Рубцов В.

Домашние задание

В качестве данных были выбраны данные по съедобным/несъедобном грибам.

https://archive.ics.uci.edu/ml/datasets/Mushroom

Всего в данных было 8124 объектов, 22 категориальных признаков. Для тестирования были созданы по 10 файлов для обучения и тестирования. Для этого для каждой пары выбирались случайные 1000 объектов, 900 — для обучения, 100 — для тестирования. Все данные переводились в бинарный формат.

Для сравнения были реализованы 4 алгоритма:

1) Тот что был реализован в lazyfca.ipynp.

2) $$f(g_\tau, +) = \frac{\sum_{g^+} a^{|g'_\tau \cap g^+|} * h(g_\tau, g^+)}{|G^+|}$$

   где $h(g_\tau, g^+) = 1$ если нет описания из $G^-$ в которое вкладывается $|g'_\tau \cap g^+|$. И $h(g_\tau, g^+) = 0$ в ином случае.

   Аналогично определяем $f(g_\tau, -)$. Если $f(g_\tau, +) > f(g_\tau, -)$ то классифицируем объект как $+$ иначе как -.

3) Наивный байесовский классификатор.

4) Тестируемый объект классифицируется таким классом, к которому принадлежит объект, пересечение описаний с которым максимально.

Так как результаты (первая таблица) для алгоритмом 1-3 получились очень хорошими, а хотелось бы их сравнить, то попробуем использовать выбоки для обцчения меньшего объема — 50 объектов. Для тестирования также выбираем 50 объектов. Для второго алгоритма наилучшего рузультата удается добится при $a = 1.5$.

| Algorithm | 1 | 2, $a = 1$ | 3 | 4 |
|---|---|---|---|---|
| True Positive | 512 | 543 | 519 | 524 |
| True Negative | 380 | 455 | 476 | 475 |
| False Positive | 96 | 1 | 0 | 1 |
| False Negative | 12 | 1 | 5 | 0 |
| True Positive Rate | 0.55 | 0.56 | 0.56 | 0.57 |
| True Negative Rate | 0.43 | 0.54 | 0.55 | 0.54 |
| Negative Predictive Value | 0.17 | 0.0 | 0.0 | 0.0 |
| False Positive Rate | 0.95 | 0.99 | 0.99 | 1.0 |
| False Discovery Rate | 0.15 | 0.0 | 0.0 | 0.0 |
| Accuracy | 0.892 | 0.998 | 0.995 | 0.999 |
| AccuracyPrecision | 0.84 | 0.99 | 1.0 | 0.99 |
| Recall | 0.98 | 0.99 | 0.99 | 1.0 |

| Algorithm | 1 | 2, $a = 1.5$ | 3 | 4 |
|---|---|---|---|---|
| True Positive | 252 | 253 | 181 | 241 |
| True Negative | 197 | 224 | 243 | 236 |
| False Positive | 46 | 15 | 0 | 7 |
| False Negative | 5 | 8 | 76 | 16 |
| True Positive Rate | 0.54 | 0.54 | 0.39 | 0.52 |
| True Negative Rate | 0.44 | 0.49 | 0.55 | 0.53 |
| Negative Predictive Value | 0.2 | 0.07 | 0.0 | 0.03 |
| False Positive Rate | 0.96 | 0.96 | 0.75 | 0.94 |
| False Discovery Rate | 0.17 | 0.06 | 0.0 | 0.03 |
| Accuracy | 0.898 | 0.954 | 0.848 | 0.954 |
| AccuracyPrecision | 0.84 | 0.94 | 1.0 | 0.97 |
| Recall | 0.98 | 0.97 | 0.7 | 0.93 |