CS 109a - Milestone #4

Alex Lin Melissa Yu

November 28, 2016

1 Possible Strangeness Functions

Let Z be the set of points already examined and let z_n be a newly added point. Given a point $z_i \in Z \cup \{z_n\}$, we wish to calculate its strangeness α_i. There are a wide variety of strangeness functions that we can consider, each with its own pros and cons. In particular, any classification method can be used in our strangeness calculations. We give a few examples here:

1.1 Nearest-Neighbors

This is the most widely used method, as it is found in two out of the three papers. Let $z_i = (x_i, y_i)$ for all i, where x_i represents the features and y_i represents the class of point i. (Note that for us, y_i will be the stock symbol of the data). The Nearest Neighbors strangeness function is presented as

$$\alpha_i = \frac{\min_{j \neq i : y_j = y_i} d(x_i, x_j)}{\min_{j \neq i : y_j \neq y_i} d(x_i, x_j)}$$

It is a ratio of the distance to the closest in-class point over the distance to the closest out-of-class point. Thus, high strangeness values will involve a point x_i being farther from points in its own class relative to points in other classes.

1.2 Support Vector Machines

This method is found in the final of the three papers. Given a series of points $(x_1, y_1) \dots (x_n, y_n)$, we use SVM to draw a separating hyperplane between the class y_i of the point i in question and all other classes. Then, we define α_i as simply the distance between x_i and the separating hyperplane. Again, high strangeness values will involve a point x_i being far from the hyperplane.

Other possibilities include logistic regression and decision trees. However, the methods by which we can compute strangeness values are not clear for these cases.