Introduction to Machine Learning, Fall 2014 - Exercise session V

Rodion “rodde” Efremov
013593012

November 23, 2014

Problem 1 (3 points)

Consider the following two sets of 80 cars each: Set ‘A’ consists of 10 Volvos, 25 Toyotas, and 45 Audis, while set ‘B’ consists of 8 Volvos, 32 Toyotas, and 40 Audis. Which set do you intuitively think is more pure (that is, has lower impurity), and why? Compute the entropy, the Gini index, and the misclassification error for each of the two sets. According to these measures, which set is more pure? Could this phenomenon (conflict among the measures) happen if there were just two classes (two types of car) rather than three? Why, or why not?

To me, it seems that the set ‘A’ is more pure than ‘B’, for “intervals” in set ‘A’ are more “equal” than in the set ‘B’.

The entropy of the set A is ($\log = \log_2$)

$$
\begin{aligned}
& -\frac{10}{80}\log\frac{10}{80} - \frac{25}{80}\log\frac{25}{80} - \frac{45}{80}\log\frac{45}{80} \\
&= \frac{1}{8}(\log 80 - \log 10) + \frac{5}{16}(\log 80 - \log 25) + \frac{9}{16}(\log 80 - \log 45) \\
&\approx 0.375 + 0.524 + 0.467 \\
&\approx 1.366.
\end{aligned}
$$

The entropy of the set B is

$$
\begin{aligned}
& -\frac{8}{80}\log\frac{8}{80} - \frac{32}{80}\log\frac{32}{80} - \frac{40}{80}\log\frac{40}{80} \\
&= \frac{1}{10}\log 10 + \frac{4}{10}(\log 80 - \log 32) + \frac{1}{2}\log 2 \\
&\approx 0.332 + 0.529 + 0.5 \\
&\approx 1.361.
\end{aligned}
$$

The Gini index of the set A is

$$\begin{aligned}
&1-\left(\frac{10}{80}\right)^2-\left(\frac{25}{80}\right)^2-\left(\frac{45}{80}\right)^2\\
&=1-(1/8)^2-(5/16)^2-(9/16)^2\\
&\approx 1-0.016-0.098-0.316\\
&\approx 0.570.
\end{aligned}$$

The Gini index of the set B is

$$\begin{aligned}
&1-\left(\frac{8}{80}\right)^2-\left(\frac{32}{80}\right)^2-\left(\frac{40}{80}\right)^2\\
&\approx 1-0.01-0.16-0.25\\
&\approx 0.58.
\end{aligned}$$

The classification error of the set A is

$$1-\frac{45}{80}=1-\frac{9}{16}=0.4375.$$

The classification error of the set B is

$$1-\frac{40}{80}=0.5.$$

In summary: According to the above table, the set A is slightly more pure

	Entropy	Gini index	Classification error
A	1.366	0.570	0.438
B	1.361	0.580	0.5

than B.

Now, let us restrict the amount of car types only to two. (Say, Lada and Rolls Royce.) Let N be the total amount of cars, and let L denote the amount of Lada's (so that we have $N-L$ RR's). Now the entropy of such a set is

$$\begin{aligned}
&-\frac{L}{N}\log\frac{L}{N}-\frac{N-L}{N}\log\frac{N-L}{N}=\\
&\frac{L}{N}\log\frac{N}{L}+\left(1-\frac{L}{N}\right)\log\frac{N}{N-L}=\\
&\frac{L}{N}\log N-\frac{L}{N}\log L+\left(1-\frac{L}{N}\right)\left(\log N-\log(N-L)\right)=\\
&-\frac{L}{N}\log L+\log N-\log(N-L)+\frac{L}{N}\log(N-l)=\\
&a
\end{aligned}$$

Problem 2 (3 points)

Problem 3 (3 points)

Problem 4 (15 points)