University of Illinois Spring 2018

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:

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

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

   Your answer:

   $$P(y^{(i)} = 1 | x^{(i)}; w) = \frac{1}{1 + e^{-\mathbf{w}^\intercal \mathbf{x}^{(i)}}}$$
   $$P(y^{(i)} = -1 | x^{(i)}; w) = \frac{1}{1 + e^{\mathbf{w}^\intercal \mathbf{x}^{(i)}}}$$

   So we can combine the two equations to get more compact.

   $$P(y^{(i)} | x^{(i)}; w) = \frac{1}{1 + e^{-y^{(i)} \mathbf{w}^\intercal \mathbf{x}^{(i)}}}$$

   If the sigmoid output $g(\mathbf{w}^\intercal \mathbf{x}) >= 0.5$:

   y = 1

   If the sigmoid output $g(\mathbf{w}^\intercal \mathbf{x}) < 0.5$:

   y = -1

   (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:
Actually, we can write the derivative of the sigmoid function as follows:

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

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

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

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

Your answer:

$$\begin{aligned}\mathbf{w}^{(j)} &:= \mathbf{w}^{(j)} - \alpha \frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} \\ &:= \mathbf{w}^{(j)} - \alpha(g(\mathbf{w}^\intercal \mathbf{x}_i) - y_i) \cdot (g(\mathbf{w}^\intercal \mathbf{x}_i)) \cdot (1 - g(\mathbf{w}^\intercal \mathbf{x}_i)) \cdot \mathbf{x}_i^{(j)}\end{aligned}$$

where $\mathbf{w}^{(j)}$ indicates the j^{th} element of $\mathbf{w}$.

(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:
The assumption is that samples/data points are i.i.d.(independent, identically distributed).