Lazy FCA Assignment, Ilia Pakhalko

In this project we will apply Lazy FCA method for classification of binary data, as well as study the impact of minimal cardinality and number of bins for feature binarization on Accuracy, F1 score and Prediction time (relative to training sample size).

Data Overview

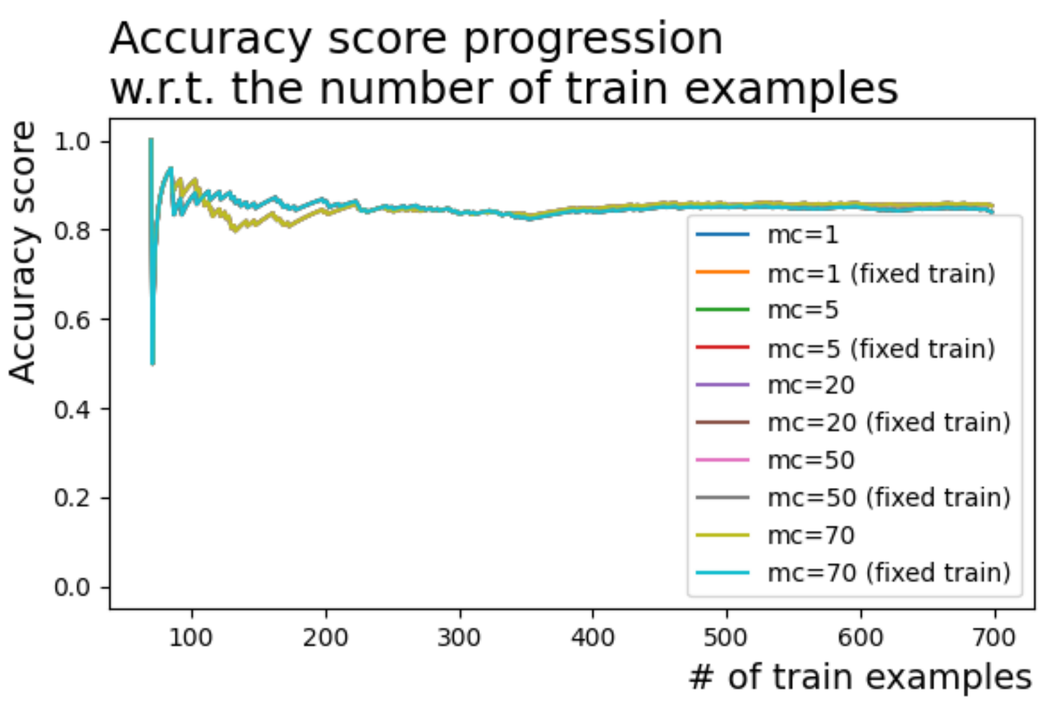
For the purposes of this task we've chosen to adopt the Wisconsin Breast Cancer dataset, made available publicly online via the UCI website: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ . Each record in the dataset contains the information about a certain tissue fragment; it is described by its sample id, the target class - whether the cell in question is benign or malignant, and 9 ordinal features: each on the scale from 1 to 10, containing such information as the thickness of the tissue, uniformity of cell shape and size, as well as the amount of bare and normal nucleoli. Firstly, we substitute the missing values with the special class 0 - this value is not present throughout the dataset and thus can be used as a placeholder. Then we binarize each feature - in this work we experiment with two configurations: namely, we use naive one-hot encoding as our baseline, and compare it to reduced binarization, where each value is assigned into one of three bins (either "low", "middle" or "high").

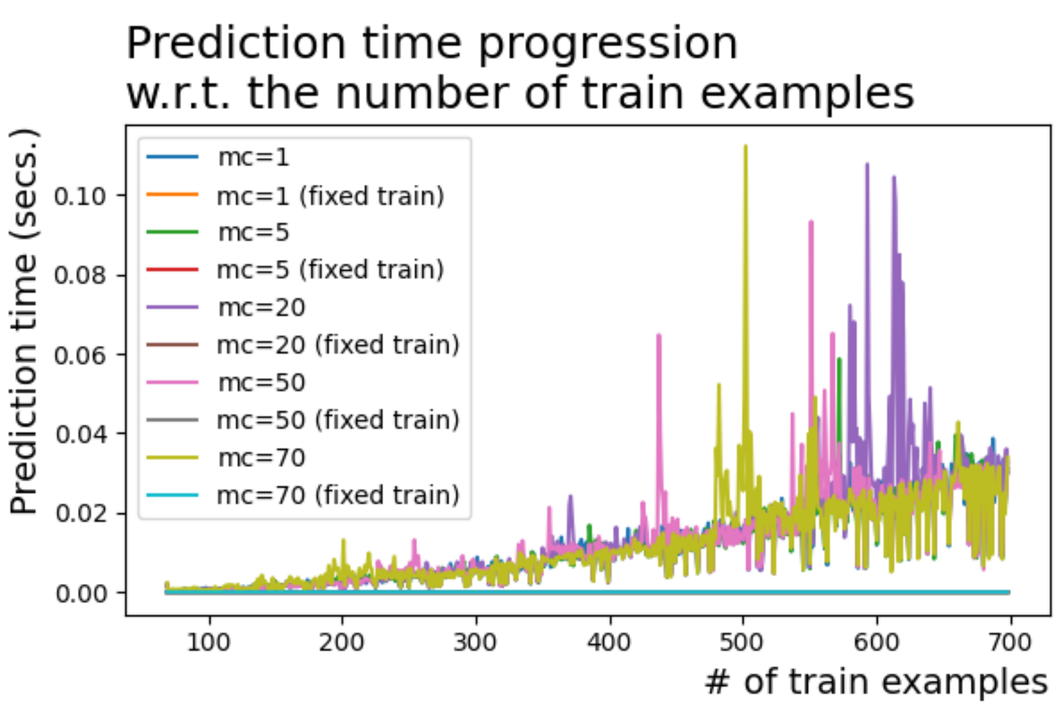
Minimal Cardinality Grid

As well as comparing coarse and fine binarization strategies, we seek to establish a relation between prediction quality and Minimal Cardinality Parameter. We start from the initial assumption that Minimal Cardinality regulates the strictness of our classification: if (binarized) feature intersection between positive (negative) sample and classified example is less than this specified threshold, then we discard the sample from contributing to overall counterexample count. Thus, we expect classification precision to correlate with Minimal Cardinality.

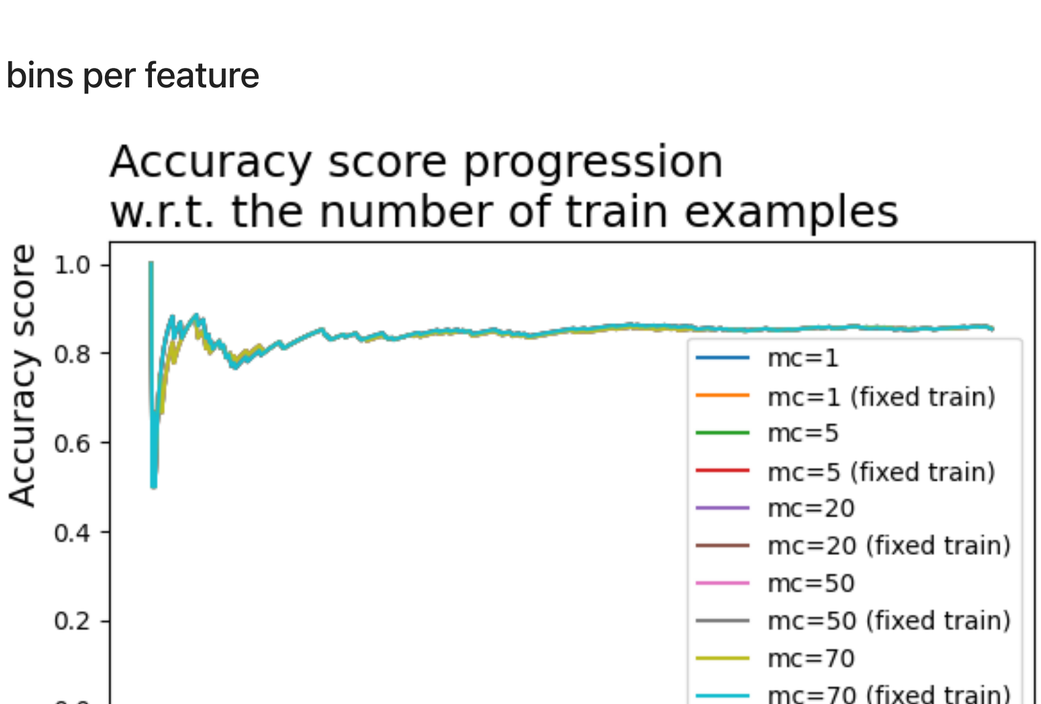
We use the `MC grid = [1, 5, 20, 50, 70]` in our experiments. Results can be observed in both figures 1 and 2. Namely, we do not see any correlation with Accuracy or F1 Score whatsoever, they seem to be perfectlu alligned. In the meantime, we see a difference between prediction times, though no clear trend is visible, so this difference is more likely to be attributed to unrelated noise.

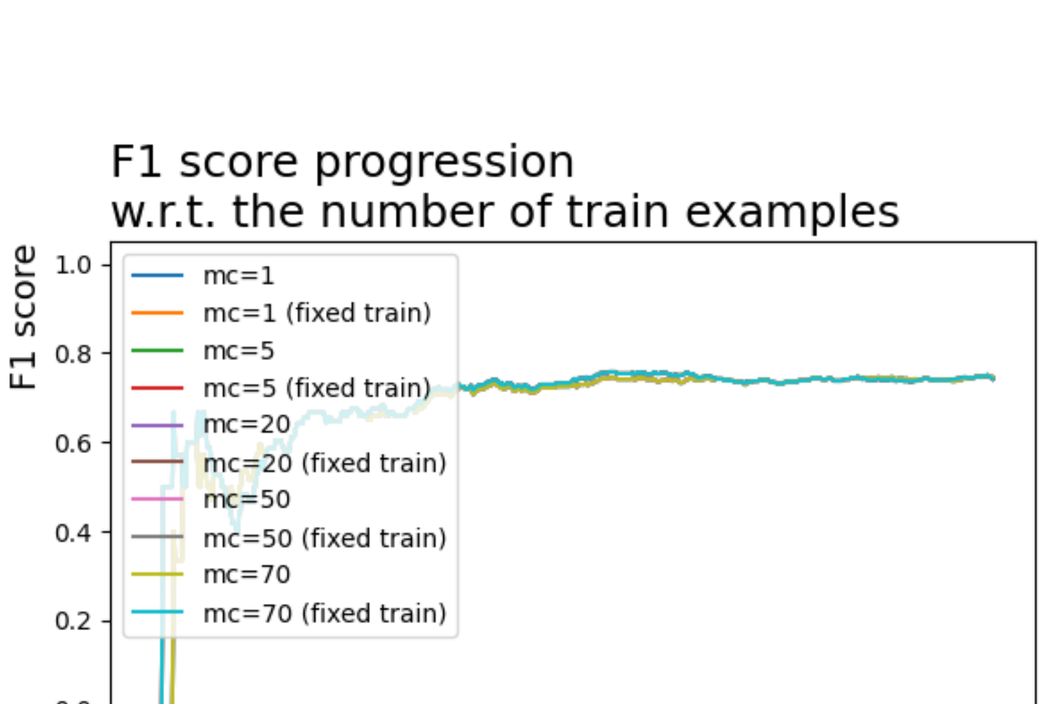
One-Hot Encoded Features

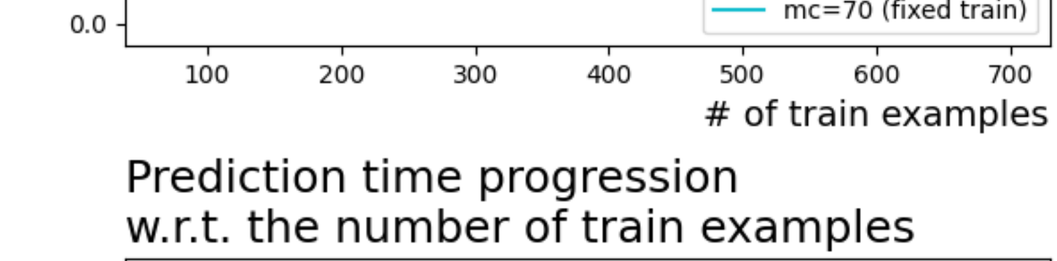




3 bins per feature







Binarization Strategy

We experiment with two binarization strategies. The first one One-Hot encodes each feature, leading to 10 creating bins per feature. This may be costly on wider datasets, so we try to reduce this number to 3 bins per feature: low, middle and high values. The results can be seen in the figures above. From the graphs it is clearly visible, that compressed binarization leads to lower classification quality (accuracy and F1 score) in lower training data setting.

In [ ]: