Introductory Applied Machine Learning, Tutorial 3

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.

(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.

(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.

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

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

logistic 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?

It will help you to remember the following facts:

- The log-likelihood in logistic regression is

$$\begin{aligned} L(\boldsymbol{w}) &= \sum_{i=1}^{n} \log p(y = y_i|\boldsymbol{x}_i) \\ &= \sum_{i=1}^{n} [y_i \log p(y = 1|\boldsymbol{x}_i) + (1 - y_i) \log p(y = 0|\boldsymbol{x}_i)] \end{aligned}$$

- The partial derivative of the log-likelihood with respect to a parameter w_j is

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} (y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)) x_{ij}$$

- To maximize a function $L(\boldsymbol{w})$, we use the gradient *ascent* rule, which is

$$\boldsymbol{w}' \leftarrow \boldsymbol{w} + \eta \nabla L$$