Introduction to Machine Learning, Fall 2015 - Exercise session IV

Rodion “rodde” Efremov, student ID 013593012

November 20, 2015

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) rathere than three? Why, or why not?

Solution

It seems like the set ’A’ is more pure than ’B’, as in the set ’B’ we have 32 Toyotas, and 40 Audis. The entropy of A is

$$\begin{aligned}
&-\frac{10}{80}\log_2\frac{10}{80}-\frac{25}{80}\log_2\frac{25}{80}-\frac{45}{80}\log_2\frac{45}{80}\\
&=\frac{1}{8}(\log_2 80-\log_2 10)+\frac{5}{16}(\log_2 80-\log_2 25)+\frac{9}{16}(\log_2 80-\log_2 45)\\
&\approx 0.375+0.524+0.467\\
&\approx 1.367.
\end{aligned}$$

The entropy of B is

$$\begin{aligned}
&-\frac{8}{80}\log_2\frac{8}{80}-\frac{32}{80}\log_2\frac{32}{80}-\frac{40}{80}\log_2\frac{40}{80}\\
&=0.1\log_2 10+0.25(\log_2 80-log_2 32)+0.5\log_2 2\\
&\approx 0.332+0.529+0.5\\
&\approx 1.361.
\end{aligned}$$

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

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

The classification error of the set A is

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

The classification error of the set B is

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

Summarizing:

	Entropy	Gini index	Classification error
A	1.367	0.570	0.438
B	1.361	0.580	0.5

It is hard to tell which set is more pure as the impurity measures contradict each other. Suppose a set contains N elements, M of which belong to one category and the other $N - M$ to the other one. Now, all the impurity measures will attain their maximum value at around $M = \frac{N}{2}$, and all measures will be monotonically increasing on $(0, \frac{N}{2})$, and monotonically decreasing on $(\frac{N}{2}, N)$, so improvement in, say, Gini index will lead to improvement in classification error and entropy, and vice versa.

Problem 2 (3 points)

Consider a binary classification problem with the following set of attributes and attribute values:

- Air conditioner = { Working, Broken }
- Engine = { Good, Bad }

- Mileage = { High, Medium, Low }
- Rust = { Yes, No }

Suppose a rule-based classifier produces the following rule set:

- (Mileage = High) ⇒ (Value = Low)
- (Mileage = Low) ⇒ (Value = High)
- ((Air conditioner = Working) and (Engine = Good)) ⇒ (Value = High)
- ((Air conditioner = Working) and (Engine = Bad)) ⇒ (Value = Low)
- (Air conditioner = Broken) ⇒ (Value = Low)

Given the above, answer the following:

(a) Is the rule set consistent?

(b) Is the rule set complete?

(c) Is the meaning of the rule set clear if you consider it as unordered?

(d) If you consider this as an ordered rule list, could the meaning be affected if the order is changed?

(e) Do you need a default rule?

Solution

(a) Yes, it is consistent as all instances are covered only by a single rule.

(b) No, it is not as instance involving { Rust = ... } is not covered by any rule in the set.

(c) Yes, it is.

(d) No, the meaning will not change upon permuting the rules.

(e) Since instances involving rust are not classified, we need a default rule.