Machine Learning, 2024 Spring
Assignment 3

Notice

Plagiarizer will get 0 points.
LATEXis highly recommended. Otherwise you should write as legibly as possible.

Problem 1 For logistic regression, show that

$$\nabla E_{\text{in}}\left(\mathbf{w}\right) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{x}_n}{1+e^{y_n\mathbf{w}^{\text{T}}\mathbf{x}_n}} = \frac{1}{N}\sum_{n=1}^{N}-y_n\mathbf{x}_n\theta\left(-y_n\mathbf{w}^{\text{T}}\mathbf{x}_n\right) \tag{1}$$

Argue that a 'misclassified' example contributes more to the gradient than a correctly classified one.
Solution
1. Let $\theta(x) = \dfrac{e^x}{1+e^x} = \dfrac{1}{1+e^{-x}}$ be the sigmoid function.
So we have $p(y_n|\mathbf{x}_n;\mathbf{w}) = \theta(y_n\mathbf{w}^T\mathbf{x}_n)$, and the maximum likelihood estimation is

$$\max \prod_{i=1}^{N} p(y_n|\mathbf{x}_n;\mathbf{w})$$

For a convinient calculation, we take the log of the likelihood function, and then take the negative of it, and take the sample number N into consideration, so we have

$$
\begin{aligned}
E_{\text{in}}\left(\mathbf{w}\right) &= -\frac{1}{N}\log[\prod_{i=1}^{N}p(y_n|\mathbf{x}_n;\mathbf{w})] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\log\theta(y_n\mathbf{w}^T\mathbf{x}_n) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}} \\
&= -\frac{1}{N}\sum_{i=1}^{N}[-\log(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n})] \\
&= \frac{1}{N}\sum_{i=1}^{N}\log(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n})
\end{aligned}
$$

So the gradient of $E_{\text{in}}(\mathbf{w})$ is

$$
\begin{aligned}
\nabla E_{\text{in}}(\mathbf{w}) &= \nabla \left[\frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}) \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \nabla \log(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{(-y_n \mathbf{x}_n) e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}} \\
&= -\frac{1}{N} \sum_{i=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \\
&= \frac{1}{N} \sum_{i=1}^{N} -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n)
\end{aligned}
$$

2. From the property of the sigmoid function, we know that the sigmoid function is monotonic increasing.
So for a correctly classified example, we have $y_n \mathbf{w}^T \mathbf{x}_n > 0$, so $\theta(-y_n \mathbf{w}^T \mathbf{x}_n) \in (0, 0.5)$.

And for a misclassified example, we have $y_n \mathbf{w}^T \mathbf{x}_n < 0$, so we have $\theta(-y_n \mathbf{w}^T \mathbf{x}_n) \in (0.5, 1)$.

From the gradient formula $\nabla E_{\text{in}}(\mathbf{w})$, the difference between the correctly classified example and the misclassified example is only the sigmoid function.
And we also have the misclassified example's $\theta(-y_n \mathbf{w}^T \mathbf{x}_n)$ is larger than the correctly classified example's $\theta(-y_n \mathbf{w}^T \mathbf{x}_n)$, so the misclassified example contributes more to the gradient than a correctly classified one.

So above all, we have proved that

$$
\nabla E_{\text{in}}(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} = \frac{1}{N} \sum_{n=1}^{N} -y_n \mathbf{x}_n \theta\left(-y_n \mathbf{w}^T \mathbf{x}_n\right)
$$

And we also proved that the 'misclassified' example contributes more to the gradient than a correctly classified one.

For linear regression, the out-of-sample error is

$$E_{\text{out}}(h) = \mathbb{E}[(h(x) - y)^2]. \tag{2}$$

Show that among all hypotheses, the one that minimizes E_{out} is given by

$$h^*(x) = \mathbb{E}[y|x]. \tag{3}$$

The function h^* can be treated as a deterministic target function, in which case we can write $y = h^*(x) + \epsilon(x)$ where $\epsilon(x)$ is an (input dependent) noise variable. Show that $\epsilon(x)$ has expected value zero.

Solution

Suppose the joint distribution of x and y is $f(x, y)$, the distribution of x is $f_X(x)$, the distribution of y is $f_Y(y)$, and the conditional distribution of y given x is $f_{Y|X}(y|x)$.

Lamma: The law of iterated Expectation

$$\mathbb{E}(Y) = \mathbb{E}_X[\mathbb{E}(Y|X)]$$

proof:

$$\begin{aligned}
\mathbb{E}_X[\mathbb{E}(Y|X)] &= \mathbb{E}_X\left[\int_{-\infty}^{+\infty} y f_{Y|X}(y|x)dy\right] \\
&= \mathbb{E}_X\left[\int_{-\infty}^{+\infty} y \frac{f(x,y)}{f_X(x)}dy\right] \\
&= \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} y \frac{f(x,y)}{f_X(x)}dy\right] f_X(x)dx \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y f(x,y)dxdy \\
&= \int_{-\infty}^{+\infty} y \left[\int_{-\infty}^{+\infty} f(x,y)dx\right] dy \\
&= \int_{-\infty}^{+\infty} y f_Y(y)dy \\
&= \mathbb{E}(Y)
\end{aligned}$$

1. Since $E_{\text{out}}(h) = \mathbb{E}[(h(x) - y)^2] = \mathbb{E}_{X,Y}[(h(x) - y)^2]$, we have

$$\begin{aligned}
\frac{\partial E_{\text{out}}(h)}{\partial h} &= 2\mathbb{E}_{X,Y}[h(x) - y] \\
&= 2\mathbb{E}_X\left[\mathbb{E}[(h(x) - y)|x]\right] \quad \text{(Lamma)} \\
&= 2\mathbb{E}_X\left[\mathbb{E}[h(x)|x] - \mathbb{E}[y|x]\right] \\
&= 2\mathbb{E}_X\left[h(x) - \mathbb{E}[y|x]\right]
\end{aligned}$$

The necessary condition for minimizing the $E_{\text{out}}(h)$ is to let $2\mathbb{E}_X[h(x) - \mathbb{E}[y|x]] = 0$. Since $\mathbb{E}_X[h(x) - \mathbb{E}[y|x]]$ is the function of x, so if $h(x) - \mathbb{E}[y|x] = 0$, then $\dfrac{\partial E_{\text{out}}(h)}{\partial h} = 0$, which minimizes $E_{\text{out}}(h)$.

So we have $h^*(x) = \mathbb{E}[y|x]$.

2. Since $y = h^*(x) + \epsilon(x)$, we have

$$\mathbb{E}[\epsilon(x)] = \mathbb{E}[y - h^*(x)] = \mathbb{E}(y) - \mathbb{E}[h^*(x)] = \mathbb{E}[y] - \mathbb{E}[\mathbb{E}[y|x]] \overset{\text{Lamma}}{=} \mathbb{E}[y] - \mathbb{E}[y] = 0$$

So above all, we have prove that $h^*(x) = \mathbb{E}[y|x]$ and $\epsilon(x)$ has expected value zero.

Problem 3 According to the iterative format provided as follow:

- Use the SUV dataset to implement (using Python or MATLAB) the Gradient Descent method to find the optimal model for logistic regression.
- What is your test error? What are the sizes of the training set and test set, respectively?
- What is your learning rate? How was it chosen? How many steps were iterated in total?
- Present the results of the last 10 steps produced by your algorithm, including the loss, learning rate, the L2 norm of the gradient, and the number of function evaluations and gradient evaluations.

Fixed learning rate gradient descent:
1: Initialize the weights at time step $t = 0$ to $\mathbf{w}(0)$.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3:     Compute the gradient $\mathbf{g}_t = \nabla E_{\text{in}}(\mathbf{w}(t))$.
4:     Set the direction to move, $\mathbf{v}_t = -\mathbf{g}_t$.
5:     Update the weights: $\mathbf{w}(t+1) = \mathbf{w}(t) + \eta \mathbf{v}_t$.
6:     Iterate to the next step until it is time to stop.
7: Return the final weights.

Dataset reference and description
Dataset and download

Solution

Each buyer's information has a unique 'User ID', so we can just drop the 'User ID' column.
To use the 'Gender' information, we map 'Male' into 1, and 'Female' into 0.
Then the 'Gender', 'Age', 'EstimatedSalary' can be used as the input features, and the 'Purchased' can be used as the output label.

1. Since we are applying the logistic regression, so the testing error is:

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n})$$

which have been proved in Problem 1. And N is the size of the testing set's size.
And the gradient is:

$$\nabla E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n)$$

$\eta = 0.1$ $\eta_t = \eta \cdot \|\nabla E_{\text{in}}(\mathbf{w})\|$