November 11, 2013

INTRODUCTION TO APPLIED MACHINE LEARNING

s1115104

Coursework 3

1. a

Simple logistic classifier (SLC) obtains 66.8258% correctly classified instances out of 419 total instances compared to 64.2005% correctly classified instances for the *Logistic classifier(LC)*. Looking at the *InfoGainAttributeEval* results, we obtain a large number of attributes with 0 information gain that do not contribute to the classifier in any way. Similar pattern is apparent from the output of SLC and LC. Not all attributes are used to calculate the logistic classifier because they are not relevant to the classifier. The main difference in the classifiers is that classes that are not contributing any information are not used at all in the case of SLC while they are assigned value of infinity for LC.

1. b

Plotting log(ridge) values against the percent correct we obtain the following graph:

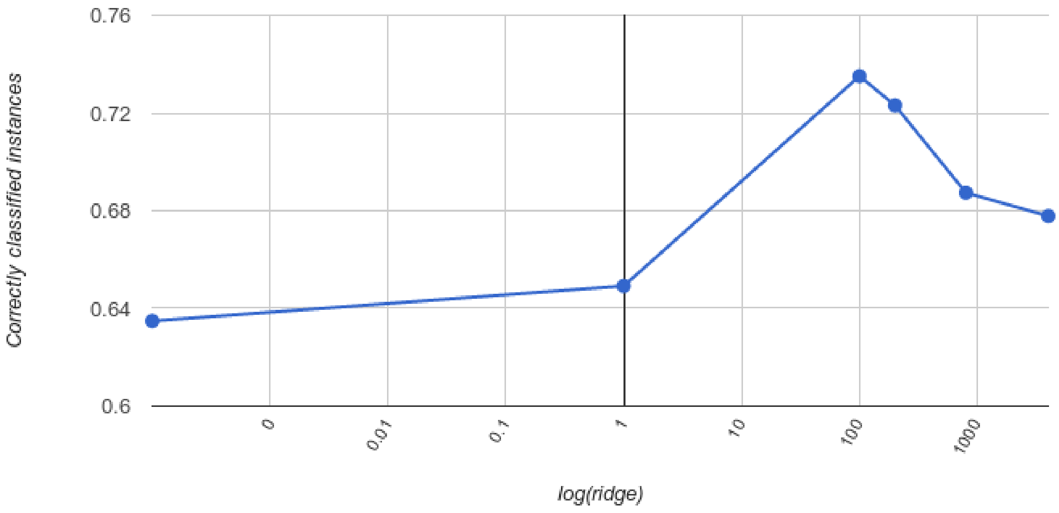


Figure 1 - Percentage correct as a function of the ridge value

Regularization is a general approach to add a “complexity parameter” to a learning algorithm and determines how much ‘wiggliness’ we’re willing to give to the curve we are trying to predict.

The main difference between regularization and feature selection is during feature selection we remove features that do not add any more useful information into the classifier while regularization maintains those features but penalizes a model based on it’s number of parameters as a way to not overfit the train data.

1. c

Plotting varying values of gamma against correctly classified instances percentage we obtain the following graph:

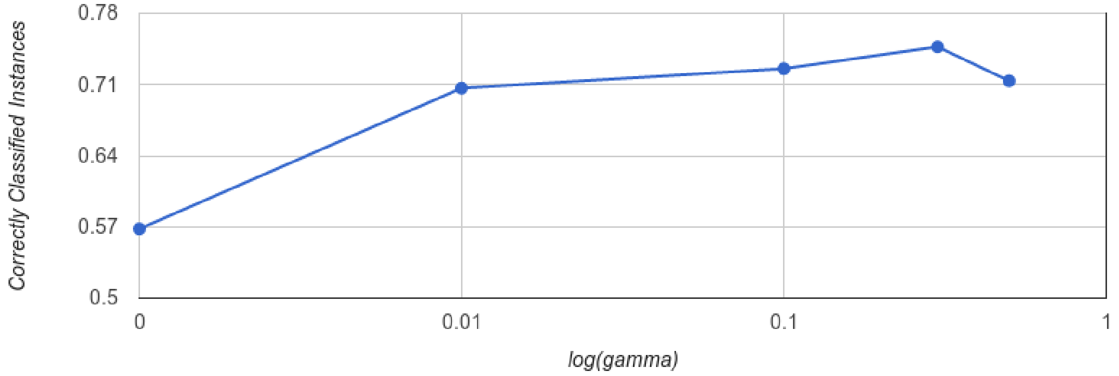


Figure 2 - Percentage correct as a function of the gamma

The percentage is obtained for *gamma = 0.3* .

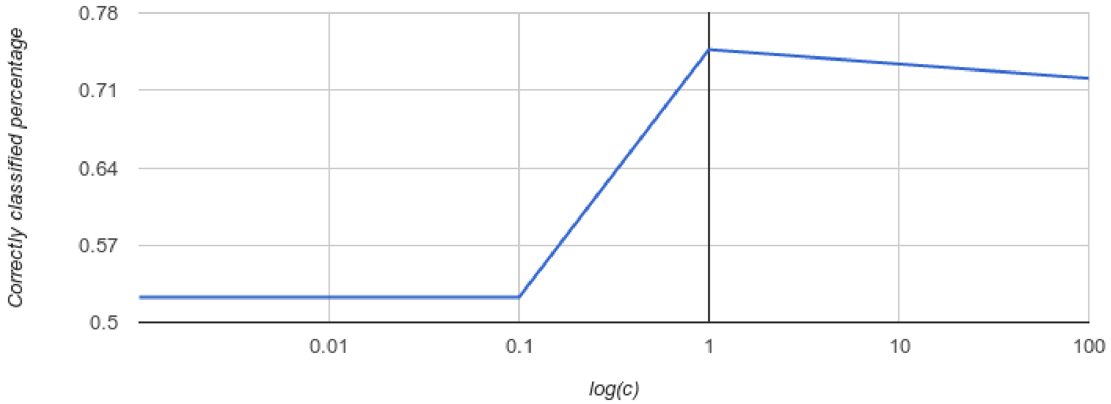


Figure 3 - Percentage correct as a function of c - the complexity parameter.

The highest percentage obtained occurs for *c = 1* .

We have found the optimal values for gamma and c, however, in order to find a combined optimal values we would have to consider all pairs in order to obtain the highest percentage correct. Having said that, considering separate values of gamma and c will provide us with an idea of their correlation.

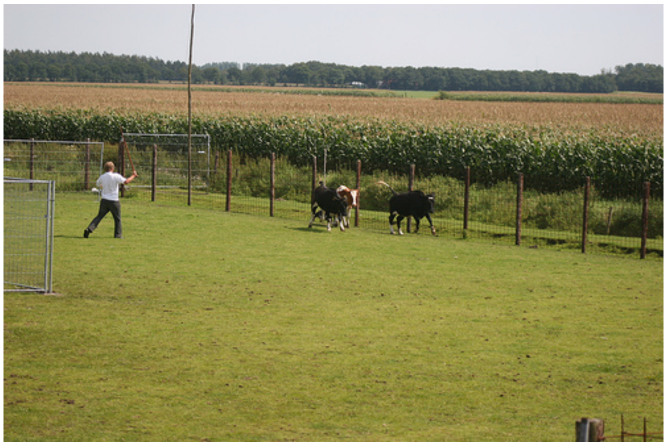
1. d

Looking at the list of the best 50 features, we notice there are classes *is_bird* and *is_aeroplane* in this list. This can be attributed to the fact that if there is a picture of a bird or an airplane, it may help us decide that there is likely to not be a person in the picture or vice versa, depending on the set of images.

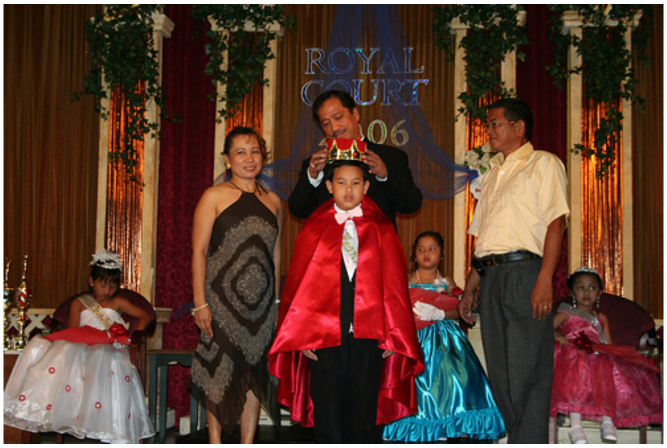
Evaluating Simple Logistic Classifier on the validation data, we obtain *PC = 76.46%*, an increase of about 10 percent. We notice that the binary class labels are added to the decision function and create a linear decision boundary for that attribute. For example, *is_cow* is *[is_cow] * 0.88* for class 0 (no person) and *[is_cow] * -0.88* for class 1 (person). Essentially, we have added complexity to the classifier and used it to improve the PC.

This implies that in practice, if we wanted to make a decision about a person being or not being in a picture, we could simply look at *is_<class>* and attempt to predict the result. Alternatively, we could look at few classes with higher importance such as *is_aeroplane*, *is_bird* and *is_cat* (higher significance classes in the classifier) and make a prediction based on those.

1. e

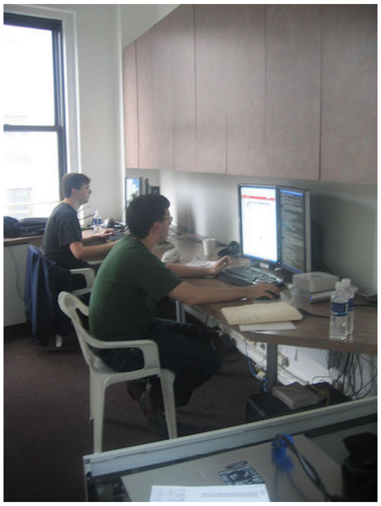


Misclassified with strong confidence - Img1



Misclassified with low confidence - Img2

Correctly classified with low confidence - Img3



Correctly classified with high confidence - Img4

Img1 - Person is not clear cut in the picture and presence of large amounts of green color may have skewed the classification.

Img2 - Large number of distinct colors could have the effect of incorrectly identifying the objects.

Img3 - Presence of the horse could have an influence on the classifier.

Img4 - Distinct colors for the people and different colors for the environment may help identify the picture correctly.

2.

Without any pre-processing, Logistic Classifier with default attributes produces *PC = 80.8511*. However, labels for *is_tvmonitor* and similar are present in the dataset and not in the test dataset. Removing these attributes and re-classifying produces *PC = 67.6123*.

Running *InfoGainAttributeEval* function, we notice some features exhibit strong influence over the final decision of the classifier.

In order to gain idea how to capture relationship of the features we can run *SimpleLogistic* classifier to get a linear expression in n terms. We obtain *0.1 + [dim23] * 0.2 + [dim88] * 0.13 + [dim92] * -0.11 + [dim97] * -0.1 + [dim123] * 0.08 + [dim130] * -0.09 + [dim171] * -0.05 + [dim206] * 0.06 + [dim311] * 0.1 + [dim387] * -0.05*.

This formula can be used to add an expression and capture the relationship of the features.

Running the classifier on train data with a 80% train and 20% test data split, we obtain *PC = 69.2671%*.