**Итоговое домашнее задание ТРАД**

**Барышев Евгений**

**Описание данных**

Для классификации были выбраны данные по подтверждению кредита:

<https://www.kaggle.com/shravan3273/credit-approval>

Датасет содержит 1000 объектов, 12 строковых, 8 числовых и 1 бинарный признак.

Целевой признак: подтверждение или отказ кредита банком.

Множество признаков:

1. **checking\_balance** - остаток на активном счете
2. **months\_loan\_duration** - длительность кредита
3. **credit\_history** - кредитная история
4. **purpose** – цель покупки
5. **amount**- размер кредита
6. **savings\_balance** – остаток на сберегательном счете
7. **employment\_length** –рабочий стаж
8. **installment\_rate** – процентная ставка
9. **personal\_status** – семейное положение и пол
10. **other\_debtors** – поручители
11. **residence\_history** – история проживания
12. **property**
13. **age** – возраст
14. **installment\_plan** – рассрочка
15. **housing** – тип жилья
16. **existing\_credits** – количество активных кредитов
17. **dependents**
18. **telephone** – указан номер телефона или нет
19. **foreign\_worker** – иностранный работник
20. **job** – тип работы
21. **default** – целевой признак. Подтверждение или отказ кредита банком

**Бинаризация данных**

preparing\_data.py

Далее представлен способ бинаризации каждого признака. Представлен график распределения значений каждого признака и таблица, согласно которой выполняется бинаризация. Для числовых признаков представлено среднее, минимальное, максимальное значение, а также распределение по процентам.

months\_loan\_duration binarize\_months\_loan\_duration\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean 21  std 12  min 4  25% 12  50% 18  75% 24  max 72 | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **mld<6** | **mld <12** | **mld <18** | **mld <24** | **mld >24** | | **0-6** | x | x | x | x |  | | **7-12** |  | x | x | x |  | | **13-18** |  |  | x | x |  | | **19-24** |  |  |  | x |  | | **>24** |  |  |  |  | x | |

checking\_balanc**e** binarize\_checking\_balance\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | unknown 394  < 0 DM 274  1 - 200 DM 269  > 200 DM 63 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **cb\_unk** | **cb <0** | **cb 1-200** | **cb >200** | | **unknown** | x |  |  |  | | **< 0 DM** |  | x |  |  | | **1 - 200 DM** |  |  | x |  | | **> 200 DM** |  |  |  | x | |

credit\_history binarize\_credit\_history\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | repaid 530  critical 293  delayed 88  fully repaid  this bank 49  fully repaid 40 | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **ch\_r** | **ch\_c** | **ch\_d** | **ch\_rb** | **ch\_fr** | | **repaid** | x |  |  |  |  | | **critical** |  | x |  |  |  | | **delayed** |  |  | x |  |  | | **FRtB** |  |  |  | x |  | | **FR** |  |  |  |  | x | |

purpose binarize\_purpose\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | radio/tv 280  car (new) 234  furniture 181  car (used) 103  business 97  education 50  repairs 22  DA 12  others 12  retraining 9 | |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | **p\_rt** | **p\_cn** | **p\_f** | **p\_cu** | **p\_b** | **p\_e** | **p\_re** | **p\_d** | **p\_o** | **p\_r** | | **radio/tv** | x |  |  |  |  |  |  |  |  |  | | **car (new)** |  | x |  |  |  |  |  |  |  |  | | **furniture** |  |  | x |  |  |  |  |  |  |  | | **car (used)** |  |  |  | x |  |  |  |  |  |  | | **business** |  |  |  |  | x |  |  |  |  |  | | **education** |  |  |  |  |  | x |  |  |  |  | | **repairs** |  |  |  |  |  |  | x |  |  |  | | **DA** |  |  |  |  |  |  |  | x |  |  | | **others** |  |  |  |  |  |  |  |  | x |  | | **retraining** |  |  |  |  |  |  |  |  |  | x | |

amount binarize\_amount\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean 3271  std 2823  min 250  25% 1365  50% 2319  75% 3972  max 18424 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **cb<1365** | **cb<2319** | **cb<3972** | **cb>3973** | | **0-1365** | x | x | x |  | | **1366-2319** |  | x | x |  | | **2320-3972** |  |  | x |  | | **>3973** |  |  |  | x | |

savings\_balance binarize\_savings\_balance\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | < 100 DM 603  unknown 183  101 - 500 DM 103  501 - 1000 DM 63  > 1000 DM 48 | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **sb\_unk** | **sb <100** | **sb 101-500** | **sb 501-1000** | **sb >1000** | | **unknown** | x |  |  |  |  | | **< 100 DM** |  | x |  |  |  | | **101 - 500 DM** |  |  | x |  |  | | **501 - 1000 DM** |  |  |  | x |  | | **> 1000 DM** |  |  |  |  | x | |

employment\_length binarize\_employment\_length\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | unemployed 62  0 - 1 yrs 172  1 - 4 yrs 339  4 - 7 yrs 174  > 7 yrs 253 | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **el\_un** | **el0-1** | **el1-4** | **el4-7** | **el>7** | | **unemployed** | x |  |  |  |  | | **0 - 1 yrs** |  | x |  |  |  | | **1 - 4 yrs** |  |  | x |  |  | | **4 - 7 yrs** |  |  |  | x |  | | **>7 yrs** |  |  |  |  | x | |

installment\_ rate binarize\_installment\_rate\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 136  2 231  3 157  4 476 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **ir1** | **ir2** | **ir3** | **ir4** | | **1** | x |  |  |  | | **2** |  | x |  |  | | **3** |  |  | x |  | | **4** |  |  |  | x | |

personal\_status binarize\_personal\_status\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | single male 548  female 310  married male 92  divorced male 50 | |  |  |  |  | | --- | --- | --- | --- | |  | **ps\_male** | **ps\_** **married** | **ps\_** **divorced** | | **single male** | x |  |  | | **female** |  |  |  | | **married male** | x | x |  | | **divorced male** | x |  | x | |

other\_debtors binarize\_other\_debtors\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | none 907  guarantor 52  co-applicant 41 | |  |  |  |  | | --- | --- | --- | --- | |  | **od\_none** | **od\_guar** | **od\_co** | | **none** | x |  |  | | **guarantor** |  | x |  | | **co-applicant** |  |  | x | |

residence\_history binarize\_residence\_history\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 130  2 308  3 149  4 413 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **rh1** | **rh2** | **rh3** | **rh4** | | **1** | x |  |  |  | | **2** |  | x |  |  | | **3** |  |  | x |  | | **4** |  |  |  | x | |

property binarize\_property\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | other 332  real estate 282  building society savings 232  unknown/none 154 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **pr\_o** | **pr\_re** | **pr\_bss** | **pr\_unk** | | **other** | x |  |  |  | | **real estate** |  | x |  |  | | **BSS** |  |  | x |  | | **unknown/none** |  |  |  | x | |

age binarize\_age\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean 35  std 11  min 19  25% 27  50% 33  75% 42  max 75 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **age<27** | **age<33** | **age<42** | **age>42** | | **<27** | x | x | x |  | | **28-33** |  | x | x |  | | **34-42** |  |  | x |  | | **>42** |  |  |  | x | |

installment\_plan binarize\_installment\_plan\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | none 814  bank 139  stores 47 | |  |  |  |  | | --- | --- | --- | --- | |  | **ps\_male** | **ps\_** **married** | **ps\_** **divorced** | | **none** | x |  |  | | **bank** |  | x |  | | **stores** |  |  | x | |

housing binarize\_housing\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | own 713  rent 179  for free 108 | |  |  |  |  | | --- | --- | --- | --- | |  | **h\_owne** | **h\_rent** | **h\_ff** | | **own** | x |  |  | | **rent** |  | x |  | | **for free** |  |  | x | |

existing\_credits binarize\_existing\_credits\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 633  2 333  3 28  4 6 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **ex1** | **ex2** | **ex3** | **ex4** | | **1** | x |  |  |  | | **2** |  | x |  |  | | **3** |  |  | x |  | | **4** |  |  |  | x | |

dependents binarize\_dependents\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 845  2 155 | |  |  | | --- | --- | |  | **dep** | | **1** |  | | **2** | x | |

telephone binarize\_telephone\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | yes 404  none 596 | |  |  | | --- | --- | |  | **tel** | | **yes** | x | | **no** |  | |

foreign\_worker binarize\_foreign\_worker\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | yes 963  no 37 | |  |  | | --- | --- | |  | **fw** | | **yes** | x | | **no** |  | |

job binarize\_job\_feature(data, verbose=False)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | skilled employee 630  unskilled resident 200  mangement self-employed 148  unemployed non-resident 22 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **j\_se** | **j\_ur** | **j\_ms** | **j\_un** | | **SE** | x |  |  |  | | **UR** |  | x |  |  | | **MS** |  |  | x |  | | **UN** |  |  |  | x | |

Алгоритмы классификации

classifiers.py

Для оценки точности алгоритмов классификации исходные данные были разделены на тестовые и обучающие в соотношении:

Алгоритм 1 algorithm1(plus, minus, x\_pred)

Алгоритм основан на нормированной сумме мощности пересечения описания классифицируемого объекта с описанием объекта из «-» или «+» контекста.

Псевдокод алгоритма:

*algorithm1(, , )*

*- объекты из плюс контекста*

*- объекты из минус контекста*

*– классифицируемый объект*

*for in*

*for in*

*if >*

*return +*

*else*

*return -*

Показатели алгоритма:

|  |  |
| --- | --- |
| Accuracy | 0.765 |
| True Positive | 45 |
| True Negative | 108 |
| False Positive | 21 |
| False Negative | 26 |
| True Positive Rate | 0.634 |
| True Negative Rate | 0.837 |
| Negative Predictive Value | 0.806 |
| False Positive Rate | 0.163 |
| False Discovery Rate | 0.318 |
| Precision | 0.682 |
| Recall | 0.634 |
| F-measure | 0.657 |

Алгоритм 2 algorithm2(plus, minus, x\_pred, threshold)

Алгоритм основан на нормированной поддержке описания классифицируемого объекта с объектами из «-» или «+» контекста. Для каждого пересечения описания классифицируемого объекта с описанием объекта из «-» или «+» контекста рассчитывается количество вложений данного пересечения в описания примеров из «-» (если классифицируемый объект был пересечен с объектом из «+» контекста) или «+» контекста (если классифицируемый объект был пересечен с объектом из «-» контекста). Рассчитанное количество вложений сравнивается с заданным пороговым значением threshold. Если количество вложений меньше threshold, то данному объекту отдается положительный или отрицательный «голос», в зависимости от контекста объекта с которым происходило пересечение.

Псевдокод алгоритма:

*algorithm1(, , , )*

*- объекты из плюс контекста*

*- объекты из минус контекста*

*– классифицируемый объект*

*for in*

*for in*

*if*

*if*

*for in*

*for in*

*if*

*if*

*if >*

*return +*

*else*

*return -*

Показатели алгоритма:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **threshold** | **0.25** | **0.4** | **0.5** | **0.55** | **0.6** | **0.75** | **1** |
| Accuracy | 0.76 | 0.76 | 0.775 | 0.77 | 0.77 | 0.76 | 0.76 |
| True Positive | 50 | 51 | 51 | 51 | 51 | 51 | 51 |
| True Negative | 102 | 101 | 104 | 103 | 103 | 101 | 101 |
| False Positive | 27 | 28 | 25 | 26 | 26 | 28 | 28 |
| False Negative | 21 | 20 | 20 | 20 | 20 | 20 | 20 |
| True Positive Rate | 0.704 | 0.718 | 0.718 | 0.718 | 0.718 | 0.718 | 0.718 |
| True Negative Rate | 0.791 | 0.783 | 0.806 | 0.798 | 0.798 | 0.783 | 0.783 |
| Negative Predictive Value | 0.829 | 0.835 | 0.839 | 0.837 | 0.837 | 0.835 | 0.835 |
| False Positive Rate | 0.209 | 0.217 | 0.194 | 0.201 | 0.201 | 0.217 | 0.217 |
| False Discovery Rate | 0.351 | 0.354 | 0.329 | 0.338 | 0.338 | 0.354 | 0.354 |
| Precision | 0.649 | 0.646 | 0.671 | 0.662 | 0.662 | 0.645 | 0.646 |
| Recall | 0.704 | 0.718 | 0.718 | 0.718 | 0.718 | 0.718 | 0.718 |
| F-measure | 0.675 | 0.680 | 0.694 | 0.689 | 0.689 | 0.680 | 0.680 |

**Оптимизация количества признаков для классификации**

features\_selection\_ga.py

Точность алгоритмов можно увеличить, если использовать для классификации подмножество признаков, а не весь их набор. Данное подмножество можно оптимизировать, используя какой либо алгоритм. Такой подход позволит выделить признаки которые влияют на значение целевого признака.

Для оптимизации использовался генетический алгоритм, в котором оптимизировалось значение точности. На каждом шаге алгоритма генерировался бинарный массив длины равной количеству исходных (не бинаризованных) признаков. Далее в соответствии с этим массивом делалась выборка признаков, которая бинаризовывалась, делилась на тестовую/обучающую части и рассчитывалась точность в соответствии с заданным алгоритмом классификации.

Расчет точности классификации с выбранными признаками на каждом шаге генетического алгоритма требует больших вычислительных затрат, в связи с этим накладывается ограничение на параметры генетического алгоритма, такие как размер популяции, количество эпох. Это ограничение не позволяет в полной мере оптимизировать множество признаков. Учитывая данные ограничения, мною был получен прирост точности в 2% на следующей выборке признаков:

months\_loan\_duration, checking\_balance, credit\_history,

purpose, employment\_length, installment\_rate,

personal\_status, other\_debtors, residence\_history,

property, existing\_credits, telephone, default,

Полученные показатели алгоритмов:

|  |  |  |
| --- | --- | --- |
| **Classiffier** | **Algorithm 1** | **Algorithm 2** |
| Accuracy | 0.785 | 0.795 |
| True Positive | 43 | 56 |
| True Negative | 114 | 103 |
| False Positive | 15 | 26 |
| False Negative | 28 | 15 |
| True Positive Rate | 0.606 | 0.789 |
| True Negative Rate | 0.884 | 0.798 |
| Negative Predictive Value | 0.803 | 0.873 |
| False Positive Rate | 0.116 | 0.201 |
| False Discovery Rate | 0.259 | 0.317 |
| Precision | 0.741 | 0.683 |
| Recall | 0.606 | 0.789 |
| F-measure | 0.667 | 0.732 |

**Итоговое сравнение точности**

Для сравнения приведены показатели наивного байесовского метода, метода k ближайших соседей и дерева решений. Реализация данных алгоритмов классификации была взята из библиотеки sklearn. Показатели данных алгоритмов представлены без оптимизации параметров и при сравнительной оценке алгоритмов необходимо учитывать данный факт.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classiffier** | **Algorithm 1** | **Algorithm 2** | **Naive Bayes** | **K Neighbors** | **Decision Tree** |
| Accuracy | 0.785 | 0.795 | 0.765 | 0.63 | 0.67 |
| True Positive | 43 | 56 | 50 | 31 | 24 |
| True Negative | 114 | 103 | 103 | 95 | 110 |
| False Positive | 15 | 26 | 26 | 34 | 19 |
| False Negative | 28 | 15 | 21 | 40 | 47 |
| True Positive Rate | 0.606 | 0.789 | 0.704 | 0.437 | 0.338 |
| True Negative Rate | 0.884 | 0.798 | 0.798 | 0.736 | 0.853 |
| Negative Predictive Value | 0.803 | 0.873 | 0.831 | 0.704 | 0.701 |
| False Positive Rate | 0.116 | 0.201 | 0.201 | 0.264 | 0.147 |
| False Discovery Rate | 0.259 | 0.317 | 0.342 | 0.523 | 0.442 |
| Precision | 0.741 | 0.683 | 0.658 | 0.477 | 0.558 |
| Recall | 0.606 | 0.789 | 0.704 | 0.437 | 0.338 |
| F-measure | 0.667 | 0.732 | 0.681 | 0.421 | 0.5 |

**Онлайн схема классификации**

online\_classification.py

Для реализации онлайн схемы классификации был построен класс. При инициализации он принимает объекты обучающего контекста и алгоритм классификации. Для классификации новых объектов вызывается метод predict. Данный метод находит класс каждого объекта используя алгоритм указанный при инициализации и добавляет в обучающую выборку новый объект в зависимости от определенного класса.

Листинг класса онлайн схемы классификации:

class OnlineClassifier(object):

def \_\_init\_\_(self, plus\_data, minus\_data, algorithm):

self.plus\_data = plus\_data

self.minus\_data = minus\_data

self.algorithm = algorithm

def predict(self, x\_pred):

y\_pred = []

for x in x\_pred:

y = self.algorithm(self.plus\_data, self.minus\_data, [x])

y\_pred.append(y[0])

if y[0] == 1:

self.plus\_data.append(x)

else:

self.minus\_data.append(x)

return y\_pred

Полученные показатели онлайн схемы:

|  |  |  |
| --- | --- | --- |
| **Classiffier** | **Algorithm 1** | **Algorithm 2** |
| Accuracy | 0.78 | 0.785 |
| True Positive | 0.738 | 55 |
| True Negative | 42 | 102 |
| False Positive | 114 | 27 |
| False Negative | 15 | 16 |
| True Positive Rate | 29 | 0.775 |
| True Negative Rate | 0.591 | 0.791 |
| Negative Predictive Value | 0.884 | 0.864 |
| False Positive Rate | 0.797 | 0.209 |
| False Discovery Rate | 0.116 | 0.329 |
| Precision | 0.263 | 0.671 |
| Recall | 0.737 | 0.775 |
| F-measure | 0.591 | 0.719 |