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 $y \in \{-1, 1\}$?

   Solution:
   (2 points)
   $$y = \text{sign}(\mathbf{w}^\top \mathbf{x})$$
   where sign is the sign function or signum function.
   $$\textit{or } y = 2\,\text{u}(\mathbf{w}^\top \mathbf{x}) - 1$$
   where u is the unit step function.

   (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 $y \in \{-1, 1\}$?

   Solution:
   (3 points)
   $$y = \text{sign}(\text{g}(\mathbf{w}^\top \mathbf{x}) - 0.5)$$
   where sign is the sign function or signum function.
   $$\textit{or } y = 2\,\text{u}(\text{g}(\mathbf{w}^\top \mathbf{x}) - 0.5) - 1$$
   where u is the unit step function.

   (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?

   Solution:
   (3 point)
   Yes
   $$\text{g}'(a) = \text{g}(a)(1 - \text{g}(a))$$

   (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?

Solution:
(5 points)

$$\mathbf{w}_{t+1} := \mathbf{w}_t - \alpha \sum_i (y_i - \mathrm{g}(\mathbf{w}_t^\top \mathbf{x}_i))\, \mathrm{g}(\mathbf{w}_t^\top \mathbf{x}_i)(1 - \mathrm{g}(\mathbf{w}_t^\top \mathbf{x}_i))\mathbf{x}_i$$

or

$$w_{j,t+1} := w_{j,t} - \alpha \sum_i (y_i - \mathrm{g}(\mathbf{w}_t^\top \mathbf{x}_i))\, \mathrm{g}(\mathbf{w}_t^\top \mathbf{x}_i)(1 - \mathrm{g}(\mathbf{w}_t^\top \mathbf{x}_i))x_{j,i}$$

(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?

Solution:
(2 points)
Independently drawn from an identical distribution, or i.i.d.