Секция «Теория вероятностей и математическая статистика»

Особенности применения регрессионных моделей в прикладных исследованиях

Чокля Дмитрий
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: dimtry.choklya@yandex.ru

При использовании модели линейной регрессии, необходимо учитывать особенности структуры данных реальных задач: анализ финансовых данных, медицинские исследования или показ рекламы в интернете. Все эти задачи объединяет то, что потенциально можно задействовать огромное количество данных. При этом может быть как большое количество элементов обучающей выборки (напр. данные о сделках на бирже), так и большое количество признаков для каждого элемента выборки (напр. полногеномное исследование человека). Представленные проблемы могут также возникать одновременно: данные о пользовательском поведении и кликах в онлайн рекламе, могут состоять из порядка 10^{12} элементов при 10^8 и более различных признаков.

Еще одна проблема возникает с применением хорошо известного метода наименьших квадратов [3]. Напомним, что если задана модель вида $Y = X\beta + \varepsilon$, то оценка наименьших квадратов параметра β вычисляется как $(X^TX)^{-1}A^TY$. При таком подходе, необходимо вычислять обратную матрицу, что сопряжено с неустойчивостью ошибки при округлении чисел в компьютере. Поэтому, в работе будут описаны используемые на практике методы, такие как: стохастический градиентный спуск [5], моментум [4], метод адаптивного градиента [1].

Также, в работе будут рассмотрены особенности применения модели разреженной регрессии, когда предполагается, что для каждого объекта выборки возможно рассмотреть только k его признаков. Данное ограничение может быть вызвано как априорным знанием, так и невозможностью (или дороговизной) получения большого числа значений признаков. В условиях ограниченности знания, трудно определить, какие признаки самые важные для алгоритма. С помощью генерирования случайных подмножеств признаков, будет построен неэффективный с вычислительной точки зрения алгоритм, предсказания которого отличаются по средней L_2 норме от лучшей k-разреженной регрессии не больше чем на $O(\sqrt{T})$, после обработки T элементов выборки, где под k-разреженной регрессией понимается регрессия, у которой не более k ненулевых коэффициентов. И более того, не существует алгоритма, работающего за полиномиальное время в зависимости от количества элементов выборки, который был бы лучше предъявленного [2].

Показано, что при отсутствии дополнительных знаний о признаках, не только сложность самого алгоритма построения модели экспоненциально зависит от размера выборки, но еще и довольно трудно получить качественную оценку на размер ошибки.

Список литературы

[1] Duchi, John; Hazan, Elad; Singer, Yoram (2011). Adaptive subgradient methods for online learning and stochastic optimization. JMLR. 12: 2121–2159

[2] Foster, Kale, Karlof (2016). Online Sparse Linear Regression. JMLR. 49:1–11.

[3] Hayashi, Fumio (2000). Econometrics. Princeton University Press. p. 27-30

[4] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). Learning representations by back-propagating errors. Nature. 323 (6088): 533–536.

[5] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. Theory of Computing, 8(1):121–164, 2012.