Домашнее задание #4

Вопрос 1: Расскажите, как работает регуляризация в решающих деревьях, какие параметры мы штрафуем в данных алгоритмах?

Когда тренировочные наборы находятся слишком близко друг к другу, они имеют тенденцию к деградации в своей способности обобщать модель. Методы регуляризации используются для уменьшения эффекта переподгонки, устраняя деградацию за счет того, что процедура подборки ограничена.

Одним из популярных параметров регуляризации является параметр M, который обозначает количество итераций повышения градиента. M обозначает количество деревьев решений во всей модели, когда дерево решений является базовым обучаемым.

Увеличение количества итераций увеличения градиента снижает количество ошибок обучающего набора. Слишком большое количество повышающих итераций градиентов увеличивает переоснащение. Мониторинг ошибки прогнозирования из отдельного набора данных проверки может помочь в выборе оптимального значения для количества градиентов, повышающих итерации.

Помимо использования количества повышающих итераций градиентов в качестве параметра регуляризации, в качестве эффективного параметра регуляризации можно использовать глубину деревьев. При увеличении глубины деревьев модель, скорее всего, переместится на тренировочные данные.

Shrinkage - это процедура регуляризации градиента, которая помогает модифицировать правило обновления, чему способствует параметр, известный как скорость обучения. Использование коэффициента обучения ниже 0.1 приводит к улучшениям, которые являются существенными для обобщения модели.

Еще один метод регуляризации градиента заключается в штрафах за сложность деревьев. Сложность модели может быть определена как количество пропорциональных листьев деревьев. Оптимизация модели может быть выполнена путем обрезки деревьев для уменьшения сложности модели, что исключает любые ветви, которые не могут достичь пороговых потерь.

Вопрос 2: По какому принципу рассчитывается “важность признака (feature_importance)” в ансамблях деревьев?

Источник: Scikit-Learn

Отдельные деревья решений можно легко интерпретировать, просто визуализируя структуру дерева. Модели с градиентным увеличением, однако, состоят из сотен деревьев регрессии, поэтому они не могут быть легко интерпретированы визуальным осмотром отдельных деревьев. К счастью, был предложен ряд методик для обобщения и интерпретации моделей повышения градиента.

Зачастую характеристики не в равной степени способствуют прогнозированию целевой реакции; во многих ситуациях большинство характеристик на самом деле не имеют отношения к делу. При интерпретации модели первый вопрос, как правило, заключается в следующем: каковы эти важные особенности и как они способствуют прогнозированию целевого ответа?

Отдельные деревья решений по своей природе выполняют выделение признаков путем выбора соответствующих точек разделения. Эта информация может быть использована для измерения важности каждого признака; основная идея заключается в следующем: чем чаще признак используется в точках разделения дерева, тем более важным является этот признак. Это понятие важности может быть распространено на ансамбли деревьев решений простым усреднением примесной значимости каждого признака в каждом дереве (более подробную информацию см. в разделе "Оценка важности признака").

Оценки важности функции для модели увеличения градиента подгонки могут быть доступны через свойство feature_importances_.