Aufgabe 5-1  Nearest neighbor classification

The 2D feature vectors in the figure below belong to two different classes (circles and rectangles). Classify the object at $(6, 6)$ — in the image represented using a triangle — using $k$ nearest neighbor classification. Use Manhattan distance ($L_1$ norm) as distance function, and use the non-weighted class counts in the $k$-nearest-neighbor set, i.e. the object is assigned to the majority class within the $k$ nearest neighbors. Perform $k$NN classification for the following values of $k$ and compare the results with your own “intuitive” result.

(a) $k = 4$

(b) $k = 7$

(c) $k = 10$

Aufgabe 5-2  Decision trees

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1 - 2$ years, $2 - 7$ years, $> 7$ years)
- Gender (male, female)
- Residential area (urban, rural)
For your analysis you have the following manually classified training examples:

<table>
<thead>
<tr>
<th>Person</th>
<th>Time since license</th>
<th>Gender</th>
<th>Area</th>
<th>Risk class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1−2</td>
<td>m</td>
<td>urban</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>2−7</td>
<td>m</td>
<td>rural</td>
<td>high</td>
</tr>
<tr>
<td>3</td>
<td>&gt;7</td>
<td>f</td>
<td>rural</td>
<td>low</td>
</tr>
<tr>
<td>4</td>
<td>1−2</td>
<td>f</td>
<td>rural</td>
<td>high</td>
</tr>
<tr>
<td>5</td>
<td>&gt;7</td>
<td>m</td>
<td>rural</td>
<td>high</td>
</tr>
<tr>
<td>6</td>
<td>1−2</td>
<td>m</td>
<td>rural</td>
<td>high</td>
</tr>
<tr>
<td>7</td>
<td>2−7</td>
<td>f</td>
<td>urban</td>
<td>low</td>
</tr>
<tr>
<td>8</td>
<td>2−7</td>
<td>m</td>
<td>urban</td>
<td>low</td>
</tr>
</tbody>
</table>

(a) Construct a decision tree based on this training data. For splitting, use information gain as measure for impurity. Build a separate branch for each attribute. The decision tree shall stop when all instances in the branch have the same class, you do not need to apply a pruning algorithm.

(b) Apply the decision tree to the following drivers:
   - Person A: 1-2, f, rural
   - Person B: 2-7, m, urban
   - Person C: 1-2, f, urban

Aufgabe 5-3 Information gain

In this exercise, we want to look more closely at the information gain measure.

Let $T$ be a set of $n$ training objects with the attributes $A_1, \ldots, A_a$ and the $k$ classes $c_1$ to $c_k$.

Let $\{T^A_i | i \in \{1, \ldots, m_A\}\}$ be the disjoint, complete partitioning of $T$ produced by a split on attribute $A$ (where $m_A$ is the number of disjoint values of $A$).

(a) Uniform distribution
   Compute $\text{entropy}(T)$, $\text{entropy}(T^A_i)$ for $i \in \{1 \ldots m_A\}$ as well as $\text{information-gain}(T, A)$ given the assumption that the class membership of $T$ is uniformly distributed and independent of the values of $A$. Interpret your result!

(b) Additional uniform distribution
   We want to analyze how the number of different values influences the information gain. For this, we compare two attributes, attribute $A$ with $m_A$ values and attribute $A'$ with $m_A' = m_A + 1$ values, where the relative frequencies in $A'$ in values 1 to $m_A$ are identical to that of $A$ and in the additional value $m_A'$ there is a uniform distribution of the classes.
   How does $\text{information-gain}(T, A)$ differ from $\text{information-gain}(T, A')$? Interpret your result!

(c) Attributes with many values
   Let $A$ be an attribute with random values, not correlated to the class of the objects. Furthermore, let $A$ have enough values, such than no two instances of the training set share the same value of $A$. What happens in this situation when building the decision tree? What is problematic with this situation?