

Каргальцев Степан, M05-894a

Домашнее задание по теории решеток.

Алгоритм:

Давайте извлечем всевозможные фичи для всех положительных примеров (power, support, confidence), а так же для симметричной задачи.

После этого возьмем различных агрегации от них (минимум, максимум, медиана, среднее), а также разности агрегаций для прямой и симметричной задач. Получим 36 агрегатных функций на каждый пример.

После этого будем искать правило в виде $Arrg[i] >(<) threshold$, где i от 1 до 36, а $threshold$ перебирается по сетке от минимального до максимального значения агрегатной функции, максимизируя $f1_score$

Среди топ-10 решающих правил:

```
Rule: median(POS_SUPPORT-NEG_SUPPORT) > -0.005758
Rule: 1.0 - median(NEG_SUPPORT) > 0.9381
Rule: mean(POS_POWER-NEG_POWER) > -0.4242
```

Лучшие метрики:

```
accuracy: 0.9186
f1_score: 0.9474
precision: 0.9403
recall: 0.9545
```

Если отложить валидационную выборку и выбирать правило по кросс-валидации на оставшемся сете, а потом применить его к ней то получится:

```
val_accuracy: 0.8854
val_f1_score: 0.9091
```

```
val_precision: 0.8462  
val_recall: 0.9821
```

То есть выбор правила по кросс-валидации не совсем корректен, все-таки мы переобучаемся

Если делать кросс-валидацию не по 10 фолдам, а по 20, то качество растёт:

```
accuracy: 0.9767  
f1_score: 0.9841  
recall: 1.0000  
precision: 0.9688
```

Значит, увеличение количества данных в контекстах сильно влияет на итоговое качество.