Data

Algorithms were tested on a part of the “Bike Buyers 1000” dataset (https://www.kaggle.com/heeraldedhia/bike-buyers), which has details of users from different backgrounds and whether or not they buy a bike. Particularly, I used 500 observations from this dataset in order to reduce processing time.

Target variable is `'Purchased Bike'`, which denotes whether a person buys a bike.

Feature description

Name	# of unique values
Marital Status	2
Gender	2
Income	16
Children	6
Education	5
Occupation	5
Home Owner	2
Cars	5
Commute Distance	5
Region	3
Age	49

“Income” and “Age” were categorized according to quartiles. All categorical features, which have more than 2 unique values, were preprocessed by one-hot-encoding method. As a result, each object has 37 features including target class.

Algorithms

All metrics for evaluating quality of an algorithm was computed based on 3-fold cross-validation.

Algorithm 1

Test object gets a vote in favor of positive (negative) class if its intersection with an object from positive (negative) context is not found in in the negative (positive) context and the size of intersection is not less than ‘min_cardinality’. Class of an object is determined by majority rule. Minimal cardinality parameter was chosen as 15 based on 3-fold validation results.

Accuracy of algorithm depending on minimal cardinality value

Min_cardinality	Mean Accuracy
5	0.613989
10	0.613989
15	0.613989
30	0.546016

Results

accuracy	**0.613989**
precision	**0.625132**
recall	**0.640180**
F1	**0.629370**
Neg_pred_rate	**0.642482**
FP_rate	**0.374199**
FN_rate	**0.359820**
F_disc_rate	**0.374868**
Relative number of contradictions	**0.025996**

Algorithm 2

Test object is intersected with each object from positive (negative) class and then support of an intersection is computed. The class of an object is determined by context’s average support value.

Results

accuracy	0.637965
precision	0.635526
recall	0.630807
F1	0.632554
Neg_pred_rate	0.637087
FP_rate	0.361166
FN_rate	0.369193
F_disc_rate	0.364474
Relative number of contradictions	0.000000

Algorithm 3

Test object is classified according to the maximal length of an intersection with objects of positive and negative contexts. So, it has the same class as an object which is most similar to it.

Results

accuracy	0.537912
precision	0.638490
recall	0.655354
F1	0.646015
Neg_pred_rate	0.659847
FP_rate	0.356602
FN_rate	0.344646
F_disc_rate	0.361510
Relative number of contradictions	0.172041

Discussion

According to the number of contradictions (and accuracy as well), algorithm 3 appeared to be the worst. However, it is the best according to recall and F1 score, which is related to the fact that it drops objects with contradiction. It is seen, that algorithm 2 leads to the lowest number of contradictions and has the highest accuracy. Algorithm 1 does not have any special results, but it is worth to be mentioned that due to its structure it requires much more time to predict one test example compare to other algorithms.