Homework 3

EEE 4774 & 6777 Data Analytics

1 Classification Experiment

Download "breast-cancer-wisconsin.dat". This dataset contains 9 attributes and a binary response variable, which is an indicator for breast cancer, for each of the 683 patients. The first column in the data holds the patients' ID numbers, which is *not* considered as an attribute. The last column holds the response variable, which takes the values 2 for healthy and 4 for sick. You can normalize them to 0 and 1, respectively. The remaining 9 columns are used as the attributes. Since the scale is the same for all attributes, you don't need to normalize them. You can find more information in the "breast-cancer-wisconsin.txt" file. Use the first 400 instances as the training set, and the remaining ones as the test set. *(Note: You can use built-in functions.)*

a) [10 pts] Use logistic regression to classify the test data. Report the misclassification rate (# misclassified test instances/# all test instances) and the root-mean-squared (RMS) margin around the decision probability 0.5, i.e.,

$$\text{RMS} = \sqrt{\frac{1}{M_c}\sum_{i=1}^{M_c}(p_i - 0.5)^2},$$

where M_c is the number of correctly classified instances, and p_i denotes the probability parameter for the correctly classified instances in the model.

b) [10 pts] Classify using the standard kNN classifier. Plot the misclassification rate as a function of the number of neighbors for $k = 1, \ldots, 10$.

c) [10 pts] Use linear discriminant analysis (LDA) to classify, and report its misclassification rate.

d) [10 pts] Classify using SVM with the following kernels: linear, poly (degrees from 2 to 5), rbf, and sigmoid. Report the misclassification rate and RMS value as in part a. Compare the RMS value with that in part a. You can get the class probability estimates by setting the probability parameter to True in scikit-learn.

e) [20 pts] Use decision tree classifiers with maximum depths $3, \ldots, 10$. Plot the misclassification rate as a function of maximum tree depth. Visualize the decision rule of the tree with maximum depth 5.

f) [20 pts] Use AdaBoost, XGBoost, and Random Forest classifiers. Try $100, 200, 300, 400, 500$ weak learners (decision trees) for each algorithm and report the misclassification rates.

g) [20 pts] Compare the classification algorithms in the previous parts in terms of misclassification rate and the following weighted cost:

$$C = \frac{1}{6916} \sum_{i=1}^{283} c_i,$$

where c_i is the cost of the decision for instance i in the test set. 283 is the total number of test instances, and 6916 is the maximum cost when all instances are misclassified with respect to the following decision costs. Use zero cost for correctly classified instances, cost 1 for healthy instance misclassified as sick, and cost 100 for sick instance misclassified as healthy. Since there are 216 healthy and 67 sick instances in the test set, the maximum cost is $67 \times 100 + 216 = 6916$. Which algorithm is the best in terms of misclassification rate and the considered weighted cost?