Отчет по БДЗ. Курс: Упорядоченные множества в анализе данных.
Иванова Полина

Задача бинарной (ленивой) классификации решалась с помощью двух алгоритмов.

Для первого алгоритма, при попытке классифицировать поступивший объект g, считаем следующие характеристики:

$$\frac{|\left(g' \cap g_i^+\right)^+|}{|G^+|}, \frac{|\left(g' \cap g_j^-\right)^-|}{|G^-|}, \tag{1}$$

где G^+ — положительный контекст тренировочной выборки, G^- — отрицательный контекст тренировочной выборки, g_i^+ и g_j^- — элементы положительного и отрицательного контекстов соответственно. Если вычисленная поддержка в положительной контексте (в отрицательном контексте) больше некоторого минимального порога (threshold), то объекту добавляется положительная (соответственно, отрицательная) метка. Если объекту присвоились и положительная, и отрицательная метки, то классификация противоречива. В данном алгоритме возможны также случаи, когда классификация не проводится: если поддержки и для положительного, и для отрицательного контекстов меньше порогового значения (threshold).

Второй алгоритм учитывает число элементов g_i^+ из положительного контекста тренировочной выборки, таких что для поступившего элемента g признаки $g' \cap g_i^+$ содержатся в некотором описании элемента положительного контекста, при этом не содержатся в описаниях элементов отрицательного контекста. Аналогично, считается элементов из отрицательного контекста, для которых $g' \cap g_j^-$ вкладывается в описание некоторого элемента отрицательного контекста и не вкладывается в элементы положительных. Затем доля таких элементов в каждом контексте сравнивается с некоторым пороговым числом threshold: если доля больше, чем threshold, то элементу присваивается соответствующая метка.

Алгоритмы проверялись на части массива **Tic-Tac-Toe** (а именно **train2.csv**, **test2.csv**).

Для первого алгоритма после применения метода скользящего контроля (k-fold cross-validation с $k = 5$) был выбран threshold= 0.125. Однако метод не показал себя с хорошей стороны. В качестве метрики качества использовалась точность. Результаты в таблице 1.

Для второго алгоритма также использоввлся метод скользящего контроля с $k = 5$ для выбора порогового значения threshold. Результаты показаны в таблице 1.

Таблица 1: Результаты метода скользящего контроля с $k = 5$. 'Not classified' — доля элементов, для которых не был определен класс (не ставилась ни одна из меток 'positive' и 'negative')

Алгоритм 1	threshold	accuracy	not classified
	0.1	0.02	0.03
	0.125	0.19	0.22
	0.13	0.16	0.33
Алгоритм 2	threshold	accuracy	not classified
	0.5	nan	1
	0.2	0.33	0.6
	0.1	0.95	0.01

В итоге, для применения Алгоритма 1 на тестовой выборке был выбран параметр 0.125, а для Алгоритма 2 — параметр 0.1. Также можем видеть, что Алгоритм 1 значительно проигрывает Алгоритму 2.

Результаты на текстовой выборке **test2.csv** показаны в таблице 2 ниже.

На тестовой выборке Алгоритм 2 сработал лучше. Также, для него можно сделать модификацию: сравнивать описанные выше доли для положительного и отрицательного контекстов не с некоторым пороговым значением, а между собой, присваивая ту метку ('positive' или 'negative'), для которой соответствующая доля выше (и присваивая объекту обе метки, если разница между долями меньше, например, 0.01).

Таблица 2: Результаты метода скользящего контроля с $k = 5$.

Алгоритм 1	threshold	accuracy	not classified
	0.125	0.216	0.44
Алгоритм 2	threshold	accuracy	not classified
	0.1	0.93	0.05