Актуальность темы

Задача отбора признаков решается во всех областях, где используется машинное обучение. Выбор признаков на основе теории информации является популярным подходом из-за его вычислительной эффективности, масштабируемости с точки зрения размерности набора данных и независимости от модели машинного обучения.

Постановка задачи

Для множества признаков $F = \{f_1, f_2, ..., f_N\}$ набора данных D размерности N необходимо определить подмножество признаков S с размерностью K, где $K \leq N$ и $S \subseteq F$. Подмножество S должно обеспечивать равную или лучшую точность классификации по сравнению с набором признаков F.

Обзор литературы

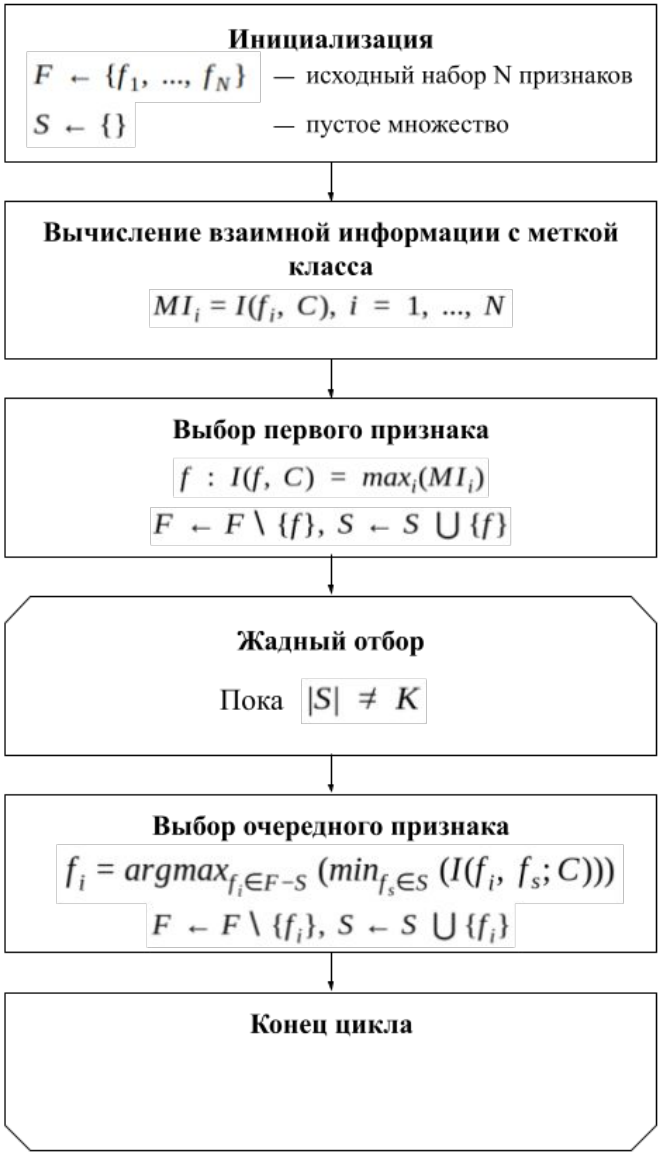
- Battiti, R. (1994). Использование взаимной информации для выбора признаков в контролируемом обучении нейронной сети
- Estévez, P. A., Tesmer, M., Perez, A., & Zurada, J. M. (2009). Нормализованный выбор признаков с использованием взаимной информации.
- Yang, H., & Moody, J. (1999). Выбор признаков на основе совместной взаимной информации.

Недостаток метода JMI

- Переоценка значимости некоторых признаков.

Метод решения

- Подход максимум минимума вместо кумулятивной суммы.



Постановка задачи практической части

- Реализовать библиотеку, позволяющую вычислять взаимную информацию и условную взаимную информацию для больших данных.
- Протестировать работу библиотеки на искусственном наборе данных с заранее известным значением тестируемых функций.
- Реализовать алгоритм, использующий максимизацию совместной взаимной информации.
- Сравнить работу предложенного алгоритма с JMI на наборе данных.

Тестирование библиотеки на искусственном наборе данных

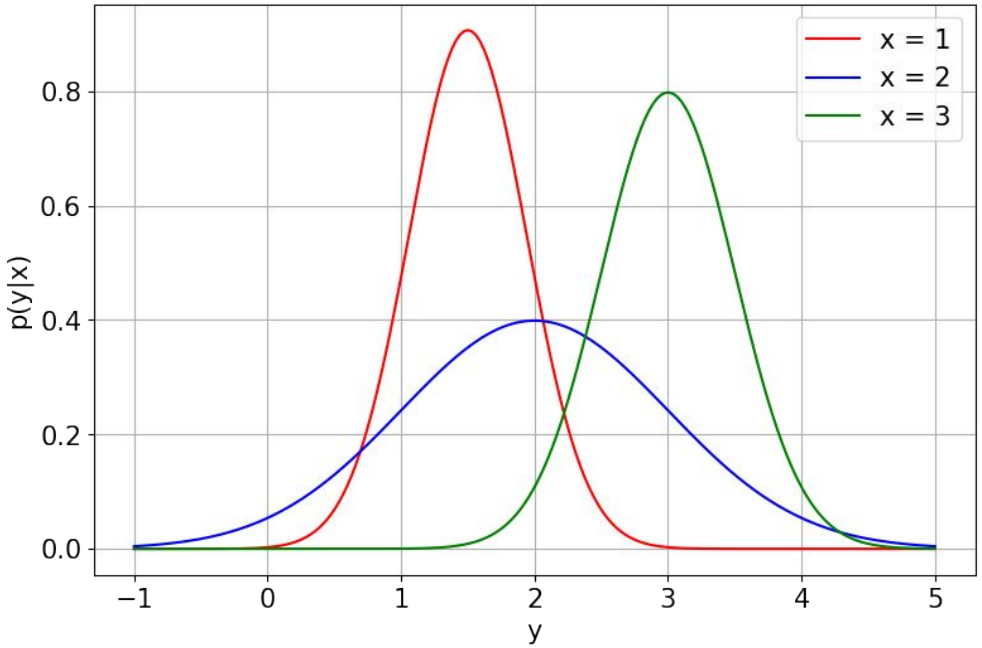
№	f1	f2	C	I(f1;C)	I(f2;C)	I(f1, f2; C)
1.	random(0, 100)	random(0, 100)	randint(0, 1)	0	0	0
2.	random(0, 1)	C - f1	randint(0, 1)	0	0	1.3876

1. Признаки и метка класса -- независимые случайные величины.
2. Метка класса является линейной комбинацией признаков, которые являются случайными величинами.

Тестирование библиотеки на искусственном наборе данных

x -- категориальный признак, принимающий значения 1, 2, 3.

y -- вещественный признак, распределение которого зависит от x.



x	μ	σ
1	1.5	0.44
2	2	1
3	3	0.5

Аналитическое решение:

I(x, y) = 0.585

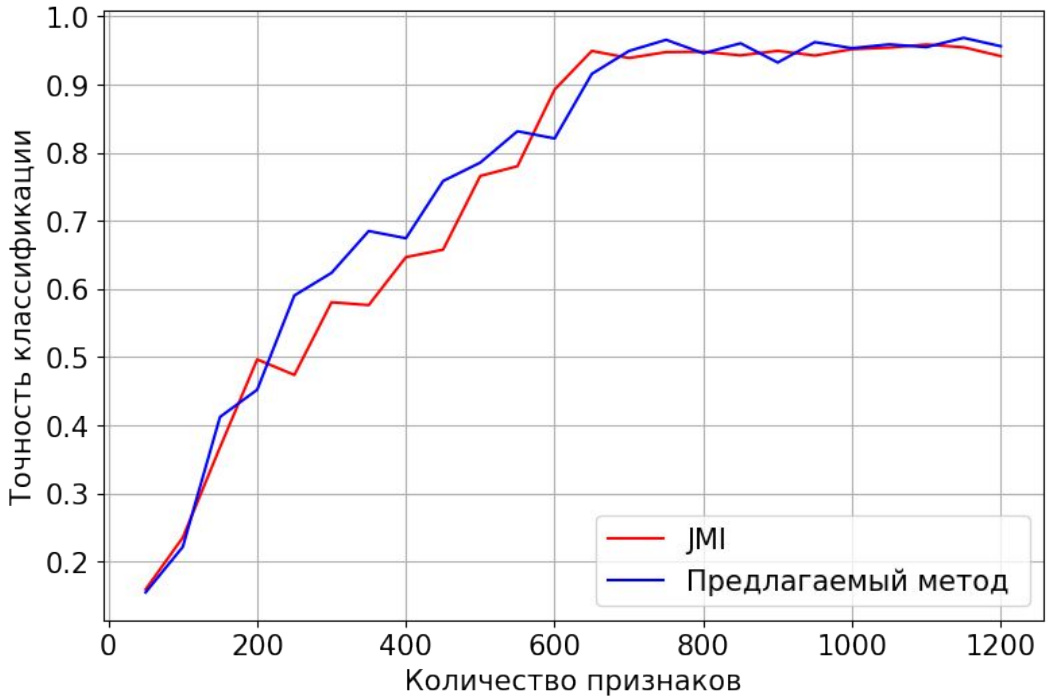
Численное решение:

I(x, y) = 0.52

Описание данных

- 10 000 000 строк данных, описывающих пользователя, которому показывается рекламный баннер.
- Описание содержит 1200 значений существующих признаков.
  - Признаки пользователя: регион, текущее время, последние N запросов в поисковую систему и другие.
  - Признаки рекламного баннера: новизна, является ли это рекламой товара/услуги, возрастная категория и другие.
- Задача бинарной классификации “Пользователь совершит клик на рекламный баннер”.

Сравнение метрик качества алгоритмов отбора признаков



Результаты

- Изучены методы отбора признаков, основанные на теории информации.
- Предложено решение для устранения недостатков метода JMI.
- Реализована и протестирована библиотека для вычисления взаимной информации.
- Реализован предложенный метод отбора признаков для больших данных.
- Проведено сравнение работы двух методов на наборе данных.

Спасибо за внимание!