Midterm Exam Solutions

1. **(3 points)** Please write down the time at the *start* and *end* of your exam. The difference should not exceed 90 minutes. Please also write down your *name* and *signature* below; by doing so, you are affirming the NYU Tandon School of Engineering student code of conduct.

Solution N/A

2. **(10 points)** This is a slight variant of a homework problem. Let $\{x_1, x_2, \dots, x_n\}$ be a set of points in d-dimensional space, and let $\{p_1, p_2, \dots, p_n\}$ denote a probability distribution over the integers $[1, 2, \dots, n]$. Suppose we wish to produce a single point estimate $\mu \in \mathbb{R}^d$ that minimizes the *weighted* squared-error:

$$L(\mu) = p_1\|x_1 - \mu\|_2^2 + p_2\|x_2 - \mu\|_2^2 + \ldots + p_n\|x_n - \mu\|_2^2$$

Find a closed form expression for the optimal μ and prove that your answer is correct.

Solution

Take the gradient with respect to μ and set to zero. The minimum mean-squared error estimate is

$$\mu^* = \sum_{i=1}^{n} p_i x_i.$$

(The denominator disappears since p is a probability distribution and hence sums to one).

3. **(15 points)** This is a variant of a homework problem. Suppose x is a d-dimensional input, w is a d-dimensional variable, and λ is a regularization parameter.

   a. Show that the minimizer of the squared-error loss with ℓ_1 regularizer:

   $$L(w) = \frac{1}{2}\|x - w\|_2^2 + \lambda\|w\|_1$$

   is given by:

   $$w_i^* = \begin{cases} x_i - \lambda & \text{if} \quad x_i > \lambda, \\ x_i + \lambda & \text{if} \quad x_i < -\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

   b. Show that the minimizer of the squared-error loss with ℓ_2 regularizer:

   $$L(w) = \frac{1}{2}\|x - w\|_2^2 + \lambda\|w\|_2^2$$

   is given by:

   $$w_i^* = \left(\frac{1}{1 + 2\lambda}\right) x_i.$$

   c. In class, we argued via contour plots that greater ℓ_2 regularization encourages "small" solutions, while greater ℓ_1 regularization encourages "sparse" solutions. Mathematically justify why that is the case by examining the structure of the optimal solutions derived above.

Solution

a. This is from homework.

b. This can be done by taking gradients.

c. In the ℓ_1 case, all coefficients with magnitude below λ are zeroed out, and hence higher λ encourages solutions with more zeros (or sparser solutions). In the ℓ_2 case, all coefficients are scaled inversely with respect to λ, and hence higher lambda encourages smaller solutions.

4. **(10 points)** The following represents python code for an algorithm that attempts to perform linear regression. (a) Identify the algorithm. (b) Explain why this algorithm may not converge as implemented below, and identify the line in the algorithm that makes this happen. (c) Suggest a way to fix this algorithm.

```
def optim_alg(init, steps, grad):
  xs = [init]
  for step in steps:
    xs.append(xs[-1] - step * grad(xs[-1]))
  return xs

def linear_reg_grad(X,y,w):
  return X.T.dot(X.dot(w)-y)

input_to_optim_alg = lambda w: linear_reg_grad(X,y,w)
learning_rates = np.arange(start=1, stop=300, step=3)
ws = optim_alg(w0,learning_rates,input_to_optim_alg)
```

Solution

This is regular gradient descent. However, the learning rates are increasing (and this may lead to instabilities). Convergence can be assured by choosing a small, low learning rate (or sequence of decreasing learning rates).

5. **(10 points)** To combat the COVID-19 pandemic, an enterprising NYU Tandon graduate student decides to build a logistic regression model to predict the conditional likelihood of a person being one of two states – *infected* or *clear* – based on daily forehead temperature measurements over the last 30 days. Fortunately, a dataset of such measurements for a population of 100,000 persons is available.

   a. Identify the parameters of the problem (number of samples n, data dimension d, number of classes k.)

   b. If X and y denote the arrays that encode the training data points and labels, what are the sizes of X and y?

   c. Starting from the definition of conditional likelihood, derive the loss function used to train the model. You can assume the probabilities can be modeled as a sigmoid function.

Solution

a. $n = 100,000$, $d = 30$, $k = 2$.

b. Dimension of X is $n \times d$ or $n \times (d+1)$ depending on whether bias is modeled or not. Dimension of y is $n \times 1$ (this is a binary classification so no need to have one-hot encoding).

c. From notes.

6. **(15 points)** Suppose we are given real-valued scalar data (i.e., $d = 1$) belonging to one of two classes. We are given a set of three data samples with negative labels, $X_- = \{0, 1, -1\}$, and a set of three data samples with positive labels, $X_+ = \{-3, 3, -2\}$. Our goal is to build a classifier for this dataset. We will show that kernel methods are particularly useful in this case.

   a. Argue that no perfect linear separator in the original space can exist.

   b. Argue that if the data is mapped via the two-dimensional feature mapping $\phi(u) = (u, u^2)$, then a perfect linear separator exists.

   c. Explicitly draw the maximum-margin linear separator in the new feature space, and mark the closest points nearest to this linear separator.

   d. Calculate the equation of the maximum-margin separator.

Solution

a. Any separator in one dimension, $d = 1$, can be uniquely specified by a scalar $a \in \mathbb{R}$. There are exactly 7 disjoint possibilities where a can occur on the real line: $3 \geq a$, $1 \leq a < 3$, $0 \leq a < 1$, $-1 \leq a < 0$, $-2 \leq a < -1$, and $-3 \leq a < -2$, and $a < -3$. By inspection one can see that each of the above choices makes at least one error while labeling the test data in X_+ and X_-.

b. Clear from the figure. (Green circles = '+', red triangles = '-')

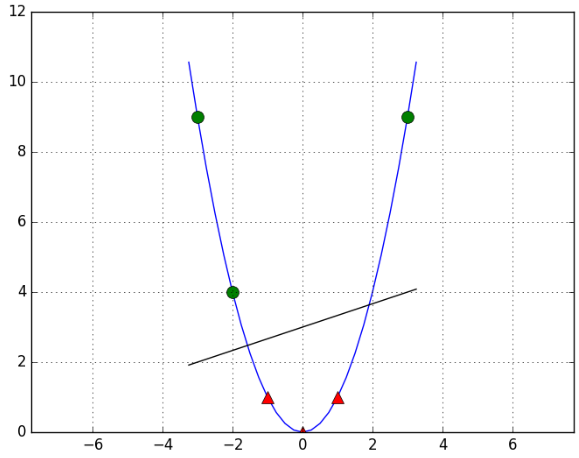

Figure 1:

c. (and d) First, observe that the optimal linear separator must lie somewhere between $(-2, 4)$ and $(-1, 1)$; everything above this line will be labeled $+$, while everything below will be labeled $-$. For the separator to be *optimal*, this line has to be as far away as possible from these two points, i.e., it has to be the *perpendicular bisector* between these points. In other words, it has to pass through the mid-point:

$$\frac{(-2,4)+(-1,1)}{2}=(\frac{-3}{2},\frac{5}{2})$$

Moreover, it should have slope 1/3, since the line joining the two points has slope -3. Therefore, the equation to the separator is given by:

$$x_2-\frac{5}{2}=\frac{1}{3}(x_1+\frac{3}{2})\rightarrow 3x_2=x_1+9.$$

Value of margin is $\sqrt{10}/2\approx 1.58$ (no need to calculate this).