Brief summary of breast cancer data

Gloria Hu

November 2019

1 Introduction

The code and associated analysis are contained in **Beast Cancer Data Analysis Part 1** and **Beast Cancer Data Analysis Part 2**.

2 Comparing Performance Measures

Classifier	Accuracy	AUROC	AUPRC	f1 score	log-loss
Logistic Regression	0.930	0.980	0.989	0.928	0.294
KNN	0.920	0.979	0.988	0.924	0.165
Decision Tree	0.870	0.868	0.930	0.869	4.545
Random Guess	0.6	0.5	0.63		

Table 1: Results for different classifiers and measuring performance

By visualisation of the data set, it is expected almost all classifiers would perform well, as data is well separated. Therefore, all scores are relatively high. All performance measures, except the log-loss score, indicate the logistic regression behaves the best. Log-loss score indicates KNN is the best classifier. However, there is no much difference between the performance of these five performing measures. Thus, we can conclude that for a "easy" data set, the difference between AUROC or AUPRC is negligible.