DS 5220 / CS6140-02 - Fall 2017 Olga Vitek

Homework 3

Each part of the problems 5 points
Due on Blackboard before midnight on Tuesday October 17, 2017.

We will compare four classification procedures: Logistic regression, linear discriminant analysis, logistic regression with Lasso regularization, and nearest shrunken centroids.

1. This question is similar to the previous homework. It revisits concepts related to binary classification.
   (a) Set random seed to 123. Simulate the training set: N=50 values of X_1, distributed Uniformly on interval (0,3) and N=50 values of X_2 independent from X_1, distributed Uniformly on interval (0,3). Simulate the values of $Y \sim Bernoulli(\pi = 1/(1+e^{-(-3+X_1+X_2)}))$. Set random seed to 456. Repeat the above to simulate the validation set.
   (b) Using training set, develop a classifier of Y as function of predictors X_1 and X_2 using an existing implementation of logistic regression (e.g., `glm` in R) and an existing implementation of LDA (e.g., `MASS::lda` in R). Make a plot where the x axis is X_1, the y axis is X_2. Overlay the simulated observations from the validation labeled with their class. Overlay the true decision boundary and the decision boundaries estimated by the two methods, report the proportion of correctly classified observations in the validation set, and interpret the results.
   (c) Implement batch gradient descent and stochastic gradient descent to estimate the parameters of logistic regression, as function of the learning rate α. Stop the iterations when the change in the parameter between two consecutive iterations is less than a pre-defined constant of your choice, or when the number of iterations exceeds a large number (say, 1000). *[Note: consider a range of α to optimize convergence on the training set.]*
   (d) Implement linear discriminant analysis.
   (e) Repeat (a) 200 times. Each time, record the % of correctly classified observations in the validation set by batch gradient descent with best α from (c), and by stochastic gradient descent from best α from (c), and by LDA from (d). Plot three histograms of % of correctly classified observations, and comment on the results.

2. This question revisits concepts related to binary classification with regularization.
   (a) Repeat question 1(a), while adding 10 additional predictors $X_3, X_4, \ldots, X_{12}$ that are not informative of $\mathrm{P}\{Y = 1|X\}$. Let us make X_3, X_4 correlated with X_2, obtained as $X_3 = 0.8 \cdot X_2 + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 = 0.75)$, and $X_4 = 0.8 \cdot X_2 + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 = 0.75)$. Let us make the remaining predictors $X_5, X_6, \ldots, X_{12}$ independent from the other predictors, each distributed Uniformly on interval (0,3).

(b) Repeat question 1(c), while adding regularization, penalizing $\sum \beta_j^2$ as in slide 4-23. Implement cross-validation to choose optimal regularization parameter. *[Note: use the optimal α from question 1(c), or consider a range of α to optimize convergence if needed.]*

(c) Repeat question 1(d), while adding regularization as in the first expression in slide 5-29. Implement cross-validation to choose optimal regularization parameter.

(d) Repeat (a) 200 times. Each time, record the % of correctly classified observations in the validation set by:

- logistic regression, 12 predictors, batch gradient descent, no regularization
- logistic regression, 12 predictors, stochastic gradient descent, no regularization
- linear discriminant analysis, 12 predictors no regularization
- logistic regression, 12 predictors, batch gradient descent and regularization
- logistic regression, 12 predictors, stochastic gradient descent and regularization
- linear discriminant analysis, 12 predictors and regularization

(e) Comment on the results. Is regularization helpful in situations where we have more candidate predictors? In which other configuration of predictors the results may change?

3. Consider a problem of classifying a response Y with 3 classes, and two predictors X_1, X_2, using logistic regression. *[Note: this is a theoretical question and does not require implementation. You can simulate data to double-check your results, but this is not required.]*

(a) Assume that X_1 and X_2 are continuous. Write the multi-class logistic regression classification, in softmax parametrization. State the total number of parameters.

(b) Assume that X_1 is categorical with 5 categories, and X_2 is continuous. Write the multi-class logistic regression classification, in softmax parametrization. State the total number of parameters.

4. *[Data analysis question]* Consider the dataset "South African Heart Disease" available on the HTF website, and used as an example in HTF Section 4. The goal is to predict the binary variable `chd`. We will compare four classification procedures: logistic regression, linear discriminant analysis, logistic regression with lasso regularization, and nearest shrunken centroids.

(a) Partition the dataset into a training and a validation subsets of equal size, by randomly selecting rows in the training set. Explore the training set: report one-variable summary statistics, two-variable summary statistics, and discuss your findings (e.g., presence of highly correlated predictors, categorical predictors, missing values, outliers etc).

(b) Fit logistic regression on the training set. Perform variable selection using all subsets selection and AIC or BIC criteria. *[Hint: it may be interesting to also consider statistical interactions]*

(c) Fit LDA on the training set, using the standard workflow.

(d) Fit logistic regression with Lasso regularization on the training set: produce and interpret the plot of paths of the individual coefficients; produce the plot of regularization parameter versus cross-validated predicted error; select regularization parameter, and the corresponding predictors; fit the model with the selected predictors only on the full raining set.

(e) Fit the nearest shrunken centroids model on the training set: use cross-validation to select the best regularization parameter; refit the model with the selected regularization parameter; visualize the centroids of the selected model.

(f) Evaluate the performance of the classifiers using ROC curves on the training and on the validation set.

(g) Summarize your findings. How do the results differ between the training and the validation set? Which approach(es) perform(s) better on the validation set? What is are the reasons for this difference in performance? Which models are more interpretable?