Introduction to Machine Learning, Fall 2014 - Exercise session V

Rodion "rodde" Efremov
013593012

November 24, 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.

Suppose a set contains N elements in total, and M of them belong to one type of elements and $N - M$ to the other one. Now, all the impurity measure 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 or entropy, and vice versa. (As can be seen in a figure of course slides.)

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 conditioiner = Broken) ← (Value = Low)

Given the above, answer the following:

(a) Are the rules mutually exclusive? Yes, they are. On each record, at most one rule is triggered.

(b) Is the rule set exhaustive? Suppose that Mileage = Medium. This is not covered by two first rules, but it all boils down to the fact that Air conditioner will "catch" all the records regardless of the state of Engine. So, yes, the rule set is exhaustive.

(c) Is ordering needed for this set of rules? No, it doesn't appear that way: mileage is a good predictor of the value of a car, so it is justified to have the first two rules where they are.

(d) Do you need a default class for the rule set? No, we don't need that since the rules are exhaustive.

Problem 3 (3 points)

Given a set of N points $y_1, \ldots, y_N$, with each $y_i \in \mathbb{R}$, show that...

(a) ... the value y^* which minimizes the sum of *squared* errors, i.e.

$$y^* = \arg\min_{\hat{y}} \sum_{i=1}^{N} (y_i - \hat{y})^2$$

is given by the *mean* of the y_i, i.e. $y^* = \sum_i y_i / N$.

Suppose we have

$$\sum_{i=1}^{N} (y_i - t)^2 = \sum_{i=1}^{N} (y_i^2 - 2y_i t + t^2).$$

Now the first derivative of the above sum with respect to t is given by

$$\sum_{i=1}^{N} 2t - 2y_i,$$

and the second derivative of the same sum is given by

$$\sum_{i=1}^{N} 2 = 2N > 0.$$

The original sum is, therefore, minimized (2nd derivative > 0) at point, where the first derivative is zero, i.e.

$$\begin{aligned}
\sum_{i=1}^{N} 2t - 2y_i &= 0 \\
\sum_{i=1}^{N} t - y_i &= 0 \\
Nt &= \sum_{i=1}^{N} y_i \\
t &= \frac{\sum_{i=1}^{N} y_i}{N},
\end{aligned}$$

which proves that the mean of the y_i minimizes the sum of squared errors.

(b) ... the value y^* which minimizes the sum of *absolute* errors, i.e.

$$y^* = \arg\min_{\hat{y}} \sum_{i=1}^{N} |y_i - \hat{y}|$$

is given by the *median* of the y_i. [Hint for part (b): Break the sum into parts corresponding to *pairs* of observations, pairing the smallest with the largest point, the second-smallest with the second-largest point, etc.]

Let us rephrase the sum. If N is even, we have

$$\sum_{i=1}^{N} |y_i - \hat{y}| = \sum_{i=1}^{N/2} |y_i - \hat{y}| + |y_{N+1-i} - \hat{y}|$$

Problem 4 (15 points)