ДИСЦИПЛИНА	**Проектирование интеллектуальных систем (часть 1/2)**
ИНСТИТУТ	**информационных технологий**
КАФЕДРА	**вычислительной техники**
ВИД УЧЕБНОГО МАТЕРИАЛА	**Материалы для практических/семинарских занятий**
ПРЕПОДАВАТЕЛЬ	**Холмогоров Владислав Владиславович**
СЕМЕСТР	**6, 2023-2024**

Практическая работа № 6

«Ансамбли методов машинного обучения»

по дисциплине «Проектирование интеллектуальных систем (часть 1/2)»

Цели: приобрести навыки создание, обучения и применения ансамблей моделей машинного обучения.

Задачи:

1) Выполнить задачу или задачи классификации, регрессии, анализа и предсказания временных рядов и/или предсказательной аналитики с помощью ансамблей машинного обучения в соответствии со следующими пунктами:

– определить предметную область решаемой задачи, может совпадать с таковыми из предыдущих работ в соответствии с тематикой заданий (главное, чтобы модели имело смысл объединять в ансамбль, см. Примечание 1);

– найти или сгенерировать набор данных, содержащий данные для задачи выбранной предметной области;

– выполнить (хотя бы) минимальную предобработку данных аналогично предыдущим практическим работам;

– реализовать вручную ансамбль алгоритмов машинного обучения на основе простых ансамблевых приёмов (см. Примечание 2);

– реализовать хотя бы один продвинутый ансамбль – бустинг, бэггинг или стекинг (см. Примечание 3), провести сравнительный анализ по основным техническим характеристикам – затрачиваемому времени, памяти и качеству работы, сделать графики сравнения результатов работы с результатами алгоритмов, сделанных вручную, и предыдущих практических работ.

2) В качестве дополнительного задания реализовать хотя бы один из следующих пунктов (один – на оценку 4, все – на оценку 5):

– реализовать не менее двух ансамблей методов машинного обучения одной группы, провести сравнительный анализ и визуализацию;

– реализовать вручную ансамбли бустинга, бэггинга и стекинга (хотя бы один, см. Примечание 3), провести сравнительный анализ и визуализацию.

ПРИМЕЧАНИЕ:

1) Ансамблевое обучение позволяет повысить точность и устойчивость решения задачи за счёт объединения прогнозов из нескольких моделей/алгоритмов/методов, выделяют три проблемы, которые позволяют решить ансамблевые методы:

– статистическая проблема: пространство гипотез (они же предсказания) слишком велико для объема доступных данных, и, следовательно, существует множество гипотез с одинаковой точностью (параметры разные, а итог один), и алгоритм обучения выбирает только одну из них, тогда существует риск того, что точность выбранной гипотезы будет низкой на основе невидимых данных (например, тех же экзогенных параметров);

– вычислительная проблема: алгоритм не может гарантировать нахождение наилучшего результата;

– репрезентативная проблема: пространство гипотез не содержит хорошей аппроксимации целевого класса или классов (приближённые методы в принципе не могут адекватно описать оригинальную модель).

Как правило, ансамбли применяется в задачах классификации и регрессии (так как они обе имеют предсказательный характер), но могут и быть применены в других задачах вроде анализа временных рядов, реже при анализе ассоциативных правил и кластеризации, так как в основе ансамблей лежит сопоставление или усреднение результатов (как правило, результатов предсказания) нескольких слабых моделей (если речь идёт не о стекинге). Так при поиске ассоциативных правил использование ансамбля не имеет смысла, так как в большинстве случаев результаты работы этих алгоритмов будут одинаковы – всё отличие в том, какой из методов какие правила находит быстрее, поэтому ансамбль мог бы иметь смысл только в сопоставлении результатов по итерациям, когда один метод, например, производил бы поиск в ширину, а в другой – в глубину.

2) Под простыми ансамблевыми приемами, как правило, подразумевают следующие подходы:

– голосование (voting): подразумевает, что предсказания от большинства моделей (max voting) используются в качестве окончательного прогноза (обычно используется для задачи классификации);

– усреднение (averaging): подразумевает расчёт среднего значения прогнозов из всех моделей и использование его для составления окончательного прогноза (может быть использовано для прогнозирования в регрессионных задачах и при вычислении вероятностей для задач классификации);

– средневзвешенное значение (weighted average): подразумевает, что всем моделям присваиваются разные весовые коэффициенты (точнее – их результатам), определяющие важность каждой модели для прогнозирования, после определения весов (обычно в сумме дают единицу) результаты, умноженные на соответствующие значения весов, складываются, что и является итоговым предсказанным значением.

3) К продвинутым ансамблевым методам обычно относят следующие 3 группы ансамблей:

– бустинг (boosting): подразумевает последовательное соединение моделей (как правило, последовательность слабых однородных моделей с сильной в конце), в котором каждая последующая модель пытается исправить ошибки предыдущей, что выполняется с помощью назначения наибольшего веса неправильному предсказанию прошлой модели для будущей (на самом же деле увеличение веса обычно означает увеличение вероятность появления неправильно предсказанного набора данных на входе следующей модели), результат представляет собой средневзвешенное значение или выбор голосованием;

– бэггинг (bagging, bootstrap aggregation): подразумевает объединение результатов нескольких (как правило, слабых и однородных/одинаковых) моделей, сначала производится создание нескольких наборов данных как подмножеств начального набора с заменой (подразумевает, что одни и те же группы данных могут быть сразу в нескольких наборах, то есть наборы

пересекаются друг с другом), все модели обучаются и делают прогнозы параллельно, после чего все прогнозы агрегируются с помощью голосования или усреднения (вся суть подхода заключается в попытке понизить влияние дисперсии слабых моделей);

– стекинг (stacking): подразумевает объединение нескольких (как правило, сильных неоднородных) моделей для уменьшения их отдельной предвзятости предсказаний, в этом подходе обучающие данные могут быть разделены на несколько частей (разделение «K-fold»), при этом каждая из моделей по очереди обучается на определённой части данных, а потом делает для неё прогноз, после чего предсказания всех моделей по всем данным отправляются в общую метамодель, которая обучается на них, и именно её предсказания становятся окончательным решением задачи; существует также подход блендинга, который имеет ту же архитектуру, что и обычный стекинг, но для прогнозирования данных для метомодели используется только контрольный (проверочный) набор из обучающей выборки – то есть прогнозы делаются только на валидационной выборке.

К наиболее популярным моделям продвинутых ансамблей (в том числе реализуемым конкретными библиотеками) относятся:

– бустинг: AdaBoost, CatBoost, eXtreme Gradient Boosting, Gradient Boosting, Gradient tree boosting, LightGBM;

– бэггинг: Random Forest (и его разновидности вроде Balanced Random Forest и Rotation Forest), Bagging meta-estimator, Extra Trees (Extremely Randomized Trees), Pasting, BagBoo (Bagging of Boosted Trees);

– стекинг: StackingClassifier, BlendingClassifier, StackingRegressor (библиотечные решения), а также стекинг других ансамблевых алгоритмов.

ОСОБЫЙ БОНУС (доступен только в том случае, если выполнены пункты 2-го задания):

Реализовать ансамбль алгоритмов кластеризации (так как кластеризация является задачей неконтролируемого обучения, то классические методы

ансамблей вроде средневзвешенного значения или голосования методом большинства не подходят, так как для задачи кластеризации неважно, находится ли конкретная точка в кластере 1 или 2, важно то, какие точки также находится в этом кластере с ней, поэтому результаты ансамбля кластеризации определяются на основе матрицы подобия – the similarity matrix, в ней может отображаться информация о том, сколько раз одни и те же точки встречались вместе в результатах разных алгоритмов кластеризации, либо вероятность их совместного появления, также существует «граф связных компонентов» – Graph connected components, который предполагает, что точки, которые появляются вместе более чем в X% случаях, должны находиться в одном и том же кластере, данные подходы можно использовать совместно, в таком случае на основе матрицы подобия строится граф связей совместных встреч объектов, а после удаляются рёбра, имеющие меньше X% случаев совместной встречи).