CS446: Machine Learning, Fall 2017, Homework 1

Name: Yiming Gao (yimingg2)

Worked individually

Logistic Regression: Deriving Gradient Descent Update Rules

Problem 1

Solution:
We have

$$log\frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x}} = \boldsymbol{w^T x},$$

which is equivalent to

$$P(y=1|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{1+e^{w_0+\sum_{i=1}^{d} w_i x_i}} = \frac{1}{1+e^{-\boldsymbol{w^T x}}},$$

$$P(y=0|\boldsymbol{x},\boldsymbol{w}) = 1 - P(y=1|\boldsymbol{x},\boldsymbol{w}) = \frac{e^{-\boldsymbol{w^T x}}}{1+e^{-\boldsymbol{w^T x}}},$$

Problem 2

Find the derivative of the sigmoid function.
Solution:
We denote Sigmoid function as $\boldsymbol{\sigma(z)}$ and derive its derivative as follows:

$$(1+e^{-z})\boldsymbol{\sigma(z)} = 1$$

$$\Rightarrow -e^{-z}\sigma + (1+e^{-z})\frac{d\sigma}{dz} = 0$$

$$\begin{aligned}
\Rightarrow \frac{d}{dz}\boldsymbol{sigm}(z) &= \sigma \cdot \frac{e^{-z}}{1+e^{-z}} \\
&= \sigma \cdot \frac{(1+e^{-z})-1}{1+e^{-z}} \\
&= \sigma \cdot [1 - \frac{1}{1+e^{-z}}] \\
&= \sigma \cdot (1-\sigma)
\end{aligned}$$

Problem 3, 4

Reference:http://web.engr.oregonstate.edu/ xfern/classes/cs534/notes/logistic-regression-note.pdf

Derive the Likelihood function of Logistic Regression.

Solution:

The log likelihood function is as follows:

$$logp(D|M) = \sum_{i=1}^{N} logp(\boldsymbol{x}_i, y_i) = \sum_{i=1}^{N} logp(y_i|\boldsymbol{x}_i)p(\boldsymbol{x}_i).$$

Let $\sigma(\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}}}$ denote the **sigmoid** function. Note that in logistic regression, we don't care about $p(\boldsymbol{x}_i)$ and only need to learn the $p(y|\boldsymbol{x})$. Thus we have

$$L(\boldsymbol{w}) = \sum_{i=1}^{N} logp(y_i|\boldsymbol{x}_i) = \sum_{i=1}^{N} log\sigma(\boldsymbol{x}_i, \boldsymbol{w})^{y_i}(1-\sigma(\boldsymbol{x}_i, \boldsymbol{w}))^{1-y_i}$$

To miximize L with respect to $\boldsymbol{w}$, we look at each example:

$$L_i(\boldsymbol{w}) = log\sigma(\boldsymbol{x}_i, \boldsymbol{w})^{y_i}(1-\sigma(\boldsymbol{x}_i, \boldsymbol{w}))^{1-y_i} = y_i log\sigma(\boldsymbol{x}_i, \boldsymbol{w}) + (1-y_i)log(1-\sigma(\boldsymbol{x}_i, \boldsymbol{w}))$$

where $\sigma(\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}}}$.

Taking gradient of L_i with respect to $\boldsymbol{w}$, we have

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L_i &= \frac{y_i}{\sigma(\boldsymbol{x}_i, \boldsymbol{w})}\nabla_{\boldsymbol{w}}\sigma - \frac{1-y_i}{1-\sigma(\boldsymbol{x}_i, \boldsymbol{w})}\nabla_{\boldsymbol{w}}\sigma \\
&= \frac{y_i}{\sigma}\sigma(1-\sigma)\boldsymbol{x}_i - \frac{1-y_i}{1-\sigma}\sigma(1-\sigma)\boldsymbol{x}_i \\
&= [y_i(1-\sigma) - (1-y_i)\sigma]\boldsymbol{x}_i \\
&= (y_i - \sigma(\boldsymbol{x}_i, \boldsymbol{w}))\boldsymbol{x}_i
\end{aligned}$$

(1) So for a **single** training example, the update rule is:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta(y_i - \sigma(\boldsymbol{x}_i, \boldsymbol{w}))\boldsymbol{x}_i)$$

(2) Consider **all** training examples, we have

$$\nabla_{\boldsymbol{w}} L = \sum_{i=1}^{N}(y_i - \sigma(\boldsymbol{x}_i, \boldsymbol{w}))\boldsymbol{x}_i$$

The update rule for Gradient Descent is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta\nabla LL(\boldsymbol{w}_k)$$

where $\nabla LL(\boldsymbol{w}_k) = \sum_{i=1}^{N}(y_i - \sigma(\boldsymbol{x}_i, \boldsymbol{w}_k))\boldsymbol{x}_i$ and η is the stepsize of Gradient Descent.