Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Новосибирский государственный технический университет»

Кафедра теоретической и прикладной информатики

расчетно-графическое задание

по дисциплине «Статистический анализ нечисловых данных»

**Работа с текстовыми данными. Интеллектуальный анализ текстов**

Направление подготовки: 02.03.03 Математическое обеспечение и администрирование информационных систем.

|  |  |
| --- | --- |
| Выполнил:  Студент: Сидоров Д. И.  Группа: ПМИ-02  Факультет: ПМИ.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  подпись  «\_\_» ноября 2023 г. | Проверил:  Преподаватель: Тимофеева А.Ю.  Балл: \_\_\_\_\_\_\_\_\_\_  Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  подпись  «\_\_» ноября 2023 г. |
|  |  |

Новосибирск, 2023

Оглавление

[Постановка задачи 3](#_Toc150463963)

[Ход работы 3](#_Toc150463964)

## Постановка задачи

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Вариант | Файл | word1 | word2 | t-test\* | Chi2\* | LHR\* | MI\* |
| 26 | text\_sent\_3 | … | boosting | + | - | - | + |

Исходный набор данных содержит текст учебника Berk R. A. Statistical learning from a regression perspective. – Switzerland: Springer, 2020. – 433 c. Для исходного текста выполнена предварительная обработка. Результат представлен в файле text\_sent\_3. Каждая строка соответствует одному предложению. Слово с апострофом считается как одно слово (например, “what’s”).

Необходимо проанализировать биграммы смежных слов (стоящие рядом в одном предложении).

## Ход работы

**1. Из всего текста отоберем комбинации смежных слов, содержащие слово “boosting” на второй позиции. На первой позиции может быть любое слово. Сразу проранжируем биграммы по частоте встречаемости и сделаем предварительные выводы о возможных коллокациях.**

Всего таких комбинаций в корпусе – 57. Построим таблицу, в которой в первых двух столбцах будут соответствующие слова из биграмм, а в остальных столбцах – абсолютные частоты, которые можно получить из таблицы сопряженности, имеющей в общем случае следующий вид:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

– абсолютные частоты. Рассмотрим каждую частоту подробнее:

* – количество биграмм в корпусе;
* – количество биграмм в корпусе, начинающихся со слова ;
* – количество биграмм в корпусе, заканчивающиеся словом ;
* – количество биграмм в корпусе, которые не являются биграммой .

Таблица 1 – Ранжированные биграммы по частоте встречаемости (столбец по убыванию)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **word1** | **word2** | **O11** | **O12** | **O21** | **O22** |
| gradient | boosting | 58 | 15 | 110 | 155917 |
| of | boosting | 13 | 5388 | 155 | 150544 |
| the | boosting | 9 | 12389 | 159 | 143543 |
| regression | boosting | 8 | 867 | 160 | 155065 |
| to | boosting | 6 | 3533 | 162 | 152399 |
| in | boosting | 5 | 3385 | 163 | 152547 |
| that | boosting | 5 | 1873 | 163 | 154059 |
| why | boosting | 3 | 72 | 165 | 155860 |
| but | boosting | 3 | 699 | 165 | 155233 |
| and | boosting | 3 | 3255 | 165 | 152677 |
| tree | boosting | 2 | 235 | 166 | 155697 |
| forests | boosting | 2 | 202 | 166 | 155730 |
| fitted | boosting | 2 | 598 | 166 | 155334 |
| other | boosting | 2 | 259 | 166 | 155673 |
| does | boosting | 2 | 162 | 166 | 155770 |
| for | boosting | 2 | 2785 | 166 | 153147 |
| or | boosting | 2 | 624 | 166 | 155308 |
| on | boosting | 2 | 870 | 166 | 155062 |
| stochastic | boosting | 1 | 51 | 167 | 155881 |
| class | boosting | 1 | 264 | 167 | 155668 |
| help | boosting | 1 | 84 | 167 | 155848 |
| moreover | boosting | 1 | 38 | 167 | 155894 |
| first | boosting | 1 | 168 | 167 | 155764 |
| by | boosting | 1 | 777 | 167 | 155155 |
| known | boosting | 1 | 40 | 167 | 155892 |
| make | boosting | 1 | 151 | 167 | 155781 |
| traditional | boosting | 1 | 12 | 167 | 155920 |
| from | boosting | 1 | 902 | 167 | 155030 |
| size | boosting | 1 | 75 | 167 | 155857 |
| probabilities | boosting | 1 | 60 | 167 | 155872 |
| quantile | boosting | 1 | 81 | 167 | 155851 |
| example | boosting | 1 | 339 | 167 | 155593 |
| usual | boosting | 1 | 94 | 167 | 155838 |
| procedure | boosting | 1 | 147 | 167 | 155785 |
| conclusions | boosting | 1 | 24 | 167 | 155908 |
| perspective | boosting | 1 | 54 | 167 | 155878 |
| third | boosting | 1 | 49 | 167 | 155883 |
| than | boosting | 1 | 302 | 167 | 155630 |
| colinearity | boosting | 1 | 0 | 167 | 155932 |
| even | boosting | 1 | 166 | 167 | 155766 |
| chapter | boosting | 1 | 78 | 167 | 155854 |
| when | boosting | 1 | 476 | 167 | 155456 |
| recent | boosting | 1 | 32 | 167 | 155900 |
| new | boosting | 1 | 227 | 167 | 155705 |
| sparse | boosting | 1 | 8 | 167 | 155924 |
| learners | boosting | 1 | 5 | 167 | 155927 |
| links | boosting | 1 | 13 | 167 | 155919 |
| well | boosting | 1 | 183 | 167 | 155749 |
| important | boosting | 1 | 236 | 167 | 155696 |
| iteration | boosting | 1 | 27 | 167 | 155905 |
| index | boosting | 1 | 25 | 167 | 155907 |
| if | boosting | 1 | 568 | 167 | 155364 |
| places | boosting | 1 | 4 | 167 | 155928 |
| all | boosting | 1 | 420 | 167 | 155512 |
| alpha | boosting | 1 | 7 | 167 | 155925 |
| called | boosting | 1 | 112 | 167 | 155820 |
| bagging | boosting | 1 | 151 | 167 | 155781 |

По результатам ранжирования биграмм по частоте встречаемости можно сделать вывод, что самой часто встречающейся биграммой является биграмма “gradient, boosting”. Это указывает на то, что в анализируемом тексте часто обсуждаются темы, связанные с техникой машинного обучения для задач классификации и регрессии. Биграмму “regression, boosting” не берем в расчет, по крайней мере сейчас, потому что частота появления этой биграммы относительно частот появления каждого её слова в отдельности мизерно.

Устойчивость коллокации не всегда определяет высокая частота встречаемости биграммы. В таблице на лидирующих позициях мы можем заметить биграммы, содержащие в себе союзы и предлоги, которые не всегда можно назвать коллокациями, ведь такие комбинации в большинстве случаев не образуют устойчивого смыслового сочетания.

Следует отметить, что частотный анализ не сможет адекватно определить потенциальные коллокации, когда все биграммы имеют одинаковые частоты. Ранжирование в таких случаях не имеет смысла.

**2. Проранжируем биграммы, отобранные в п. 1, в соответствие с критериями.**

Используем **t-критерий** проверки равенства вероятностей для поиска потенциальных коллокаций. Для каждой отобранной биграммы вычислим статистику:

– вероятность появления биграммы;

– среднее при условии справедливости гипотезы, в которой вероятность совстречаемости двух слов равна произведению вероятности каждого слова в отдельности;

– дисперсия для биномиального распределения;

– общее число биграмм, которое можно определить как сумма всех абсолютных частот из таблицы сопряженности.

Данная статистика используется для определения, является ли разница между наблюдаемой частотой совместного вхождения слов w1 и w2 и ожидаемой частотой статистически значимой. Сразу сравним ранжированные статистики с критическим значением – квантилем распределения Стьюдента с степенью свободы уровня . Уровень значимости был взят равным 0.005.

Таблица 2 – Ранжированные биграммы по значению t-статистики

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **word1** | **word2** | **t\_stat** | **hypothesis** | **Квантиль** |
| gradient | boosting | 7,606870336 | Отвергается | 2,80707 |
| regression | boosting | 2,495548418 | Не отвергается |  |
| of | boosting | 1,993471545 | Не отвергается |  |
| why | boosting | 1,68546474 | Не отвергается |  |
| that | boosting | 1,332196435 | Не отвергается |  |
| but | boosting | 1,29586607 | Не отвергается |  |
| does | boosting | 1,28941589 | Не отвергается |  |
| forests | boosting | 1,258975224 | Не отвергается |  |
| tree | boosting | 1,233861674 | Не отвергается |  |
| other | boosting | 1,215597274 | Не отвергается |  |
| colinearity | boosting | 0,998926966 | Не отвергается |  |
| places | boosting | 0,99462202 | Не отвергается |  |
| learners | boosting | 0,993545783 | Не отвергается |  |
| alpha | boosting | 0,99139331 | Не отвергается |  |
| sparse | boosting | 0,990317073 | Не отвергается |  |
| traditional | boosting | 0,986012127 | Не отвергается |  |
| links | boosting | 0,98493589 | Не отвергается |  |
| conclusions | boosting | 0,973097287 | Не отвергается |  |
| index | boosting | 0,972021051 | Не отвергается |  |
| iteration | boosting | 0,969868577 | Не отвергается |  |
| recent | boosting | 0,964487394 | Не отвергается |  |
| moreover | boosting | 0,958029974 | Не отвергается |  |
| fitted | boosting | 0,957612628 | Не отвергается |  |
| known | boosting | 0,955877501 | Не отвергается |  |
| third | boosting | 0,946191372 | Не отвергается |  |
| stochastic | boosting | 0,944038898 | Не отвергается |  |
| perspective | boosting | 0,940810188 | Не отвергается |  |
| or | boosting | 0,937826194 | Не отвергается |  |
| probabilities | boosting | 0,934352769 | Не отвергается |  |
| size | boosting | 0,918209219 | Не отвергается |  |
| chapter | boosting | 0,914980509 | Не отвергается |  |
| quantile | boosting | 0,911751799 | Не отвергается |  |
| help | boosting | 0,908523089 | Не отвергается |  |
| usual | boosting | 0,897760723 | Не отвергается |  |
| to | boosting | 0,89457524 | Не отвергается |  |
| called | boosting | 0,878388464 | Не отвергается |  |
| procedure | boosting | 0,840720182 | Не отвергается |  |
| make | boosting | 0,836415235 | Не отвергается |  |
| bagging | boosting | 0,836415235 | Не отвергается |  |
| even | boosting | 0,820271686 | Не отвергается |  |
| first | boosting | 0,818119212 | Не отвергается |  |
| well | boosting | 0,801975663 | Не отвергается |  |
| new | boosting | 0,754621251 | Не отвергается |  |
| on | boosting | 0,750616097 | Не отвергается |  |
| important | boosting | 0,744935122 | Не отвергается |  |
| class | boosting | 0,714800496 | Не отвергается |  |
| than | boosting | 0,673903504 | Не отвергается |  |
| example | boosting | 0,634082749 | Не отвергается |  |
| in | boosting | 0,604449939 | Не отвергается |  |
| all | boosting | 0,546907581 | Не отвергается |  |
| when | boosting | 0,48663833 | Не отвергается |  |
| if | boosting | 0,38762456 | Не отвергается |  |
| by | boosting | 0,162691104 | Не отвергается |  |
| from | boosting | 0,028161525 | Не отвергается |  |
| and | boosting | -0,292354345 | Не отвергается |  |
| for | boosting | -0,706730803 | Не отвергается |  |
| the | boosting | -1,44775474 | Не отвергается |  |

Данный критерий интуитивно понятен. Однако проблем может доставить частое отклонение нулевой гипотезы, чего можно миновать, варьируя уровень значимости. t-критерий наиболее полезен в качестве ранжировщика коллокаций.

Один из самых важнейших минусов t-критерия в том, что он основан на предположении о нормальности распределения данных. Однако в реальности многие наборы данных не следуют нормальному распределению. Например, в случае анализа текста, некоторые слова могут встречаться очень часто, в то время как большинство слов встречаются редко, что приводит к скошенному распределению частот.

Найдем **оценки точечной взаимной информации**. По определению, данную оценку можно интерпретировать как объем информации о встречаемости слова , получаемой от появления слова . Для каждой отобранной биграммы она вычисляется по следующей формуле:

Здесь мы сравниваем вероятности совместного появления от произведения вероятностей появления каждого слова в отдельности. В случае независимости слов, вероятности будут равными и логарифм будет равен 0.

Таблица 3 – Ранжированные биграммы по значению оценки точечной взаимной информации

|  |  |  |
| --- | --- | --- |
| **word1** | **word2** | **PMI** |
| colinearity | boosting | 9,859793589 |
| gradient | boosting | 9,527950025 |
| places | boosting | 7,537865494 |
| learners | boosting | 7,274831088 |
| alpha | boosting | 6,859793589 |
| sparse | boosting | 6,689868588 |
| traditional | boosting | 6,159353871 |
| links | boosting | 6,052438667 |
| why | boosting | 5,215937399 |
| conclusions | boosting | 5,215937399 |
| index | boosting | 5,159353871 |
| iteration | boosting | 5,052438667 |
| recent | boosting | 4,81539947 |
| moreover | boosting | 4,57439137 |
| known | boosting | 4,502241584 |
| third | boosting | 4,215937399 |
| stochastic | boosting | 4,159353871 |
| perspective | boosting | 4,078433875 |
| probabilities | boosting | 3,929056251 |
| size | boosting | 3,611866076 |
| chapter | boosting | 3,556012841 |
| does | boosting | 3,502241584 |
| quantile | boosting | 3,502241584 |
| help | boosting | 3,450402653 |
| usual | boosting | 3,289937981 |
| forests | boosting | 3,187368247 |
| regression | boosting | 3,086654382 |
| called | boosting | 3,039614627 |
| tree | boosting | 2,97105034 |
| other | boosting | 2,831887592 |
| procedure | boosting | 2,650340223 |
| make | boosting | 2,611866076 |
| bagging | boosting | 2,611866076 |
| even | boosting | 2,476089296 |
| first | boosting | 2,458914153 |
| well | boosting | 2,336231633 |
| new | boosting | 2,026903575 |
| but | boosting | 1,989428869 |
| important | boosting | 1,97105034 |
| class | boosting | 1,80994504 |
| fitted | boosting | 1,630974898 |
| than | boosting | 1,616619606 |
| or | boosting | 1,569774742 |
| example | boosting | 1,450402653 |
| that | boosting | 1,306740336 |
| of | boosting | 1,161222474 |
| all | boosting | 1,142117166 |
| on | boosting | 1,091609264 |
| when | boosting | 0,961948133 |
| if | boosting | 0,707508747 |
| to | boosting | 0,655630043 |
| in | boosting | 0,454652126 |
| by | boosting | 0,256167244 |
| from | boosting | 0,041211411 |
| and | boosting | -0,225014799 |
| the | boosting | -0,568101198 |
| for | boosting | -0,584703698 |

Данный способ оценивания биграмм особенно чувствителен к неточным оценкам вероятностей, возникающим из-за разреженности данных. Часто критерий хи квадрат позволяет найти более подходящие пары слов.

Найдем **оценки средней взаимной информации**.

Оценка средней взаимной информации использует все абсолютные частоты из таблицы сопряженности, в отличие от оценки , что делает её мерой взаимосвязи, аналогичной статистике хи квадрат.

Таблица 4 – Ранжированные биграммы по значению оценки средней взаимной информации

|  |  |  |
| --- | --- | --- |
| **word1** | **word2** | **MI** |
| gradient | boosting | 0,003426315 |
| regression | boosting | 9,46184E-05 |
| why | boosting | 7,40306E-05 |
| colinearity | boosting | 6,31908E-05 |
| places | boosting | 4,01065E-05 |
| learners | boosting | 3,82555E-05 |
| alpha | boosting | 3,54029E-05 |
| sparse | boosting | 3,42545E-05 |
| of | boosting | 3,18209E-05 |
| traditional | boosting | 3,07268E-05 |
| links | boosting | 3,00251E-05 |
| does | boosting | 2,82049E-05 |
| conclusions | boosting | 2,46243E-05 |
| forests | boosting | 2,45422E-05 |
| index | boosting | 2,42644E-05 |
| iteration | boosting | 2,35863E-05 |
| recent | boosting | 2,20917E-05 |
| tree | boosting | 2,20829E-05 |
| moreover | boosting | 2,05851E-05 |
| other | boosting | 2,05291E-05 |
| known | boosting | 2,01368E-05 |
| third | boosting | 1,83712E-05 |
| stochastic | boosting | 1,80249E-05 |
| but | boosting | 1,76631E-05 |
| perspective | boosting | 1,75313E-05 |
| probabilities | boosting | 1,66254E-05 |
| size | boosting | 1,47268E-05 |
| that | boosting | 1,45958E-05 |
| chapter | boosting | 1,43963E-05 |
| quantile | boosting | 1,40794E-05 |
| help | boosting | 1,3775E-05 |
| usual | boosting | 1,28403E-05 |
| called | boosting | 1,1407E-05 |
| procedure | boosting | 9,25011E-06 |
| make | boosting | 9,04243E-06 |
| bagging | boosting | 9,04243E-06 |
| fitted | boosting | 8,44529E-06 |
| even | boosting | 8,31842E-06 |
| first | boosting | 8,22787E-06 |
| the | boosting | 7,9518E-06 |
| or | boosting | 7,91657E-06 |
| well | boosting | 7,58827E-06 |
| new | boosting | 6,03766E-06 |
| important | boosting | 5,76826E-06 |
| to | boosting | 5,09111E-06 |
| class | boosting | 5,01157E-06 |
| on | boosting | 4,21254E-06 |
| than | boosting | 4,14752E-06 |
| example | boosting | 3,44783E-06 |
| all | boosting | 2,27356E-06 |
| in | boosting | 2,12553E-06 |
| for | boosting | 1,77536E-06 |
| when | boosting | 1,67367E-06 |
| if | boosting | 9,55323E-07 |
| and | boosting | 3,63044E-07 |
| by | boosting | 1,3833E-07 |
| from | boosting | 3,76113E-09 |

**3. Для анализируемых биграмм зададим истинные значения переменной класса: 1 – биграмма является коллокацией (устойчивым словосочетанием), 0 – иначе.**

**Сразу построим ROC-кривые для оценки качества классификации биграмм как коллокаций/не коллокаций с помощью критериев.**

На основе частотного анализа, с помощью t-критерия и средней взаимной информации выделена одна биграмма, которую можно считать коллокацией: **gradient, boosting.**

Эта биграмма имеет высокую частоту встречаемости, значимую t-статистику и оценку взаимной информации. Она значительно выделяется среди остальных биграмм своей высокой долей вхождения относительно вхождения каждого слова в отдельности.

Если проанализировать ранжирование биграмм на основе t- критерия и средней взаимной информации, то дополнительно к найденной биграмме можно выделить биграмму **regression, boosting**.

Используя знания предметной области и глоссарий (в конце учебника), можно также выделить биграммы **index, boosting** и **colinearity, boosting.** Ещёвторую биграмму выделяет показатель точечной взаимной информации (однако к данному показателю надо относиться с осторожностью, ведь он чувствителен к разреженным данным).

Зададим для отобранных коллокаций значение переменной класса, равное 1. Для остальных анализируемых биграмм – 0.

Перейдем к построению ROC-кривой. Это график, позволяющий оценить качество бинарной классификации. По вертикальной оси данного графика располагается чувствительность , а по горизонтальной - частота ложноположительных результатов .

Рассмотрим алгоритм нахождения и , необходимых для построения графика, при использовании - статистики:

1. Формирование исходных данных: массива полученных - статистик и массива истинных значений переменной класса.
2. Формирование вектора значений - уникальных упорядоченных по возрастанию значений – статистик.
3. Цикл по всем значениям вектора . Для каждого :

* классифицировать все объекты обучающей выборки по правилу:

eсли -статистика > , то положительный класс, иначе – отрицательный;

* на основе предсказанных классов построить таблицу неточностей;
* используя таблицу неточностей, рассчитать и сохранить и .

1. По значениям и построить ROC-кривую.

Матрица ошибок (матрица неточностей) — это таблица, которая позволяет визуализировать эффективность алгоритма классификации путем сравнения прогнозируемого значения переменной класса с ее фактическим значением. Столбцы матрицы представляют наблюдения в прогнозируемом классе, а строки — наблюдения в фактическом классе.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* – истинно положительные результаты;
* – истинно отрицательные результаты;
* - ложноотрицательные результаты;
* - ложноположительные результаты.

Благодаря матрице неточностей можно вычислить ряд показателей, однако рассмотрим только нужные для построения ROC-кривой:

* Чувствительность. Это доля истинно положительных результатов среди всех действительно положительных случаев. Она вычисляется как ;
* Специфичность. Это доля истинно отрицательных результатов среди всех действительно отрицательных случаев. Она вычисляется как ;
* Частота ложноположительных результатов: .

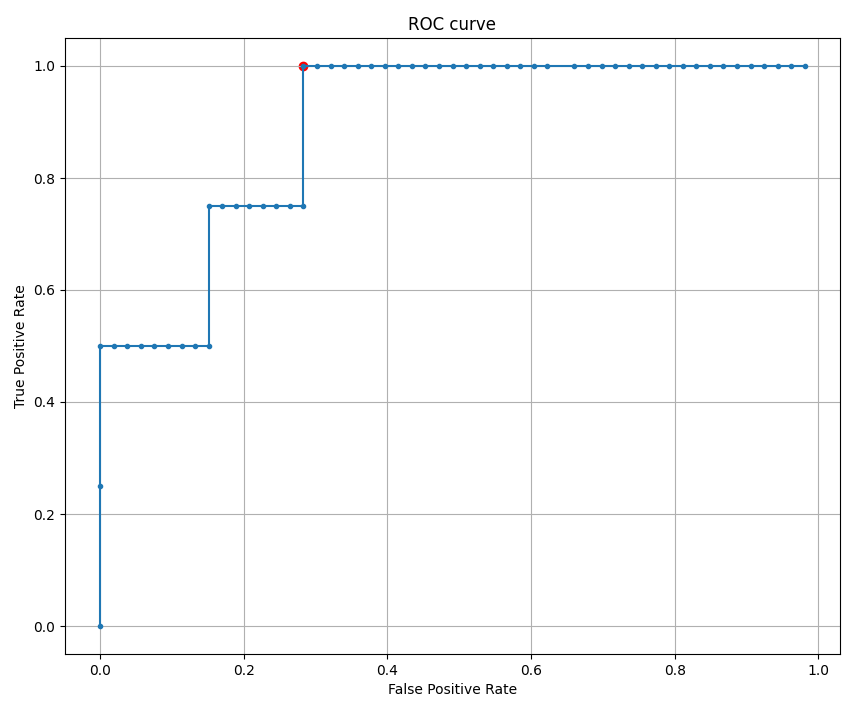


Рисунок 1 – ROC-кривая для бинарного классификатора, использующего t-критерий

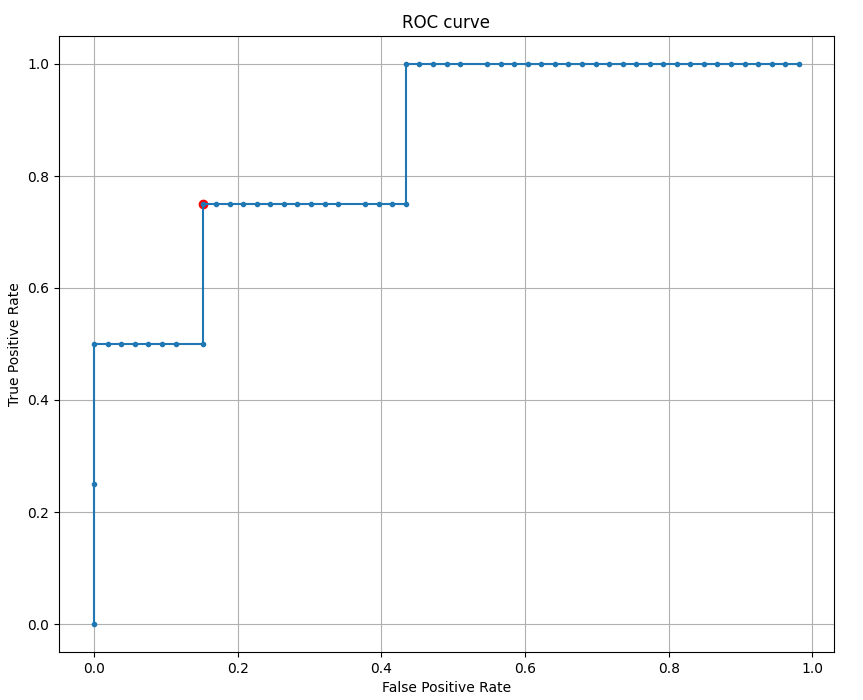


Рисунок 2 – ROC-кривая для бинарного классификатора, использующего показатель точечной взаимной информации

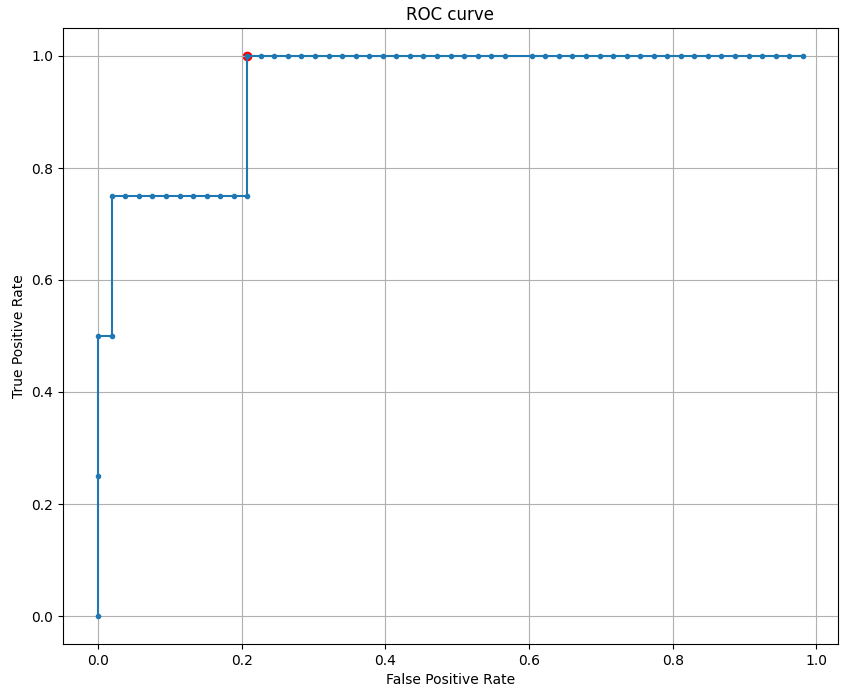


Рисунок 3 – ROC-кривая для бинарного классификатора, использующего показатель средней взаимной информации

Красная точка на графиках - оптимальное пороговое значение . Она определялась по правилу максимизации специфичности и чувствительности классификатора. Для графиков оптимальная точка соответственно: [0.28302, 1.0], [0.15094, 0.75], [0.20755, 1.0].

Оптимальный порог для t-критерия и критерия максимального правдоподобия определялся как значение , соответствующее оптимальной точке. Для t-критерия оптимальный порог - 0.84072, для точечной взаимной информации - 1.14212, для средней взаимной информации - 7.58827.

**4. Сравним критерии по качеству классификации на основе показателя AUC (площади под ROC-кривой). Сделаем выводы.**

Показатель AUC – это площадь под ROC-кривой. Чем выше показатель AUC, тем лучше качество классификатора. Если значение AUC=0.5 - непригодность выбранного метода классификации (соответствует случайному гаданию).

AUC для первой ROC-кривой - 0.87264.

AUC для второй ROC-кривой - 0.83491.

AUC для третьей ROC-кривой - 0.92453.

Значения AUC для критериев довольно высоки, что указывает на хорошее качество всех классификаторов. Однако классификатор, основанный на средней взаимной информации, имеет наибольшее значение AUC. Это указывает на то, что он эффективнее всех предсказывает то, является ли биграмма коллокацией или нет.