

МЕТРИКИ КАЧЕСТВА КЛАССИФИКАЦИИ

Задание основано на применении метрик качества в составе пакета Python Scikit-learn.

Цели практики:

- освоить основы вычисления различных мер качества классификации: долю правильных ответов, точность, полноту, AUC-ROC и пр.
- научиться сравнивать алгоритмы классификации при наличии ограничений на точность или полноту

Начало работы

В этом задании мы предлагаем вам самостоятельно реализовать градиентный спуск.

В качестве метрики качества будем использовать **AUC-ROC** (Area Under ROC-Curve). Она предназначена для алгоритмов бинарной классификации, выдающих оценку принадлежности объекта к одному из классов. По сути, значение этой метрики является агрегацией показателей качества всех алгоритмов, которые можно получить, выбирая какой-либо порог для оценки принадлежности.

В Scikit-Learn метрика **AUC** реализована функцией **sklearn.metrics.roc_auc_score**. В качестве первого аргумента ей передается вектор истинных ответов, в качестве второго — вектор с оценками принадлежности объектов к первому классу.

Задание

Различные метрики качества реализованы в пакете **sklearn.metrics**.

1. Загрузите файл **classification.csv**. В нем записаны истинные классы объектов выборки (колонок **true**) и ответы некоторого классификатора (колонок **predicted**).

2. Заполните таблицу ошибок классификации:

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Необходимо подсчитать **TP, FP, FN и TN** согласно их определениям. Например, FP – это количество объектов, имеющих класс 0, но отнесенных алгоритмом к классу 1.

Представьте ответ в виде четыре чисел, через пробел.

3. Посчитайте основные метрики качества классификатора:

- Аккуратность (доля верно угаданных) – **sklearn.metrics.accuracy**
- Precision (точность) – **sklearn.metrics.accuracy.precision_score**
- Recall (полнота) – **sklearn.metrics.recall_score**
- F-мера – **sklearn.metrics.f1_score**

4. Имеется четыре обученных классификатора.

Загрузите файл **scores.csv**, в котором записаны истинные классы и значения степени принадлежности положительному классу для каждого классификатора на некоторой выборке:

- для логистической регрессии – вероятность положительного класса (колонок **score_logreg**),
- для SVM – отступ от разделяющей поверхности (колонок **score_svm**),
- для метрического алгоритма — взвешенная сумма классов соседей (колонок **score_knn**),
- для решающего дерева — доля положительных объектов в листе (колонок **score_tree**).

5. Посчитайте площадь под **ROC-кривой** для каждого классификатора. Какой классификатор имеет наибольшее значение метрики **AUC-ROC** (укажите название столбца с ответами этого классификатора)?

Воспользуйтесь функцией `sklearn.metrics.roc_auc_score`.

6. Какой классификатор достигает наибольшей точности (***Precision***) при полноте (***Recall***) не менее **70%** (укажите название столбца с ответами этого классификатора)?

Какое значение точности при этом получается?

Ответ на каждое задание — текстовый файл, содержащий ответ в первой строчке (#.txt). Обратите внимание, что отправляемые файлы не должны содержать перевод строки в конце.

Уточнения по выполнению задания:

Чтобы получить ответ на этот вопрос, найдите все точки **precision-recall** кривой с помощью функции `sklearn.metrics.precision_recall_curve`.

Она возвращает три массива: **precision, recall, thresholds**. В них записаны точность и полнота при определенных порогах, указанных в массиве **thresholds**.

Найдите максимальное значение точности среди тех записей, для которых полнота не меньше, чем **0.7**.

При необходимости округляйте ответ до двух знаков после запятой.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

В задачах классификации может быть много особенностей, влияющих на подсчет качества: различные цены ошибок, несбалансированность классов и т.д. Из-за этого существует большое количество метрик качества — каждая из них рассчитана на определенное сочетание свойств задачи и требований к ее решению.

Меры качества классификации можно разбить на две большие группы: предназначенные для алгоритмов, выдающих номера классов, и для алгоритмов, выдающих оценки принадлежности к классам.

К первой группе относятся доля правильных ответов, точность, полнота, F-мера.

Ко второй — площади под ROC- или PR-кривой.