Data Mining and Machine Learning

Practice Assignment 1, II Semester, 2024–2025

3 February, 2025

1. A number of new vaccines are being deployed to treat a recently discovered disease. Reports are emerging of patients having side effects caused by vaccinations. Some side effects are vaccine-specific, some occur across vaccines.

   For each reported case, there is information available about the nature of the side effect, the vaccine used, demographic details about the patient (age, gender, race, ...) as well as information about prevailing health conditions of the patient (diabetes, hypertension, ...) that may create complications.

   Explain how market-basket analysis can help doctors determine risk factors associated with vaccinations, in general, and specific vaccines, in particular.

2. In the market-basket analysis problem, suppose the set of items I has size 10^7, the number of transactions T is 10^{10} and each transaction $t \in T$ contains at most 10 distinct items. Compute upper bounds for F_1 and F_2, the number of frequent itemset of size 1 and 2, respectively, for a support value of 0.1%.

3. You are building a decision tree on tabular data with attributes $\{A_1, A_2, \ldots, A_7\}$ where $\{A_3, A_5\}$ are numeric and the other five attributes are categorical. Attribute A_3 takes integer values in the range $[-100, 100]$ and attribute A_5 takes integer values in the range $[1, 10000]$. There are 2000 items in the training set. You adopt a pre-pruning strategy to build the tree where you do not split any node with 50 items or fewer. Across all possible decision trees that can be built on this training set, what is the maximum height of the resulting tree? Explain your answer.

4. The algorithm we described to build a decision tree is deterministic. However, we saw that the decision tree library implemented in Python's `sklearn` library uses a random seed. Why should a random seed be needed? (Hint: Consider the example from the iris dataset.)

5. An airport security system consists of a full body scanner followed by manual frisking. If the full body scanner beeps, the passenger is checked manually and then allowed to proceed if there is nothing amiss. If the full body scanner does not beep, no frisking is done.

   (a) In terms of the entries in the confusion matrix, what ratio should the full body scanner maximize to ensure that no suspicious person is let through unchecked?

   (b) Similarly, what ratio should manual frisking maximize?