CS 5350/6350, DS 4350: Machine Learning Fall 2024

Homework 6

April 23, 2025

Jordy A. Larrea Rodriguez

1 Logistic Regression

We saw Maximum A Posteriori (MAP) learning of the logisitic regression classifier in class. In particular, we showed that learning the classifier is equivalent to the following optimization problem:

$$\min_{\mathbf{w}} \sum_{i=1}^{m} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w}$$

In this question, you will derive the stochastic gradient descent algorithm for the logistic regression classifier.

1. [5 points] What is the derivative of the function $g(\mathbf{w}) = \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$ with respect to the weight vector? Your answer should be a vector whose dimensionality is the same as $\mathbf{w}$.

   Response: $\frac{\partial g(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i))$

   Let $z = -y_i \mathbf{w}^T \mathbf{x}_i$ and $u = 1 + exp(z)$.

   So, $\frac{\partial g(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} log(u)$

   $= \frac{1}{u} \frac{\partial}{\partial \mathbf{w}} (1 + exp(z))$

   $= \frac{-y_i \mathbf{x}_i exp(z)}{1+exp(z)}$

2. [5 points] The inner most step in the SGD algorihtm is the gradient update where we use a single randomly chosen example instead of the entire dataset to compute a stochastic estimate of the gradient. Write down the objective where the entire dataset is composed of a single example, say $(\mathbf{x}_i, y_i)$.

   Response: For a single example $(\mathbf{x}_i, y_i)$, the objective function is the negative logarithm of the likelihood function.

$$J(\mathbf{w}) = -\text{Ł}(\mathbf{w}) = -(y_i log(P(y_i = +1|\mathbf{x}_i; \mathbf{w})) + (1 - y_i) log(1 - P(y_i = +1|\mathbf{x}_i; \mathbf{w})))$$

Where, $P(y_i = +1|\mathbf{x}_i; \mathbf{w}) = \frac{1}{1+exp(z)}$

3. [5 points] Derive the gradient of the SGD objective for a single example (from the previous question) with respect to the weight vector.

   Response: $\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = -\frac{\partial}{\partial \mathbf{w}}(y_i log(P(y_i = +1|\mathbf{x}_i;\mathbf{w})) + (1-y_i)log(1-P(y_i = +1|\mathbf{x}_i;\mathbf{w})))$
   $$= \mathbf{x}_i(y_i - P(y_i = +1|\mathbf{x}_i;\mathbf{w}))$$

4. [15 points] Write down the pseudo code for the stochastic gradient algorithm using the gradient from previous part.

   Hint: The answer to this question will be an algorithm that is similar to the SGD based learner we developed in the class for SVMs.

   Response:

   Given a training set $S = (\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \Re^d$, $y \in \{-1, +1\}$

   1. Init $\mathbf{w} = 0$
   2. For epoch = 1...T:
      1. Get rand example $(\mathbf{x}_i, y_i)$ in S
      2. Treat $(\mathbf{x}_i, y_i)$ as a full dataset and take derivative of objective $J(\mathbf{w})$ at $\mathbf{w}^{t-1}$
      3. Update $\mathbf{w}^t = \mathbf{w}^{t-1} - \gamma \Delta J(\mathbf{w}^{t-1})$
   3. Return $\mathbf{w}$