SIF8015 Logic

Exercise 4, Solutions

Applied Logic: Data Mining and Knowledge Discovery

Task 1

(a) The equivalence classes with respect to the condition attributes are:

$$\begin{aligned}
E_1 &= [\text{Smith}] \\
&= \{\text{Smith}\} \\
E_2 &= [\text{Jones}] \\
&= \{\text{Jones}\} \\
E_3 &= [\text{Parker}] \\
&= [\text{Hanson}] \\
&= \{\text{Parker}, \text{Hanson}\} \\
E_4 &= [\text{Moore}] \\
&= [\text{Starr}] \\
&= \{\text{Moore}, \text{Starr}\} \\
E_5 &= [\text{Fields}] \\
&= \{\text{Fields}\}
\end{aligned}$$

(b) The values for the generalized decision attribute are:

$$\begin{aligned}
G_{\text{Walk}}(E_1) &= \{\text{Yes}\} \\
G_{\text{Walk}}(E_2) &= \{\text{No}\} \\
G_{\text{Walk}}(E_3) &= \{\text{No}, \text{Yes}\} \\
G_{\text{Walk}}(E_4) &= \{\text{No}\} \\
G_{\text{Walk}}(E_5) &= \{\text{Yes}\}
\end{aligned}$$

Since $|G_{\text{Walk}}(E_3)| > 1$, the table is inconsistent.

(c) The discernibility matrix with respect to the condition attributes is:

$$\begin{array}{c}
E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5
\end{array}
\begin{bmatrix}
\emptyset & \{\text{L}\} & \{\text{A}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \{\text{S}, \text{L}\} \\
\{\text{L}\} & \emptyset & \{\text{A}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \{\text{S}, \text{L}\} \\
\{\text{A}, \text{L}\} & \{\text{A}, \text{L}\} & \emptyset & \{\text{A}, \text{S}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} \\
\{\text{A}, \text{S}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \emptyset & \{\text{A}\} \\
\{\text{S}, \text{L}\} & \{\text{S}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \{\text{A}\} & \emptyset
\end{bmatrix}$$

The discernibility matrix computed modulo the decision attribute is obtained by setting entries between objects with the same generalized decision value to the empty set:

$$\begin{array}{cc} \begin{array}{c} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} & \left[\begin{array}{ccccc} \emptyset & \{\text{L}\} & \{\text{A}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \emptyset \\ \{\text{L}\} & \emptyset & \{\text{A}, \text{L}\} & \emptyset & \{\text{S}, \text{L}\} \\ \{\text{A}, \text{L}\} & \{\text{A}, \text{L}\} & \emptyset & \{\text{A}, \text{S}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} \\ \{\text{A}, \text{S}, \text{L}\} & \emptyset & \{\text{A}, \text{S}, \text{L}\} & \emptyset & \{\text{A}\} \\ \emptyset & \{\text{S}, \text{L}\} & \{\text{A}, \text{S}, \text{L}\} & \{\text{A}\} & \emptyset \end{array}\right] \end{array}$$

(d) The desired discernibility function is obtained by running over the lower (or upper) diagonal of the discernibility matrix computed modulo the decision attribute:

$$\begin{aligned} f(\{E_1, \ldots, E_5\}) &= (\text{LEMS}) \wedge \\ &\quad (\text{Age} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age} \vee \text{Sex} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age} \vee \text{Sex} \vee \text{LEMS}) \wedge \\ &\quad (\text{Sex} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age} \vee \text{Sex} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age}) \\ &= (\text{Age} \wedge \text{LEMS}) \end{aligned}$$

Attribute Sex is hence not needed to determine the outcome for any person in the decision table. However, this does not mean that attribute Sex can never be used, just that in those cases where Sex can be used (alone or together with other attributes), alternatives exists that do not involve the Sex attribute.

(e) The desired discernibility function is obtained by running over the row (or column) that corresponds to Ms. Fields in the discernibility matrix computed modulo the decision attribute:

$$\begin{aligned} f(E_5) &= (\text{Sex} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age} \vee \text{Sex} \vee \text{LEMS}) \wedge \\ &\quad (\text{Age}) \\ &= (\text{Age} \wedge \text{Sex}) \vee \\ &\quad (\text{Age} \wedge \text{LEMS}) \end{aligned}$$

Ms. Fields thus gives rise to the following two minimal decision rules:

if (Age = 16-30) and (Sex = Female) then (Walk = Yes)
if (Age = 16-30) and (LEMS = 26-49) then (Walk = Yes)

(f) Note that there are three persons in the decision table that match the if-part of the rule, while there are four persons that match the then-part of the rule. There is only one person that matches both the if-part and the then-part simultaneously.

$$\text{Accuracy} = 1/3$$
$$\text{Coverage} = 1/4$$

Hence, the probability that the then-part of the rule is correct given that the if-part matches is 0.333, while the probability that the if-part of the rule matches given that the then-part is correct is 0.25. The rule is in other words not a very good rule.

Task 2

(h) Obviously, the performance estimate is a function of the split. Different splits may result in different performance estimates.

One way of obtaining a more reliable performance estimate is through a process called cross-validation: In k-fold cross-validation one partitions the data into k disjoint “blocks”. A classifier is then induced using $k-1$ of the blocks, and the induced model is subsequently applied to the remaining hold-out block to obtain a performance estimate. This is done k times with each block being used once as a hold-out block, and an average performance estimate can then be computed along with an estimate of its variability.

(i) Accuracy may not be a good performance measure if the costs of making different types of classification errors are unequal. For example, in a hypothetical apple-grenade sorting problem, it would be far more dangerous to classify a grenade as an apple than to classify an apple as a grenade.

Secondly, accuracy may not be a very suitable performance measure if the distribution of cases from each class is very skewed. For instance, an accuracy of 99% may sound impressive, but if we are trying to detect a rare event and 99% of all examples are known to belong to a particular decision class, then an accuracy of 99% does not really tell us very much.