CS-GY 6923: Assignment 1

September 12, 2015

COLLABORATION POLICY: You may discuss general concepts relating to the homework questions with other students, but you must write up your solutions on your own, in your own words.

Hard Deadline: Assigment 1 is due on September 25.

1 Linear Classification (10 points)

Show that the logistic (sigmoid) function

$$\delta(a) = \frac{1}{1 + e^{-a}}$$

satisfies the property $\delta(-a) = 1 - \delta(a)$ and that its inverse is given by $\delta^{-1}(y) = \ln \frac{y}{1-y}$.

2 Linear Regression (15 points)

Suppose the samples are $x_1, x_2, \ldots, x_n \in R^d$, and $X = [x_1, x_2, \ldots, x_n] \in R^{d \times n}$, and the corresponding output variables are $y = [y_1, y_2, \ldots, y_n] \in R^{1 \times n}$. We would like to learn a vector of coefficient $w \in R^d$. Using the technique of Lagrange multipliers, show that minimization of the regularized error function

$$w = \arg \min_{w \in R^d} (\|w^T X - y\|_2^2 + \lambda \|w\|_q^q)$$

is equivalent to minimizing the unregularized sum-of-squares error

$$w = \arg \min_{w \in R^d} \|w^T X - y\|_2^2$$

subject to the constraint

$$\|w\|_q^q \leq \eta$$

Discuss the relationship between the parameters η and λ.

3 Linear Regression (15 points)

In linear regression suppose your samples are $x_1, x_2, \ldots, x_n \in R^d$, and $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$. The corresponding output variables are $y = [y_1, y_2, \ldots, y_n] \in R^{1\times n}$. We would like to learn a vector of coefficient $w \in R^d$, such that

$$w = \arg\min_{w\in R^d} \|w^T X - y\|_2^2$$

As in our discussion, in the end it reduces to solving a group of linear equations.

$$XX^T w = Xy^T.$$

Prove if $n < d$, the matrix XX^T is not invertible.

4 Ridge Regression (20 points)

In ridge regression, suppose your samples are $x_1, x_2, \ldots, x_n \in R^d$, and $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$. The corresponding output variables are $y = [y_1, y_2, \ldots, y_n] \in R^{1\times n}$. We would like to learn a vector of coefficient $w \in R^d$, such that

$$w = \arg\min_{w\in R^d} (\|w^T X - y\|_2^2 + \lambda\|w\|_2^2)$$

I mentioned in the lecture in the end it boils down to solving a group of linear equations.

$$(XX^T + \lambda I_d)w = Xy^T.$$

1. Prove the above statement.
2. Prove the matrix $XX^T + \lambda I_d$ is invertible for $\lambda > 0$.

5 GD and SGD for Ridge Regression (20 points)

If you would like to use Gradient Descent or Stochastic Gradient Descent to solve the ridge regression problem, what is your algorithm? (Please state your algorithm for both GD and SGD).

6 MAP and Lasso (20 points)

In lecture 2 we discussed the connection between the Maximum A Posterior Estimation (MAP) and the ridge regression. Can you come up with a prior of w for Lasso using the Bayesian framework?

$$w = \arg\min_{w\in R^d} (\|w^T X - y\|_2^2 + \lambda\|w\|_1)$$

7 Bonus Points

7.1 Growth Rate of Functions(5 points)

First read the wikipedia about growth rate of functions `http://en.wikipedia.org/wiki/Big_O_notation`. Indicate, for each example, whether $f = O(g)$, $f = \Omega(g)$ or $f = \Theta(g)$. No justification is necessary.

1. $f = 5x^3 + 3x^2, g = 14x^3$.
2. $f = n2^n, g = 3n$
3. $f = \log(n^3), g = 10 \log n$
4. $f = n^{20}, g = n^2$
5. $f = n^3, g = 4 \log_2 n$
6. $f = n!, g = 3n$
7. $f = 1.0005^n, g = n^{1.005}$
8. $f = n^2 + 5 \log n, g = 3n^2 + 0.5n$
9. $f = n \log n, g = n + \log n$

7.2 Posterior Probability(5 points)

Suppose you have two coins. The first coin, has probability $= 1/2$ of heads, and the second, has probability $= 1/3$ of heads. Consider the following random process:

1. You randomly choose one of the two coins, with equal probability of choosing either one. You throw away the other coin.
2. You toss the chosen coin 3 times in a row.

Now you observe the outcome of the 3 tosses were Heads, Heads, Tails, what is the probability that the coin you chose is the first coin? (assume the outcomes are independent, given that you've chosen the coin.)