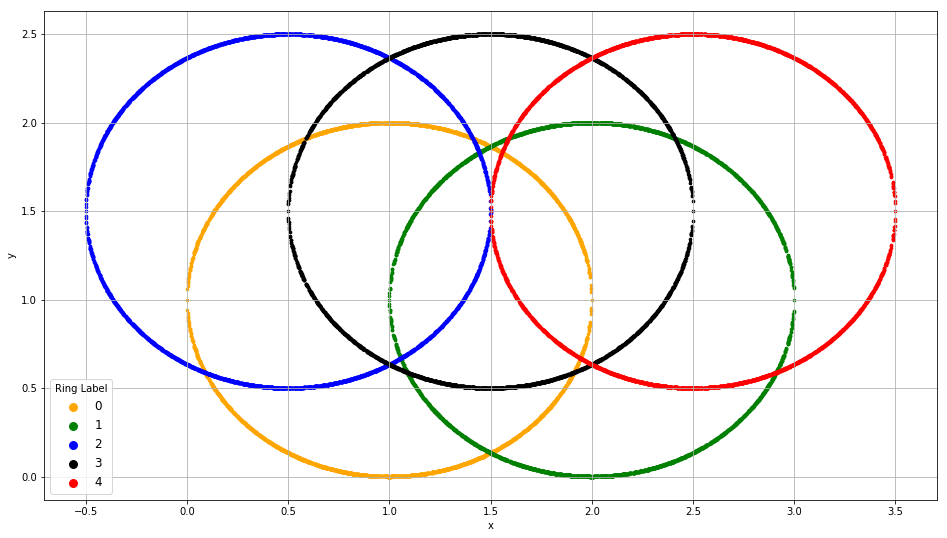
CS 584-04: Machine Learning

Spring 2019 Assignment 5

# The Five Rings Data

1. You will analyze the FiveRing.csv data for all the questions.
2. This data has 20,010 observations and three numeric fields, namely, x, y, and ring.
3. The fields x and y are the x-coordinate and the y-coordinate of the rings respectively.
4. The field ring indicates to which ring the coordinates belong.
5. The rings are labelled 0, 1, 2, 3, and 4.
6. The graph below shows the five rings.



# Misclassification Rate

Let be the predicted probability that the *i*-th observation will belong to the *j*-th ring. The predicted ring for the *i*-th observation is the smallest ring label which has the highest predicted probability. The following examples illustrate how the predicted ring is determined.

* Suppose , then the predicted ring is 3 because is the highest value among the five probabilities.
* Suppose , then the predicted ring is 0. Although are tied for the highest probability, the smallest ring label is 0.

An observation is *misclassified* if the predicted ring label is different from the observed ring label. The Misclassification Rate is the proportion of all the observations which are misclassified.

# Root Average Squared Error (RASE)

The Root Average Squared Error is

where

* is the number of observations.
* if the ring label of the *i*-th observation is *j*. Otherwise, .

# Question 1 (100 points)

You will build the multinomial logistic model according to the specifications below. You will use the Misclassification Rate and the Root Average Squared Error to assess the performance of your model.

* The nominal target variable is ring
* The predictors are x and y.
* The model will have the Intercept terms.
* The maximum number of iterations is 1000.

Build and assess the multinomial logistic model using all 20,010 observations **without** bagging and answer the following questions.

1. (10 points). List the parameter estimates (round to four decimal places) in a table. The rows are the Intercept, the predictor x, and the predictor y. The columns are the ring labels.
2. (10 points). What is the Misclassification Rate?
3. (10 points). What is the Root Average Squared Error?
4. (10 points). Redraw the above picture (i.e., the field y on the vertical axis and the field x on the horizontal axis), however, use the predicted ring label for coloring. The coloring scheme is 0 = orange, 1 = green, 2 = blue, 3 = black and 4 = red.

Apply the Bagging technique, build and assess the multinomial logistic model using all 20,010 observations. The initial random seed is 20190430. Try number of bootstraps equals to 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100.

**Note: you are not allowed to use any functions (e.g., BaggingClassifier) in the sklearn.ensemble module to perform Bagging. Instead, you must write your Python codes to implement the Bagging algorithm**.

1. (40 points). List the Misclassification Rate and the Root Average Squared Error of the bootstrap results. The columns are the two metrics. The rows are the number of bootstraps. Also, include the no-bootstrap (i.e., zero number of bootstrap) metrics.
2. (10 points). Redraw the above picture (i.e., the field y on the vertical axis and the field x on the horizontal axis), however, use the predicted ring label for coloring. The coloring scheme is 0 = orange, 1 = green, 2 = blue, 3 = black and 4 = red. There should be ten pictures, one for each set of bootstraps.
3. (10 points). Compare the results between the bagging results and the non-bagging results. Briefly comment on the comparison.

# Question 2 (100 points)

You will build the classification tree model and then apply the Adaptive Boosting technique. You will use the Misclassification Rate and the Root Average Squared Error to assess the performance of your model. The classification tree model should be built according to the specifications below.

* The nominal target variable is Ring
* The predictors are x and y.
* The splitting criterion is Entropy
* The maximum depth is 2.
* The random state value is 20190415

Build and assess the classification tree model using all 20,010 observations **without** boosting and answer the following questions.

1. (10 points). What is the Misclassification Rate?
2. (10 points). What is the Root Average Squared Error?
3. (10 points). Redraw the above picture (i.e., the field y on the vertical axis and the field x on the horizontal axis), however, use the predicted ring label for coloring. The coloring scheme is 0 = orange, 1 = green, 2 = blue, 3 = black and 4 = red.

Build and assess the classification tree model using all 20,010 observations **with** boosting with initial random seed 20190430. Try the maximum number of iterations equals to 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000. The case weights are determined as follows.

* The case weights are initialized to 1.
* After each iteration, the case weight is the for a misclassified observation. The case weight is for a correctly classified observation.

The iteration stops if either the Misclassification Rate is zero or the maximum number of iterations is reached. The aggregated predicted probabilities from a set of boosting are the weighted mean of the predicted probabilities of the iterations. The weights are the accuracy (i.e., one minus the misclassification rate) of the iterations.

**Note: you are not allowed to use any functions (e.g., AdaBoostClassifier) in the sklearn.ensemble module to perform boosting. Instead, you must write your Python codes to implement the Boosting algorithm**.

1. (50 points). List the Misclassification Rate and the Root Average Squared Error of the aggregated boosting results. The columns are the number of iterations performed and the two metrics. The rows are the maximum number of iterations. Also, include the no-boosting metrics.
2. (10 points). Redraw the above picture (i.e., the field y on the vertical axis and the field x on the horizontal axis), however, use the predicted ring label for coloring. The coloring scheme is 0 = orange, 1 = green, 2 = blue, 3 = black and 4 = red. There should be ten pictures, one for each set of boosting.
3. (10 points). Compare the results between the boosting results and the non-boosting results. Briefly comment on the comparison.