Lecturer: Jie Wang   Homework 4
Posted: Apr. 5, 2020   Due: Apr. 12, 2020
Name: San Zhang   ID: PBXXXXXXXX

Notice, to get the full credits, please show your solutions step by step.

Exercise 1: Logistic Regression 40pts

Given the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Let

$$\mathcal{I}^+ = \{i : i \in [n], y_i = 1\},$$
$$\mathcal{I}^- = \{i : i \in [n], y_i = 0\},$$

where $[n] = \{1, 2, \ldots, n\}$. We assume that $\mathcal{I}^+$ and $\mathcal{I}^-$ are not empty.
    Then, we can formulate the logistic regression as:

$$\min_{\mathbf{w}} \; L(\mathbf{w}) = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \log \left(\frac{\exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle)}{1 + \exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle)} \right) + (1 - y_i) \log \left(\frac{1}{1 + \exp(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle)} \right) \right), \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^{d+1}$ is the model parameter to be estimated and $\overline{\mathbf{x}}_i^\top = (1, \mathbf{x}_i^\top)$.

1. (5pts) Find the gradient and the Hessian of $L(\mathbf{w})$.

2. (10pts) Suppose that $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_n)^\top \in \mathbb{R}^{n \times (d+1)}$ and $\mathbf{rank}(\overline{\mathbf{X}}) = d + 1$. Show that $L(\mathbf{w})$ is strictly convex, i.e., for all $\mathbf{w}_1 \neq \mathbf{w}_2$,

$$L(t\mathbf{w}_1 + (1 - t)\mathbf{w}_2) < tL(\mathbf{w}_1) + (1 - t)L(\mathbf{w}_2), \forall \, t \in (0, 1).$$

3. (10pts) Suppose that the training data is strictly linearly separable, that is, there exists $\hat{\mathbf{w}} \in \mathbb{R}^{d+1}$ such that

$$\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i \rangle > 0, \; \forall \, i \in \mathcal{I}^+,$$
$$\langle \hat{\mathbf{w}}, \overline{\mathbf{x}}_i \rangle < 0, \; \forall \, i \in \mathcal{I}^-.$$

    Show that problem (1) has no solution.

Let $\overline{\mathbf{z}}_i = (2y_i - 1)\overline{\mathbf{x}}_i$.

4. (5pts) Show that

$$L(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \mathbf{w}, \overline{\mathbf{z}}_i \rangle)).$$

5. (20pts) Suppose that the training data is NOT linearly separable. That is, for all $\mathbf{w} \in \mathbb{R}^{d+1}$, there exists $i \in [n]$ such that

$$\langle \mathbf{w}, \bar{\mathbf{z}}_i \rangle < 0.$$

Show that problem (1) always admits a solution.

Solution: ∎

Exercise 2: Programming Exercise: Naive Bayes 20pts

We provide you with a data set that contains spam and non-spam emails ("hw4_nb.zip"). Please use the Naive Bayes Classifier to detect the spam emails. Finish the following exercises by programming. You can use your favorite programming language.

1. Remove all the tokens that contain non-alphabetic characters.

2. Train the Naive Bayes Classifier on the training set according to Algorithm 1.

3. Test the Naive Bayes Classifier on the test set according to Algorithm 2.

4. Compute the confusion matrix, precision, recall, and F1 score. Please report your result.

Algorithm 1 Training Naive Bayes Classifier

Input: The training set with the labels $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$.

1:   $\mathcal{V} \leftarrow$ the set of distinct words and other tokens found in $\mathcal{D}$
2:   **for** each target value c in the labels set $\mathcal{C}$ **do**
3:     $\mathcal{D}_c \leftarrow$ the training samples whose labels are c
4:     $P(c) \leftarrow \frac{|\mathcal{D}_c|}{|\mathcal{D}|}$
5:     $T_c \leftarrow$ a single document by concatenating all training samples in $\mathcal{D}_c$
6:     $n_c \leftarrow |T_c|$
7:     **for** each word w_k in the vocabulary $\mathcal{V}$ **do**
8:       $n_{c,k} \leftarrow$ the number of times the word w_k occurs in T_c
9:       $P(w_k|c) = \frac{n_{c,k}+1}{n_c+|\mathcal{V}|}$
10:    **end for**
11: **end for**

Algorithm 2 Testing Naive Bayes Classifier

Input: An email $\mathbf{x}$. Let x_i be the ith token in $\mathbf{x}$. $\mathcal{I} = \emptyset$.

1:   **for** $i = 1, \ldots, |\mathbf{x}|$ **do**
2:     **if** $\exists w_{k_i} \in \mathcal{V}$ such that $w_{k_i} = x_i$ **then**
3:       $\mathcal{I} \leftarrow \mathcal{I} \cup k_i$
4:     **end if**
5:   **end for**
6:   predict the label of $\mathbf{x}$ by

$$\hat{y} = \arg\max_{c \in \mathcal{C}} P(c) \prod_{i \in \mathcal{I}} P(w_{k_i}|c)$$

Solution: ■

Exercise 3: Example in Stochastic Gradient Descent 15pts

Consider a simple linear regression model $f(x; w) = wx$ with samples $\{(x_i, y_i)\}_{i=1}^2$, where $x_i, y_i \in \mathbb{R}$ for $i = 1, 2$. We use SGD algorithm to minimize the average fitting error

$$L(w) = \frac{1}{2} \sum_{i=1}^{2} (y_i - wx_i)^2.$$

We uniformly sample a data instance $\xi_k = (x_{i_k}, y_{i_k})$ from $\{(x_i, y_i)\}_{i=1}^2$ at k^{th} iteration. The sequence (w_k) is generated by the stochastic gradient descent algorithm.

1. (5pts) Please write down the derivative $L'(w_k)$ and stochastic derivative g_k.

2. (5pts) Please write down the variance of the stochastic derivative $\mathbb{V}_{\xi_k}[g_k]$.

3. (5pts) We assume the upper bound of $\mathbb{V}_{\xi_k}[g_k]$ takes the form of

$$\mathbb{V}_{\xi_k}[g_k] \leq M + M_V |L'(w_k)|^2.$$

   Please find the corresponding M and M_V in this problem. Specifically, when will M become zero, and when will M_V become zero?

Solution: ■

Exercise 4: Convergence of SGD for Convex Function 25pts

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable, and it attains its minimum at $\mathbf{x}^*$. Suppose the second moment of stochastic gradient $\mathbf{g}$ is bounded, i.e.,

$$\mathbb{E}_\xi[\|\mathbf{g}(\mathbf{x}, \xi)\|_2^2] \leq G^2, \ \forall \mathbf{x} \in \mathbb{R}^n.$$

Suppose $(\mathbf{x}_k)$ is a sequence generated by SGD algorithm with a fixed stepsize α. Define $\tilde{\mathbf{x}}_K = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_k$ and $f^* = f(\mathbf{x}^*)$.

1. Let $X, Y,$ and Z be random variables.

    (a) (5pts) Show that the tower property holds, i.e.,

    $$\mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}[X|Y, Z]|Y].$$

    (b) (5pts) The variance decomposition formula holds, i.e.,

    $$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X|Y]] + \mathbb{V}[\mathbb{E}[X|Y]].$$

    Hint: if you do not know measure theory well, you can assume that X, Y, and Z are continuous random variables.

2. (5pts) Suppose the stochastic gradient at k^{th} iteration is $\mathbf{g}_k$. Please show that

$$\mathbb{E}_{\xi_1:\xi_k}[f(\mathbf{x}_k) - f^*] \leq \mathbb{E}_{\xi_1:\xi_k}[\langle \mathbf{g}_k, \mathbf{x}_k - \mathbf{x}^* \rangle].$$

3. (5pts) Please show that

$$\mathbb{E}_{\xi_1:\xi_k}[f(\mathbf{x}_k) - f^*] \leq \frac{1}{2\alpha} \mathbb{E}_{\xi_1:\xi_k}[\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 + \alpha^2 \|\mathbf{g}_k\|_2^2]. \qquad (2)$$

4. (5pts) Please show that

$$\mathbb{E}_{\xi_1:\xi_K}[f(\tilde{x}_K) - f^*] \leq \frac{\|x_1 - x^*\|_2^2 + \alpha^2 G^2 K}{2\alpha K}$$
$$\xrightarrow{O(1/K)} \frac{\alpha G^2}{2}.$$

Solution: ∎

Exercise 5: Programming Exercise: Logistic Regression 20pts

We provide you with a dataset of handwritten digits[1] that contains a training set of 60000 examples and a test set of 1960 examples ("hw4_lr.zip"). Each image in this dataset has 28×28 pixels and the associated label is the handwritten digit—that is, an integer from the set $\{0, 1, \cdots, 9\}$—in the image. In this exercise, you need to build a logistic regression classifier to predict if a given image has the handwritten digit 0 in it or not. You can use your favorite programming language to finish this exercise.

1. (a) Choose a proper normalization method to process the data matrix. Please report the normalization method you use.

   (b) Find a Lipschitz constant of $\nabla L(\mathbf{w})$, where $L(\mathbf{w})$ is the objective function of the logistic regression after normalizing and $\mathbf{w}$ is the model parameter to be estimated. Please report your result.

2. (a) Use GD and SGD to train the logistic regression classifier on the training set, respectively. Evaluate the classification accuracy on the training set after each iteration. Stop the iteration when Accuracy $\geq 97\%$. Please plot the accuracy of these two classifiers (the one trained by GD and the other trained by SGD) versus the iteration step on one graph.

   (b) Compare the total iteration counts and the total time cost of the two methods (GD and SGD), respectively. Please report your result.

   (c) Compare the confusion matrix, precision, recall and F1 score of the two classifiers (the one trained by GD and the other trained by SGD). Please report your result.

3. (a) The training set is imbalanced as the majority class has roughly ten times more images than the minority class. Imbalanced data can hurt the performance of the classifiers badly. Thus, please undersample the majority class such that the numbers of images in the two classes are roughly the same.

   (b) Use GD to train the logistic regression classifier on the new training set after undersampling. Stop the iteration when Accuracy $\geq 97\%$.

   (c) Evaluate the two classifiers (the one trained with GD on the original training set and the other trained on the new training set after undersampling) on the test set. Compare the confusion matrix, precision, recall and F1 score of the two classifiers. Please report your result.

Solution: ∎

---

[1]This dataset is modified from the MNITS dataset: http://yann.lecun.com/exdb/mnist/