CZ4041: Tutorial Week 9

Due on March 18, 2021 at 8:30am

Assoc Prof Pan, Sinno Jialin - CS4

Pang Yu Shao
U1721680D

18/03/2021

Problem 1

Why the condition that the base classifiers should do better than a classifer that performs random guessing is necessary for ensemble learning?

Solution
Ensemble classifiers perform well when the base classifiers do better than random guessing (i.e., having a error rate of $< 50\%$). This is because the ensemble classifier only makes a wrong prediction if the majority of the base classifiers predict incorrectly.

When the base classifers have an error rate of more than 50%, this results in a higher chance of the majority of the classifiers making a wrong prediction – therefore this causes the ensemble classifier to perform worse than the base classifiers.

$$P(N) = \sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$$

In the example given in the lecture notes, N = 3, and error rate = 0.35

$$P(3) = \binom{3}{2} * 0.35^2 * (1-0.35)^1 + \binom{3}{3} * 0.35^3$$

$$= 0.28175$$

This performs better than the base classifier, since the error rate of 28% is less than that of 35%.
However, if the base classifier has an error rate of 0.8, we get the following:

$$P(3) = \binom{3}{2} * 0.8^2 * (1-0.8)^1 + \binom{3}{3} * 0.8^3$$

$$= 0.896$$

Therefore, the base classifier should perform better than random guessing in order for ensemble learning to be effective.

Problem 2

Suppose we have trained 5 base binary classifiers: f_1, f_2, f_3, f_4 and f_5. Their predictions on a validation dataset are shown in Table 1, where the last column denotes the ground-truth class labels. Which base classifiers would you choose to construct an ensemble learner.

Table 1: Data set for Question 2.

ID	f_1	f_2	f_3	f_4	f_5	Ground Truth
P1	+	+	-	-	+	+
P2	+	+	-	+	-	+
P3	-	-	+	+	-	+
P4	-	-	+	-	+	+
P5	-	-	+	+	-	-
P6	-	-	-	+	+	+
P7	+	+	+	+	-	+
P8	-	+	+	-	+	-
P9	+	+	-	+	+	+
P10	-	-	-	+	-	-

Solution

First, get the prediction accuracy of the base classifiers on the dataset:

$f_1 : 7/10 = 70\%$

$f_2 : 6/10 = 60\%$

$f_3 : 4/10 = 40\%$

$f_4 : 6/10 = 60\%$

$f_5 : 6/10 = 60\%$

Since f_3 has a prediction accuracy of ¡50% it will be excluded from the ensemble learner.

It can also be seen that f_1 and f_2 have the same predictions for 9 out of 10 of the samples. Therefore, it can be suggested that these two classifiers are highly co-related. We exclude f_2 from the ensemble since it has a lower prediction accuracy.

Therefore, the ensemble classifier would be built from f_1, f_4 and f_5.