MIS 382N: Advanced Predictive Modelling

Assignment #3

Due: Wed, Oct 22, 2014 3.30 pm; Total points: 40

Upload instructions: Your homework should be a single pdf file containing all the answers, plots/pictures and code. You may however insert equations by hand if you wish. Name the pdf file as <your eid>.pdf The included code should be straightforward to run by just copying it into the interpretrator or the command line without any changes. So add any include statements for libraries etc. in the code as well. Homeworks are due at the beginning of class on the due date, submitted through Canvas. If the choice of a programming language is specified in the problem, please don't use any other language.

1. **(4+4+2+2=12) Ridge/Lasso** In this question, you will explore the application of Lasso and Ridge using `sklearn` in `Python`. `autos.csv` (available on canvas) contains the data in the following format : the first column is the target variable, the next 6 columns are the features, and the last column is the train/test split.

   (a) Use `sklearn.linear_model.Lasso` and `sklearn.linear_model.Ridge` classes to do a 5-fold cross validation. The fold splits of the training data are provided in the file `auto.folds`. For the sweep of the regularization parameter, use $[0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]$ for ridge and $[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1]$ for lasso. Report the best parameter chosen based on cross-validation.

   (b) Run ridge and lasso for the all of the parameters specified above, and plot the coefficients learnt for each of them - there should be one plot each for lasso and ridge, so a total of two plots; the plots for different features for a method should be on the same plot (e.g. Fig 6.6 of JW). What do you qualitatively observe when value of the regularization parameter is changed?

   (c) Run least squares regression, ridge, and lasso on the full training data. For ridge and lasso, use only the best regularization parameter selected above. Report the prediction error on the test data for each.

   (d) For the best lasso parameter, determine the variables that were not dropped. Using only these variables, run least squares regression on full training data and report the prediction error on the test data.

2. **(1+2+2+2=7) Multilevel Modelling for categorical data** In this question you will use `lme4` package in `R` to explore multilevel modelling. Download the simulated dataset `mlm.Rdata` from canvas, and load it in R. The data is based on a behavioural study of students across a few schools. The data already has `train` and `test` splits. The target variable is `Y`. There are three numerical variables and two categorical variables.

   (a) Build a simple least squares model (use `lm`) using `open + agree + social` as independent variables. Report the prediction MSE on the test data.

   (b) The default way to encode categorical variables is to dummy code them. Build three regression models using `glm`, by including dummy coded `cat1` and `cat2` variables separately in addition to the ones used in part (a), and finally by including both of them as additive effects. Report the prediction accuracy on the test data again. Build a fourth model which has a new variable that is a combination of every possible pairs of `cat1` and `cat2` variables in addition to ones in part (a). (such terms are called interaction terms). You can use the operator `:` to do this in `glm`.

Comparing the prediction errors from the above five experiments, which one does best? Why do you think that is the case?

(c) Use `lmer` (package `lme4`) to build a model with varying intercept on `cat1` and `cat2`, in addition to the fixed intercepts for all the numerical variables. Also, fit a hierarchical model with `cat1` $\rightarrow$ `cat2` as the nested group random effect, with numerical variables as the fixed effect. Report the prediction accuracy and compare with corresponding models in part(a).

(d) Consider the combination of categories model in part(b) and the nested group model of part (c). When would one be preferable over the other ?

3. **(4+4+4=12 pts) Stochastic Gradient Descent**

In this problem you will implement SGD to estimate the parameters of a Ridge Regression problem. The dataset is derived from `https://archive.ics.uci.edu/ml/datasets/Forest+Fires`. However, the data is already partitioned into `forestfire-train.csv` and `forestfire-test.csv` (available on canvas). It contains various features, where the last column (`area`) is the target variable. You should standardize the training data (make each column including the target variable mean 0 and variance 1).

In this problem you will use the template script `hw3sgd.R` (available on canvas) to write an R function

```
sgd_ridge = function(learn.rate, lambda, train, epoch=100),
```

which solves the Ridge Regression problem using stochastic gradient descent:

$$\min_w \frac{1}{2} \|Xw - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2$$

(a) Write and implement an SGD algorithm to solve the Ridge Regression problem instead of using some library. The function *sgd_ridge* should return the optimal weight vector. You will have to fill in the placeholders marked as `TODO` in the script.

(b) Consider the following different learning rates 0.000025, 0.00055, and 0.0075 for this problem. For each learning rate and $\lambda = 0.1$, as you make multiple passes (epochs) over the training data, record the Root Mean Squared Error (RMSE) obtained for both training and test data, and plot the number of epochs vs RMSE. What is a reasonable number of epochs (one answer per learning rate) after which you stop training?

(c) Report the RMSE obtained for both training and test data, for learning rate = 0.000025 and $\lambda = \{0.01, 0.1, 1\}$. How does the model corresponding to the optimal λ value compare in terms of RMSE on the test data to (i) a simple MLR model obtained using SGD (ii) a batch solution (i.e. using MLR directly on the entire data)?

4. **(3+2=5 pts) Decision Trees**

(a) Write one paragraph on how decision trees can be implemented in a (reasonably) scalable way on a distributed computing system. You can use any resource, but should cite the resource(s) and write in your words - don't copy verbatim.

(b) In this problem you will model the data using decision trees to perform classification task. Load the breast cancer dataset `BreastCancer.csv` (available on canvas) in R. The dependent variable of interest is *diagnosis*. Using the package *rpart*, build two different trees with a maximum depth of two using the split criteria (i) Gini and (ii) Entropy. Plot the two trees. At which node do they differ?

5. **(2+2=4 pts) Finding Decision Boundary**

(a) Suppose samples in $\mathbb{R}^2$ (the two-dimensional Cartesian space) are being obtained from two classes, C1 and C2, both of which are normally distributed with means at (1.5,1) and (1,1.5) respectively. The covariance matrix for each class is the same:

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad (1)$$

If the priors of C1 and C2 are 4/7 and 3/7 respectively, what is the ideal (i.e. Bayes Optimal) decision boundary? (derive the equation for this boundary)

(b) Suppose the cost of misclassifying an input actually belonging to C2 is twice as expensive as misclassifying an input belonging to C1. Correct classification does not incur any cost. If the objective is to minimize the expected cost rather than expected misclassification rate, what would be the best decision boundary? (obtain the equation describing this boundary).