CS 446: Machine Learning
Homework 3: Binary Classification

Due on Tuesday, Feb 06, 2018, 11:59 a.m. Central Time

1. **[15 points]** Binary Classifiers

(a) In order to use a linear regression model for binary classification, how do we map the regression output $\mathbf{w}^\top\mathbf{x}$ to the class labels $y \in \{-1, 1\}$?

Your answer: We simply take the sign of the predicted real value:

$$y = \text{sign}(\mathbf{w}^\intercal\mathbf{x})$$

(b) In logistic regression, the activation function $\text{g}(a) = \frac{1}{1+e^{-a}}$ is called sigmoid. Then how do we map the sigmoid output $\text{g}(\mathbf{w}^\top\mathbf{x})$ to binary class labels $y \in \{-1, 1\}$?

Your answer: Since the sigmoid function maps real numbers to numbers $\in [0, 1]$ its value can be interpreted as a probability. Therefore,

$$y = \begin{cases} 1 & \text{if } \text{g}(\mathbf{w}^\intercal\mathbf{x}) > \alpha \\ -1 & \text{if } \text{g}(\mathbf{w}^\intercal\mathbf{x}) \leq \alpha \end{cases}$$

Where $\alpha \in [0, 1]$ is the threshold chosen. A typical choice is $\alpha = 0.5$.

Another way to write this is

$$y = \text{sign}(\text{g}(\mathbf{w}^\intercal\mathbf{x}) - \alpha)$$

(c) Is it possible to write the derivative of the sigmoid function g w.r.t a, i.e. $\frac{\partial g}{\partial a}$, as a simple function of itself g? If so, how?

Your answer: Yes. By the chain rule we have that

$$\begin{aligned}
\frac{\partial g}{\partial a} &= \frac{e^{-a}}{(1+e^{-a})^2} \\
&= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} \\
&= \frac{1}{1+e^{-a}} \cdot \frac{1+e^{-a}-1}{1+e^{-a}} \\
&= \frac{1}{1+e^{-a}} \cdot \left(1 - \frac{1}{1+e^{-a}}\right) \\
&= g(a) \cdot (1 - g(a))
\end{aligned}$$

(d) Assume quadratic loss is used in the logistic regression together with the sigmoid function. Then the program becomes:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \frac{1}{2}\sum_i \left(y_i - \text{g}(\mathbf{w}^\top\mathbf{x}_i)\right)^2$$

where $y \in \{0, 1\}$. To solve it by gradient descent, what would be the $\mathbf{w}$ update equation?

Your answer: Just like with the least square loss function the update is

$$\mathbf{w} = \mathbf{w} - \alpha \nabla \mathbf{f}$$

Where ∇f is defined as

$$\nabla f \triangleq \left\langle \frac{\partial f}{\partial \mathbf{w}_k} \right\rangle_{k=1}^{K}$$

where K is the number of dimensions. Now,

$$\frac{\partial f}{\partial \mathbf{w}_k} = \sum_{i=1} \frac{\partial f}{\partial \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right)} \cdot \frac{\partial \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right)}{\partial \mathbf{w}_\mathbf{k}}$$

But

$$\frac{\partial f}{\partial \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right)} = \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right) - y_i$$

And

$$\begin{aligned} \frac{\partial \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right)}{\partial \mathbf{w}_k} &= \frac{\partial \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right)}{\partial \mathbf{w}^\intercal \mathbf{x}_i} \cdot \frac{\partial \mathbf{w}^\intercal \mathbf{x}_i}{\partial \mathbf{w}_k} \\ &= \left(\mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right) \mathrm{g}\left(-\mathbf{w}^\intercal \mathbf{x}_i\right)\right) \cdot \mathbf{x}_i^k \end{aligned}$$

Therefore,

$$\frac{\partial f}{\partial \mathbf{w}_k} = \sum_{i=1} \left(\left(\mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right) - y_i\right) \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right) \left(1 - \mathrm{g}\left(\mathbf{w}^\intercal \mathbf{x}_i\right)\right) \mathbf{x}_i^k \right)$$

(e) Assume $y \in \{-1, 1\}$. Consider the following program for logistic regression:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \sum_i \log\left(1 + \exp(-y^{(i)} \mathbf{w}^T \phi(x^{(i)}))\right).$$

The above program for binary classification makes an assumption on the samples/data points. What is the assumption?

Your answer: Lets remember that the expression above comes from using the negative log likelihood of the entire dataset. The expression is derived from

$$\prod_i p(y^{(i)} | x^{(i)})$$

where p follows a logistic distribution. The above is true **if** the data samples are i.i.d. That's the assumption made.