithm 1. Your implementation should take input as $X = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top]^\top \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \{-1, 1\}^n$, an initialization of the hyperplane parameters $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, and the maximum number of passes of the training set [suggested $\mathsf{max\_pass} = 500$]. Run your perceptron algorithm on the spambase dataset (use the version on the course website), and plot the number of mistakes (y-axis) w.r.t. the number of passes (x-axis).

Ans:

Exercise 2: Perceptron Questions (5 pts)

1. (3 pts) The perceptron algorithm makes an update every time it witnesses a mistake. What if it makes an update on every point, regardless of whether or not that point is correctly classified? For simplicity, consider a setting where where b is fixed to 0. Give an example of an infinite sequence of points (x_i, y_i) with the following properties:

   - The sequence is strictly linearly separable with $b = 0$ (i.e., the margin is some constant $\gamma > 0$),
   - The sequence has max $\|x_i\|_2 \le 1$,
   - The modified perceptron algorithm makes an infinite number of mistakes on this sequence.

   Prove that it has these properties. Note that the perceptron convergence theorem and the first two conditions imply that, at some point, the unmodified perceptron algorithm would only make a finite number of mistakes.

   Ans:

2. (1 pt) Give examples of where the perceptron algorithm converges to a 0 margin halfspace, and a maximum margin halfspace.

   Ans:

3. (1 pt) Suppose that in each iteration of the perceptron algorithm, a point is chosen uniformly at random from the dataset. Show how the perceptron algorithm can be viewed as an instantiation of stochastic gradient descent (SGD). In particular, you must provide a loss function and a learning rate such that SGD with these parameters and perceptron are identical.

Ans:

Exercise 3: Regression Implementation (8 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

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

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

   Ans:

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). Use the train and test splits 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 η, e.g. make it smaller.]

Algorithm 2: Gradient descent for ridge regression.

```
Input: $X \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{w}_0 = \mathbf{0}_d$, $b_0 = 0$, max_pass $\in \mathbb{N}$, $\eta > 0$, tol $> 0$
Output: $\mathbf{w}, b$
for $t = 1, 2, \ldots,$ max_pass do
    $\mathbf{w}_t \leftarrow$
    $b_t \leftarrow$
    if $\|\mathbf{w}_t - \mathbf{w}_{t-1}\| \le$ tol then          // can use other stopping criteria
        break
$\mathbf{w} \leftarrow \mathbf{w}_t,\ b \leftarrow b_t$
```

   Ans:

   For the next part, you may use the Python package scikit-learn.

3. (5 pts) Train (unregularized) linear regression, ridge regression, and lasso on the mystery datasets A, B, and C on the course website (using X_train and Y_train for each dataset). For ridge regression and lasso, use parameters 1 and 10 (note that you will have to divide these by n for lasso in scikit-learn – why?). Report the average mean squared error on the test set for each method. Which approach performs best in each case? Plot the five parameter vectors obtained for each dataset on the same histogram, so they're all visible at once (change the opacity setting for the bars if necessary): specifically, for each parameter vector, plot a histogram of its value in each coordinate. Given which approach performs best, and how its parameter histogram looks, how do you suspect the true parameter vector and responses might have been generated?

   Ans:

Exercise 4: An Alternative to Least Squares? (3 pts)

Suppose we are in a setting with a dataset $X \in \mathbb{R}^{n\times d}$, and labels are generated according to $Y = XW + \varepsilon$, where $W \in \mathbb{R}^d$, $Y \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^n$ is a random vector, where all entries are independent random variables with 0 mean and variance σ^2 (you can imagine Gaussian, but it isn't necessary). As we saw in class, the least-squares solution $\hat{W}$ can be written as $\hat{W} = (X^T X)^{-1} X^T Y$ – this is a linear transformation of the response vector Y. Consider some *different* linear transformation $\left((X^T X)^{-1} X^T + N\right) Y$, where $N \in \mathbb{R}^{d \times n}$ is a non-zero matrix.

1. (1 pt) Show that the expected value of this linear transformation is $(I_d + NX)W$. Conclude that its expected value is W if and only if $NX = 0$. (1 pt)

   Ans:

2. (2 pts) Compute the covariance matrix of this linear transformation when $NX = 0$, and show that it is equal to $\sigma^2 (X^T X)^{-1} + \sigma^2 N N^T$. Since the former term is the covariance of the least squares solution[a] and the latter matrix is positive semi-definite, this implies that this alternative estimator only increased the variance of our estimate.

   Ans:

[a] Verify this for yourself, but no need to submit it.

Exercise 5: Sample Statistics (2 pts)

1. (1 pt) Suppose there is a dataset $x_1, \ldots, x_n$ sampled from a distribution with mean μ and variance σ^2. Compute the expected value of the sample mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Describe any modifications that might be required to make the expected value μ (recall that μ and σ^2 are unknown).

   Ans:

2. (1 pt) Suppose there is a dataset $x_1, \ldots, x_n$ sampled from a distribution with mean μ and variance σ^2. Compute the expected value of the sample variance: $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, where $\bar{x}$ is the sample mean from the previous part. Describe any modifications that might be required to make the expected value σ^2 (recall that μ and σ^2 are unknown).

   Ans: