Introductory Applied Machine Learning, Tutorial 3 Solutions

School of Informatics, University of Edinburgh, Instructor: Nigel Goddard

September 2017

1. Consider using logistic regression for a two-class classification problem in two dimensions:

$$p(y = 1|\boldsymbol{x}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$$

Here σ denotes the logistic (or sigmoid) function $\sigma(z) = 1/(1 + \exp(-z))$, y is the target which takes on values of 0 or 1, $\boldsymbol{x} = (x_1, x_2)$ is a vector in the two-dimensional input space, and $\boldsymbol{w} = (w_0,\ w_1, w_2)$ are the parameters of the logistic regressor.

(a) Consider a weight vector $\boldsymbol{w}_A = (-1, 1, 0)$. Sketch the decision boundary in $\boldsymbol{x}$ space corresponding to this weight vector, and mark which regions are classified with labels 0 and 1.
Solution: The decision boundary is given by $-1+x_1 = 0$ or $x_1 = 1$. So in $\boldsymbol{x}$ space this is a vertical line through $x_1 = 1$, with the halfspace to the right being class 1.

(b) Consider a second weight vector $\boldsymbol{w}_B = (5, -5, 0)$. Again sketch the decision boundary in $\boldsymbol{x}$ space corresponding to this weight vector, and mark which regions are classified with labels 0 and 1.
Solution: The decision boundary is given by $5 - 5x_1 = 0$ or $x_1 = 1$. So in $\boldsymbol{x}$ space this is a vertical line through $x_1 = 1$, with the halfspace to the left being class 1.

(c) Plot $p(y = 1|\boldsymbol{x})$ as a function of x_1 for both $\boldsymbol{w}_A$ and $\boldsymbol{w}_B$, and comment on any differences between the two.
Solution: For $\boldsymbol{w}_A$ we have a logistic function which goes to 0 as $x_1 \to -\infty$, and to 1 as $x_1 \to \infty$. It has value 0.5 at $x_1 = 1$. For $\boldsymbol{w}_B$ we have a logistic function which goes to 1 as $x_1 \to -\infty$, and to 0 as $x_1 \to \infty$. It has value 0.5 at $x_1 = 1$. Most importantly the logistic function for $\boldsymbol{w}_B$ is steeper than that for $\boldsymbol{w}_A$ around $x_1 = 1$.

2. Consider the logistic regression setup in the previous questions, but with a new weight vector $\boldsymbol{w}_A = (0, -1, 1)$. Consider the following data set: Compute the gradient of the log likelihood of the logistic

Instance	x_1	x_2	Class
0	0.5	-0.35	$-$
1	-0.1	0.1	$-$
2	-1.2	1.0	$+$

regression model for this data set. Suppose that we take a single gradient step with $\eta = 1.0$; what is the new parameter setting? Do the new parameters do a better job of classifying the training data?

Solution: For each training instance, first compute $p(y = 1|\boldsymbol{x}_i) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$ and $y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$:

Instance	x_1	x_2	$\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$	$y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$
0	0.5	-0.35	0.30	-0.30
1	-0.1	0.1	0.55	-0.55
2	-1.2	1.0	0.90	0.1

Then the partial derivatives are

$$\frac{\partial L}{\partial w_1} = (-0.3 \cdot 0.5) + (-0.55 \cdot -0.1) + (0.1 \cdot -1.2) = -0.215$$
$$\frac{\partial L}{\partial w_2} = (-0.3 \cdot -0.35) + (-0.55 \cdot 0.1) + (0.1 \cdot 1.0) = 0.15$$
$$\frac{\partial L}{\partial w_0} = (-0.3 \cdot 1) + (-0.55 \cdot 1) + (0.1 \cdot 1) = -0.75$$

Note that w_0 is special, because it corresponds to a feature that is always 1. (Without it, the decision boundary would always pass through the origin.)

The new parameter vector is $\boldsymbol{w}' = (-0.75, -1.215, 1.15)$. (In practice, we would never choose a step size like $\eta = 1.0$; it is much too large, but it makes for a better example.) If we reproduce the above table at the new weight setting, we get

Instance	x_1	x_2	$p(y=1\|\boldsymbol{x}_i)$
0	0.5	-0.35	0.15
1	-0.1	0.1	0.37
2	-1.2	1.0	0.87

So the two negative instances have seen their classification improve, while the positive instance has gotten slightly worse. This will be set right with further gradient steps.