Характеристики автоматной модели обучения.

Автор: Ricky Ho
Перевод: Черниченко Е. А.
Источник: http://horicky.blogspot.com/2012/02/characteristics-of-machine-learning.html

Аннотация

Статья посвящена обзору автоматной модели обучения и её характеристикам.

Для задачи классификации и регрессии есть различные варианты автоматной модели обучения, каждый из которых можно рассматривать как черный ящик для решения одинаковой проблемы. Однако все модели состоят из различных подходов и алгоритм будет выполняться по-разному в различных набора данных. Самый лучший способ заключается в использовании кросс-проверки, чтобы определить, какая модель наиболее эффективна в тестовых данных.

Дерево принятия решений на основе методов.

Фундаментальный подход обучения рекурсивно разделяет подготовку данных в ячейки однородных членов по наиболее дискриминационным критериям деления. Измерение "однородности" основано на выходной метке. Во время обучения различные критерии деления основываются на входных данных. Например, когда вход категории (Пн, Вт, Ср ...), то он сначала будет превращен в бинарный (isMon, isTue , isWed ...), а затем будет использоваться True/False в качестве решающей границы для оценки однородности; когда вход числовой или порядковый номер, LessThan, GreaterThan на каждом входе, то значение обучения будет использоваться в качестве решающей границы. Учебный процесс останавливается, когда нет значительного выигрыша в однородности дальнейшего раскол дерева. Члены ячеек, представленые на конечном узле будут голосовать за предсказание; большинство выигрывает, когда на выходе получаются категории и средний член, если выход числовой.

Преимущество дерева в том, что оно имеет очень гибкие с точки зрения типов данных входные и выходые переменные, которые могут быть категоричными, бинарными или числовыми. Уровень решения узлов также указывает на степень влияния различных входных переменных. Ограничением является каждое решение, которое граничит с каждой точкой разделения и является конкретным решением. Кроме того, критерии принятия решения рассматривают только один входной атрибут в момент времени, а не сочетание нескольких входных переменных. Другим недостатком является то, что дерево не может быть обновляться с приращением. При подготовке новых данных текстов, вы должны удалить старое дерево и подготовить все данные с нуля.

Тем не менее, дерево при смешивании с ансамблем методов (например, случайного леса, наращивание деревьев) затрагивает многие из ограничений, упомянутых выше. Например, градиент дерева решений повышает последовательность выполнения других моделей ML во многих задачах и является одним из самых популярных методов в наши дни.

Линейная регрессия на основе метода.

Основное предположение состоит в том, что выходную переменную (числовое значение) можно выразить в виде линейной комбинации (взвешенная сумма) набора входных переменных (числовое значение).
y = w1x1 + + w2x2 w3x3 ....

Вся цель этапа обучения состоит в том, чтобы узнать вес w1, w2 ... путем минимизации функции ошибки (y, w1x1 + w2x2 + ...). Градиентный спуск является классической техникой решения этой задачи с общей идеей регулирования W1, W2 ... в направлении максимального градиента функции потерь.

Входные переменные должны быть числовыми. Для бинарных переменных это будет представлено как 0, 1. Для категориальных переменных каждое возможное значение будет представлено как отдельная двоичная переменная 0, 1, соответственно). На выходе, если это двоичная переменная (0, 1), функция используется для преобразования диапазона от бесконечности до плюс бесконечности в 0 до 1. Это называется логистической регрессией на основе максимального правдоподобия.

Чтобы избежать переобучения, регуляризации (L1 и L2), используемого для указания большого значения W1, W2 ... L1, добавляются абсолютные значения w 1 в функцию потерь, в то время как L2 является добавлением квадрата w1 в функцию потерь. L1 обладает тем свойством, что оно будет указывать на избыточные функции или неактуальные функции (с очень малым весом), а также является хорошим инструментом для выбора весьма влиятельной особенности.

Сила линейной модели в том, что она имеет очень высокую производительность в обучении. Алгоритм стохастического градиентного спуска ориентированного обучения отличается высокой масштабируемостью и может обрабатывать дополнительные модели обучения.

Слабость линейной модели в том, что линейное предположение входных функций бывает ложным. Поэтому важным усилием инженерной функция является необходимость в преобразовании каждого входа, в котором обычно участвует эксперт предметной области. Другим распространенным способом является обычное выбрасываение различных функций преобразования 1 / x, x ^ 2, log (x) в надежде, что одна из них будет иметь линейную зависимость с выходом. Линейность может быть проверена путем наблюдения (y - predicted_y) нормального или не нормального распределения(с использованием QQplot с гауссовым распределением).

Нейронные сети.

Нейронные сети можно рассматривать как многослойный персептрон (каждая логистическая единица регрессии с несколькими двоичными входами и одним двоичным выходом). При наличии нескольких слоев это равносильно: Z = logit (v1.y1 v2y2 + + ...), a y1 = logit (w11x1 w12x2 + + ...)

Эта многослойная модель позволяет нейронной сети определить нелинейную зависимость между x входных и выходных Z. Типичная методика обучения является «отсталой распространенной ошибкой", где ошибка распространяется от выходного слоя назад к входному слой для регулировки веса.

Таким образом, следует обратите внимание, что нейронная сеть имеет двоичный вход, который указывает на то, что мы должны преобразовать входной сигнал на несколько двоичных переменных. Для числовых входных переменных мы можем преобразовать их в бинарную 101010 строку. Категоричные и числовые выводы могут быть преобразованы таким же образом.