Literate Statistical Analysis Assignment

A. I. McLeod

University of Western Ontario

Abstract

We illustrate the use of formatted R output as well as figures and tables in literate statistical analysis. For cropping and adjusting Figures, please my other tutorial in the file `AIMSweave.Rnw`. The use of BibTeX is also demonstrated.

Keywords: classification, logistic regression, regression.

1. Question 1

In this report we compare using linear and logistic regression for class prediction with the mixture data discussed by Hastie, Tibshirani, and Friedman (2009, Chapter 2). This data was obtained from the R package `ElemStatLearn` Halvorsen (2010). When loaded the R package `ElemStatLearn` provides a list variable `mixture.example` with components `x` and `y`. The inputs are `x`, a matrix with two columns and 200 rows. The output `y` is a vector of length 200 comprised 100 0's followed by 100 1's representing the classes.

The fitted regression is summarized in the table below

```
(Intercept)          x1          x2
  0.3290614  -0.0226360   0.2495983
```

and for logistic regression the estimated parameters are,

```
(Intercept)          x1          x2
 -0.9780417  -0.1343739   1.3980688
```

Although, at a superficial level, the estimated parameters look very different for the two models, both models give essentially an equivalent fit in this case as shown in Figure 1. Figure 1 compares the decision boundaries for regression and logistic regression. We see they are almost the same.

In Tables 1 and 2, we compare the confusion matricies regression and logistic regression using the training data. The misclassification rates shown in Table 3 are the same although the confusion matricies are slightly different.

	$y = 0$	$y = 1$
$\hat{y} = 0$	70	24
$\hat{y} = 1$	30	76

Table 1: Confusion matrix, linear regression, training data

	$y = 0$	$y = 1$
$\hat{y} = 0$	72	26
$\hat{y} = 1$	28	74

Table 2: Confusion matrix, logistic regression, training data

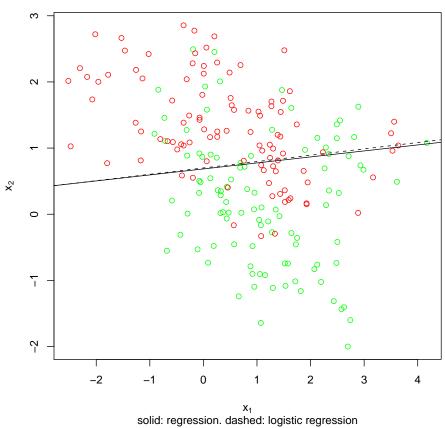**Comparing Regression and Logistic Regression**
Scenario 2

solid: regression. dashed: logistic regression

Figure 1: Mixture Data

	Regression	Logistic Regression
η	0.27	0.27

Table 3: Mis-classification rates, training data

References

Halvorsen K (2010). *ElemStatLearn: Data sets, functions and examples ...* URL http://CRAN.R-project.org/package=ElemStatLearn.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, New York, 2nd edition.

Affiliation:

A.I. McLeod
University of Western Ontario
E-mail: aimcleod@uwo.ca