ВМК

Задание 1. Метрические алгоритмы классификации

Практикум 317 группы, 2017

Начало выполнения задания: 7 октября 2017 года.
Мягкий дедлайн: **22 октября 2017 года, 23:59.**
Жёсткий дедлайн: **5 ноября 2017 года, 23:59.**

Формулировка задания

Данное задание направлено на ознакомление с метрическими алгоритмами классификации, а также методами работы с изображениями. В задании необходимо:

1. Написать на языке Python собственные реализации метода ближайших соседей и кросс-валидации. Реализации должны соответствовать требованиям, описанным ниже.
2. Провести описанные ниже эксперименты с датасетом изображений цифр MNIST.
3. Написать отчёт о проделанной работе (формат PDF). Отчёт должен быть подготовлен в системе LaTeX.
4. В систему проверки anytask сдаётся .zip архив с модулями с написанным кодом, jupyter-notebook с кодом экспериментов (может быть не структурирован, проверяется при наличии вопросов к результатам экспериментов) и отчёт. Архив необходимо назвать `task1_фамилия_имя.zip`.

Метод k ближайших соседей

Рассмотрим задачу многоклассовой классификации. Пусть дана обучающая выборка $X = (x_i, y_i)_{i=1}^{l}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{Y} = \{1, \ldots, k\}$. Пусть также введена *функция расстояния* $\rho : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$. От функции не требуется соответствие всем аксиомам метрики, достаточно лишь неотрицательности. Для каждого объекта x' можно расположить объекты обучающей выборки в порядке неубывания расстояний:

$$\rho(x', x_{(1)}) \leqslant \rho(x', x_{(2)}) \leqslant \ldots \leqslant \rho(x', x_{(l)})$$

Метрический алгоритм k ближайших соседей (k nearest neighbours, kNN) относит объект x' к тому классу, представителей которого окажется больше всего среди его k ближайших соседей:

$$a(x') = \arg\max_{y \in Y} \sum_{i=1}^{k} \mathbb{I}[y_{(i)} = y]$$

Рассмотренный алгоритм никак не учитывает степень близости объекта x' к его ближайшим соседям. Алгоритм k взвешенных ближайших соседей позволяет учесть расстояния до соседей с помощью введения весов объектов, например так:

$$a(x') = \arg\max_{y \in Y} \sum_{i=1}^{k} \frac{1}{\varepsilon + \rho(x', x_{(i)})} \mathbb{I}[y_{(i)} = y]$$

Эффективная реализация подсчёта расстояний между объектами

Существуют различные алгоритмы поиска ближайших соседей. Для евклидовой метрики можно на этапе обучения строить различные структуры данных, позволяющие упростить сложность поиска ближайших соседей, например kd-деревья. Самый простой и универсальный способ — полный перебор всех объектов обучающей выборки.

Пусть l_1 — размер обучающей выборки, l_2 — размер тестовой выборки. Подсчитаем количество операций, требующихся для подсчёта квадратов всех попарных евклидовых расстояний между объектами обучающей и тестовой выборки. Квадрат евклидова расстояния между объектами x и z записывается так:

$$\rho(x, z) = \sum_{s=1}^{d} (x^s - z^s)^2$$

Можно вычислить все попарные расстояния за $3dl_1l_2$ операций: dl_1l_2 вычитаний, dl_1l_2 умножений (возведений в квадрат) и dl_1l_2 сложений. Число операций можно уменьшить с помощью простого трюка. Раскроем скобки в выражении выше:

$$\rho(x,z) = \sum_{s=1}^{d}(x^s)^2 + \sum_{s=1}^{d}(z^s)^2 - 2\sum_{s=1}^{d}x^s z^s$$

Первые две суммы достаточно вычислить для каждого объекта, а не пересчитывать отдельно по каждой паре объектов, это можно сделать за $2d(l_1 + l_2)$ операций, для вычисления третьей суммы потребуется $2dl_1l_2$ операций. Заметим, что вычисление третьей суммы по всем парам объектов можно записать как матричное произведение, которое за счёт эффективной реализации требует ещё меньше операций. Таким образом, суммарное число операций меньше, чем для первого способа.

Эффективный подбор параметров в методе kNN

Одним из основных параметров метода kNN является число ближайших соседей, которое в процессе обучения приходится подбирать по отложенной выборке или кросс-валидации. В общем случае, чтобы замерить качество алгоритма при m значениях параметра на отложенной выборке, необходимо m раз обучить и применить алгоритм. В случае k ближайших соседей это будет стоить $O(mdl^2)$ операций.

Отсортируем значения параметра, которые мы хотим проверить, по неубыванию:

$$k_1 \leqslant k_2 \leqslant \ldots \leqslant k_m$$

Найдём k_m ближайших соседей в обучающей выборке для объектов отложенной выборки. Заметим, что подсчёт ближайших соседей — самая вычислительно затратная часть алгоритма kNN. Очевидно, что для всех остальных значений параметра не нужно проводить поиск ближайших соседей заново, достаточно выбрать из уже найденных соседей нужное количество. Таким образом, стоимость алгоритма подбора параметров была сокращена до $O(dl^2)$ операций. Предложенная схема легко обобщается на случай оценивания алгоритма по кросс-валидации.

Преобразования объектов для улучшения качества алгоритма

Рассмотрим два простых способа улучшить качество алгоритма kNN с помощью некоторых элементарных преобразований изображений. Пусть у нас есть m элементарных преобразований $\phi_j(x) : \mathbb{R}^d \to \mathbb{R}^d$, $j \in \{1, \ldots, m\}$, введём $\phi_0(x) = x$.

Создадим новую обучающую выборку:

$$X_{new} = \left(\bigcup_{j=0}^{m} (\phi_j(x_i), y_i)_{i=1}^{l} \right)$$

Качество алгоритма, обученного выборке X_{new}, при удачно выбранных преобразованиях будет превышать исходное.

Другой способ основан на преобразовании объектов тестовой выборки, а не обучающей. Для каждого объекта тестовой выборки x' получим множество объектов $\Phi(x') = \{\phi_j(x') | j \in \{1, \ldots, m\}\}$. Введём «расстояние» от множества $\Phi(x')$ до объекта обучающей выборки x:

$$\rho(x, \Phi(x')) = \min_j \{\rho(x, \phi_j(x') | j \in \{1, \ldots, m\}\}$$

Вычисление k ближайших соседей для $\Phi(x')$ можно организовать следующим образом:

1. Для каждого объекта из $\Phi(x')$ найдём его ближайших соседей из обучающей выборки X
2. Среди всех найденных соседей выберем k объектов с наименьшим расстоянием

По полученному множеству ближайших соседей вычисляется ответ алгоритма. Так как размер тестовой выборки обычно меньше размера обучающей, второй способ более эффективен по памяти.

Список экспериментов

Эксперименты этого задания необходимо проводить на датасете MNIST. Загрузить датасет можно при помощи функции `sklearn.datasets.fetch_mldata("MNIST original")`. Датасет необходимо разбить на обучающую выборку (первые 60 тыс. объектов) и тестовую выборку (10 тыс. последних объектов).

1. Исследовать, какой алгоритм поиска ближайших соседей будет быстрее работать в различных ситуациях. Для каждого объекта тестовой выборки найти 5 его ближайших соседей в обучающей для евклидовой метрики. Для выборки нужно выбрать подмножество признаков, по которому будет считаться расстояние, размера 10, 20, 100 (подмножество признаков выбирается один раз для всех объектов, случайно). Необходимо проверить все алгоритмы поиска ближайших соседей, указанные в спецификации к заданию.

   Замечание. Для оценки времени долго работающих функций можно пользоваться либо функциями из модуля `time`, либо magic-командой `%time`, которая запускает код лишь один раз.

2. Оценить по кросс-валидации с 3 фолдами точность (долю правильно предсказанных ответов) и время работы k ближайших соседей в зависимости от следующих факторов:

   (a) k от 1 до 10 (только влияние на точность).

   (b) Используется евклидова или косинусная метрика.

3. Сравнить взвешенный метод k ближайших соседей, где голос объекта равен $1/(distance + \varepsilon)$, где $\varepsilon - 10^{-5}$, с методом без весов при тех же фолдах и параметрах.

4. Применить лучший алгоритм к исходной обучающей и тестовой выборке. Подсчитать точность. Сравнить с точностью по кросс-валидации. Сравнить с указанной в интернете точностью лучших алгоритмов на данной выборке. Построить и проанализировать матрицу ошибок (confusion matrix). Визуализировать несколько объектов из тестовой выборки, на которых были допущены ошибки. Проанализировать и указать их общие черты.

   Замечание. Можно воспользоваться функцией `sklearn.metrics.confusion_matrix`. Для визуализации можно воспользоваться `pyplot.subplot`, и `pyplot.imshow` с параметром `cmap="Greys"`. Также можно убрать оси координат при помощи команды `pyplot.axis("off")`.

5. Размножить обучающую выборку с помощью поворотов, смещений и применений гауссовского фильтра. Разрешается использовать библиотеки для работы с изображениями. Подобрать по кросс-валидации с 3 фолдами параметры преобразований. Рассмотреть следующие параметры для преобразований и их комбинации:

   (a) Величина поворота: 5, 10, 15 (в каждую из двух сторон)

   (b) Величина смещения: 1, 2, 3 пикселя (по каждой из четырёх размерностей)

   (c) Дисперсия фильтра Гаусса: 0.5, 1, 1.5

   Проанализировать, как изменилась матрица ошибок, какие ошибки алгоритма помогает исправить каждое преобразование.

   Замечание. Не обязательно хранить все обучающие выборки в процессе эксперимента. Достаточно вычислить ближайших соседей для каждой из выборок, а затем выбрать из них ближайших соседей.

6. Реализовать описанный выше алгоритм, основанный на преобразовании объектов тестовой выборки. Проверить то же самое множество параметров, что и в предыдущем пункте. Проанализировать как изменилась матрица ошибок, какие ошибки алгоритма помогает исправить каждое преобразование. Качественно сравнить два подхода (5 и 6 пункты) между собой.

Требования к реализации

Прототипы функций должны строго соответствовать прототипам, описанным в спецификации и проходить все выданные тесты. Задание, не проходящее все выданные тесты, приравнивается к невыполненному. При написании необходимо пользоваться стандартными средствами языка Python, библиотеками numpy и matplotlib. Библиотеками scipy и scikit-learn пользоваться запрещается, если это не обговорено отдельно в пункте задания. Для экспериментов по последним двум пунктам разрешается пользоваться любыми открытыми библиотеками, реализующими алгоритмы обработки изображений.

Замечание 1. Далее под выборкой объектов будем понимать `numpy array` размера $N \times D$ или разреженную матрицу `scipy.sparse.csr_matrix` того же размера, под ответами для объектов выборки будем понимать `numpy array` размера N, где N — количество объектов в выборке, D — размер признакового пространства.

Замечание 2. Для всех функций можно задать аргументы по умолчанию, которые будут удобны вам в вашем эксперименте.

Среди предоставленных файлов должны быть следующие модули и функции в них:

1. Модуль `nearest_neighbors`, содержащий собственную реализацию метода ближайших соседей.

   Класс `KNNClassifier`

   Описание методов:

   (a) `__init__(self, k, strategy, metric, weights, test_block_size)` — конструктор класса.

   - `k` — число ближайших соседей в алгоритме ближайших соседей
   - `strategy` — алгоритм поиска ближайших соседей. Может принимать следующие значения:
     - `'my_own'` — собственная реализация (например, на основе кода подсчёта евклидова расстояния между двумя множествами точек из задания №1)
     - `'brute'` — использование `sklearn.neighbors.NearestNeighbors(algorithm='brute')`
     - `'kd_tree'` — использование `sklearn.neighbors.NearestNeighbors(algorithm='kd_tree')`
     - `'ball_tree'` — использование `sklearn.neighbors.NearestNeighbors(algorithm='ball_tree')`
   - `metric` — название метрики, по которой считается расстояние между объектами. Может принимать следующие значения:
     - `'euclidean'` — евклидова метрика
     - `'cosine'` — косинусная метрика
   - `weights` — переменная типа `bool`. Значение `True` означает, что нужно использовать взвешенный метод k ближайших соседей. Во взвешенном методе ближайших соседей голос одного объекта равен $1/(distance + \varepsilon)$, где $\varepsilon = 10^{-5}$.
   - `test_block_size` — размер блока данных для тестовой выборки

   Замечание 1. Для некоторых алгоритмов поиска ближайших соседей вам потребуется хранить обучающую выборку и ответы на ней. Некоторые алгоритмы не требуют хранения выборки, но требуют хранения дополнительной информации о её структуре.

   Замечание 2. При поиске k ближайших соседей некоторые методы строят в памяти матрицу попарных расстояний обучающей выборки и тестовой выборки. Рекомендуется написать функцию, которая ищет ближайших соседей блоками, то есть делает запросы ближайших соседей для первых N тестовых объектов, затем для следующих N, и так далее, и в конце объединяет полученные результаты.

   (b) `fit(self, X, y)`

   Описание параметров:

   - `X` — обучающая выборка объектов
   - `y` — ответы объектов на обучающей выборке

   Метод производит обучение алгоритма с учётом стратегии указанной в параметре `strategy`.

   (c) `find_kneighbors(self, X, return_distance)`

   Описание параметров:

   - `X` — выборка объектов
   - `return_distance` — переменная типа `bool`

   Метод возвращает `tuple` из двух `numpy array` размера (`X.shape[0], k`). `[i, j]` элемент первого массива должен быть равен расстоянию от `i`-го объекта, до его `j`-го ближайшего соседа. `[i, j]` элемент второго массива должен быть равен индексу `j`-ого ближайшего соседа из обучающей выборки для объекта с индексом `i`.

   Если `return_distance=False`, возвращается только второй из указанных массивов. Метод должен использовать стратегию поиска указанную в параметре класса `strategy`.

   (d) `predict(self, X)`

   Описание параметров:

   - `X` — тестовая выборка объектов

   Метод должен вернуть одномерный `numpy array` размера `X.shape[0]`, состоящий из предсказаний алгоритма (меток классов) для объектов тестовой выборки.

2. Модуль `cross_validation` с реализациями функций для применения кросс-валидации:

(a) `kfold(n, n_folds)`

Описание параметров:

- `n` — количество объектов в выборке
- `n_folds` — количество фолдов на которые делится выборка

Функция реализует генерацию индексов обучающей и валидационной выборки для кросс-валидации с `n_folds` фолдами. Функция возвращает список длины `n_folds`, каждый элемент списка — кортеж из двух одномерных `numpy array`. Первый массив содержит индексы обучающей подвыборки, а второй валидационной.

(b) `knn_cross_val_score(X, y, k_list, score, cv, **kwargs)`

Описание параметров:

- `X` — обучающая выборка
- `y` — ответы объектов на обучающей выборке
- `k_list` — список из проверяемых значений для числа ближайших соседей, числа в списке упорядочены по возрастанию
- `score` — название метрики, по которой оценивается качество алгоритма. Обязательно должна быть реализована метрика 'accuracy' (доля правильно предсказанных ответов)
- `cv` — список из кортежей, содержащих индексы обучающей и валидационной выборки — выход функций `kfold` или `stratified_kfold`. Если параметр не задан, необходимо внутри функции реализовать генерацию индексов с помощью функции `kfold`
- `**kwargs` — параметры конструктора класса KNNClassifier

Функция для измерения метрики качества `score` алгоритма ближайших соседей, реализованного через класс `KNN_classifier` на кросс-валидации, заданной списком индексов `cv` для обучающей выборки `X`, ответов на ней `y`. Оценку качества метода ближайших соседей нужно рассчитать для нескольких значений k: $[k_1, \ldots, k_n]$, $k_1 < k_2 < \cdots < k_n$, заданных в `k_list`. Сложность алгоритма для одного объекта из валидационной выборки должна иметь порядок $O(k_n)$.

Функция должна возвращать словарь, где ключами являются значения k, а элементами — `numpy array` размера `len(cv)` с качеством на каждом фолде.

Замечание. Для тестирования алгоритма удобно использовать функцию `cross_val_score` из библиотеки `scikit-learn`.

Бонусная часть

1. (до 4 баллов) Написать юнит-тесты к написанному коду. Юнит-тесты должны быть реализованы в отдельном модуле, в отчёте должно быть прописано, как именно юнит-тесты должны быть запущены.

2. (до 3 баллов) Написать параллельную реализацию поиска ближайших соседей (например, с помощью библиотеки joblib)

3. (до 5 баллов) Улучшить качество работы метрических алгоритмов на датасете MNIST с помощью средств, не использующихся в задании. Например, можно реализовать ансамбль метрических алгоритмов, реализовать новые метрики, новые признаковые описания объектов. Размер бонуса зависит от величины улучшения и от изобретательности подхода.

4. (до 5 баллов) Качественное проведение дополнительного (не пересекающегося с основным заданием) исследования по теме метрических алгоритмов: формулируется изучаемый вопрос, ставятся эксперименты, позволяющие на него ответить, делаются выводы. Перед исследованием необходимо обсудить тему с преподавателем.