Evaluating Classification Models in a Binary Classification Problem with a Small Number of Samples - Proposal

Avery Tan(altan)

Introduction

We are going to be assessing the performance of 3 learning algorithms on the ‘Breast Cancer Wisconsin Diagnostic Data Set’ obtained from the UCI Machine Learning Repository[0]. We will be performing classification using Naïve Bayes, Logistic Regression and Support Vector Machines. Among these 3 algorithms, we will statistically determine which algorithm generalizes best to the dataset.

The Dataset

The dataset under consideration is multivariate dataset consisting of 569 instances, of which 357 are benign tumors, which we will label as the negative case and 212 malignant instances, which we will label as the positive case. This results in our dataset having a proportion such that approximately 63% of the instances are for the benign or negative case and roughly 37% of instances are for the corresponding malignant or positive case. This information will be relevant when we split our dataset into the corresponding training and test datasets.

Each data instance consists of 30 features: the ID of the instance, the diagnosis or label (M = malignant, B = benign) and 30 real values computed from digitized images of a FNA of breast mass. They describe characteristics of the cell nuclei in a digitized image[0]. Before splitting the dataset, we will perform feature scaling and mean normalization on the entire dataset with the goal of reducing the range of each feature to 1. This is performed due to certain features having varying ranges which could slow gradient descent. Mean normalization involves calculating the mean for each feature and subtracting it from each of the features. Feature scaling involves dividing each feature by its range across the entire dataset.

The Learning Algorithms

We will be performing learning using Naïve Bayes , Logistic Regression and SVMs as mentioned above. Naïve Bayes is a generative classifier; it learns via Bayesian theorem and is a fairly simple classifier. Despite this simplicity, Naïve Bayes is still quite powerful in the sense that it often is able to outperform more complex classification methods.

We will also run the dataset with Logistic Regression. Logistic Regression is a discriminative classifier which learns a weight vector w which minimizes some defined loss function via stochastic gradient descent in our case. We will be using L2 or mean squared error as our loss function. The properties of this error make this a desirable cost function since it guarantees convexity. We will run this algorithm with varying values of alpha, the stepwise parameter and will also be employing regularization in our logistic regression implementation with varying values of lambda.

Support Vector Machines work by plotting each data instance in n-dimensional space where n is the number of features which in our case, n = 30. Classification is performed by finding the hyperplane that best differentiates the 2 labels. In our implementation of SVM on this dataset, we will vary the tolerance (epsilon), the kernel parameters and the regularization parameter.

The goal then is to statistically determine which of these 3 algorithms if any, perform best at generalizing the dataset.

Experimental Design and Evaluation

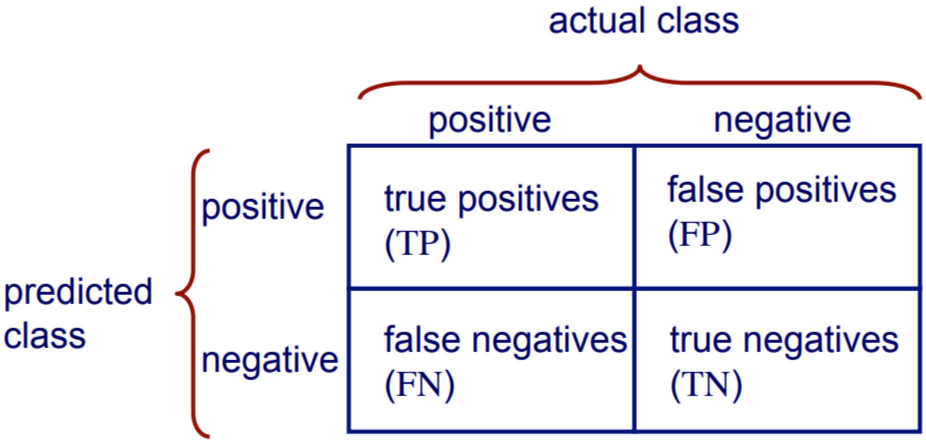
The first part of our experiment involves splitting our mean normalized and feature scaled dataset into training and test sets respectively. We will perform this split via random resampling with stratification to preserve the proportion of positive to negative labeled data in a 90/10 partition. We do so, pretending that the remaining 10% of the original dataset is not available yet and will be randomly generated later.

Although this maintains the proportions according to the dataset's label, we must recognize that there is still a change in the underlying sample statistics along the feature axes. The training dataset will be used to train models and perform cross-validation via the Leave-One-Out-Cross-Validation(LOOCV) method.

LOOCV is a special case of k-fold holdout validation in which k = n where n is the number of samples in a dataset. LOOCV is performed due to the small size of our dataset(569 instances). Although this process is expensive computationally, it is useful in cases where data is limited and withholding data from the training set is too wasteful[1]. Furthermore the pessimistic bias will be low since n-1 training samples are available for model fitting.

Pessimistic bias results when the model may not have reached full capacity due to low amounts of training samples. Another advantage of LOOCV is that each generated test set is independent of the others[1]. Thus we will use LOOCV to select the best hyperparameters of each of our respective learning algorithms considered above and train on this model. We will then estimate the generalization performance by evaluating the model's performance on the test dataset.

In estimating the predictive performance, it is not enough to just measure accuracy or error. This is due to the fact that there is a differential misclassification cost; the cost of predicting a false negative is quite higher than the cost of predicting a false positive[3]. Such is the case in cancer diagnosis.



		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

Figure 1. Confusion matrix[1]

Instead, we will use as the evaluation metric, the true positive rate or recall, defined as

$$recall = \frac{TP}{TP + FN}$$

And the false positive rate defined as :

$$false\ positive\ rate = \frac{FP}{TN + FP}$$

With these evaluation metric, the false positive rate and recall, we can plot a Receiver Operating Characteristic curve is the plot of recall vs the false positive rate for every possible classification threshold. It that allows us to visualize the performance of a binary classifier[4]. An ROC curve could also be used to quantify performance as the *area under the curve*(AUC). A AUC value close to one is considered one possible way to evaluating a very good binary classifier[2]. We can generate an ROC plot for the best model selected via cross-validation of each of the 3 algorithms.

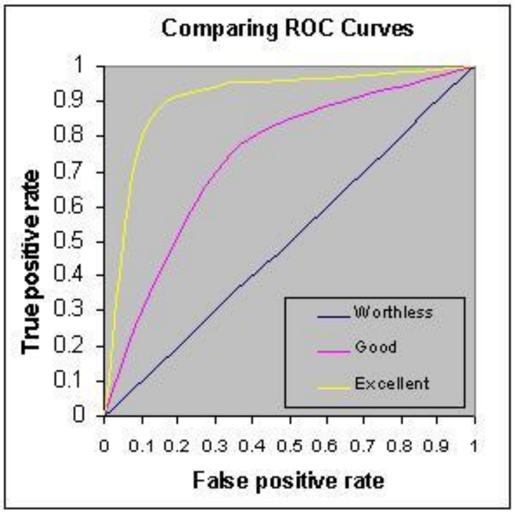


Figure 2. Receiver Operating Characteristic curve[6]

We define a set of 3 separate null hypothesis and alternative hypothesis in order to compare each of the three algorithms amongst themselves to determine which algorithm generalizes best to the dataset via a paired t test on the recall metric. A paired t test involves calculating the t statistic to determine its corresponding p-value. This is reasonable in this case as opposed to the more robust ANOVA since we are only comparing 3 learning algorithms[5].

Our null hypothesis states that there is no significant difference in the results of 2 of our learning algorithms. Since we are comparing 3 algorithms to determine which one generalizes best, we will have 3 such null hypothesis. The corresponding alternative hypothesis then states that one learning algorithm does indeed perform better than another[5].

We will set our p-value threshold to be 0.05. If the corresponding p-value is sufficiently small < 0.05, it means that there is only a 0.05 probability of the algorithm providing those results if the null hypothesis were true. We will thus we can reject the null hypothesis should this condition be met. The alternative hypothesis then states that one of the learning algorithms considered is a significant improvement over the other.

Conclusion

We will reflect on the results of our 3 learning algorithms on the ‘Breast Cancer Wisconsin Diagnostic Data Set’ postulate possible reasons or explanations as to the merit of the classifier that best generalized to this dataset. We can perhaps explore adding features or modeling complex feature interaction that may better model the distribution of the population that our dataset is drawn from.

References

[0] Archive.ics.uci.edu, 2017. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names. [Accessed: 29- Nov- 2017].

[1] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning", Sebastian Raschka's Website, 2017. [Online]. Available: https://sebastianraschka.com/blog/2016/model-evaluation-selection-part3.html. [Accessed: 29- Nov- 2017].

[2] B.E. Andrew "The Use of the Area Under The ROC Curve in the Evaluation of Machine Learning Algorithms", *Pattern Recognition*, Vol. 30, No. 7, pp. 1145-1159, 1997.

[3] D.Page, Pages.cs.wisc.edu, 2017. [Online]. Available: http://pages.cs.wisc.edu/~dpage/cs760/evaluating.pdf. [Accessed: 29- Nov- 2017].

[4] "ROC curves and Area Under the Curve explained (video)", Data School, 2017. [Online]. Available: http://www.dataschool.io/roc-curves-and-auc-explained/. [Accessed: 29- Nov- 2017].

[5] G.T. Dietterich, Citeseerx.ist.psu.edu, 2017. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.3325&rep=rep1&type=pdf. [Accessed: 29- Nov- 2017].

[6] "The Area Under an ROC Curve", Gim.unmc.edu, 2017. [Online]. Available: http://gim.unmc.edu/dxtests/roc3.htm. [Accessed: 29- Nov- 2017].