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: $\mathbf{y} = sign(\mathbf{w}^T\mathbf{x})$. Positive values will be treated as 1 and negative values will be treated as -1.

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

Your answer: $\mathbf{y} = sign(g(\mathbf{w}^T\mathbf{x}) - 0.5)$. Values above 0.5 will be treated as 1 and values under 0.5 will be treated as -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:

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

(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:

$$\begin{aligned}
k^{(i)} :=& g(\mathbf{w}^T\mathbf{x}^{(i)}) \\
\nabla_{\mathbf{w}} =& -(\mathbf{y}^{(i)} - k^{(i)})k^{(i)}(1-k^{(i)})\mathbf{x}^{(i)} \\
\mathbf{w}_{n+1} =& \mathbf{w}_n + \eta\nabla_{\mathbf{w}_n}
\end{aligned}$$

(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: Whenever we use logistic regression we do so under the assumption that all the independent variables are truly independent of one another and are identically distributed.