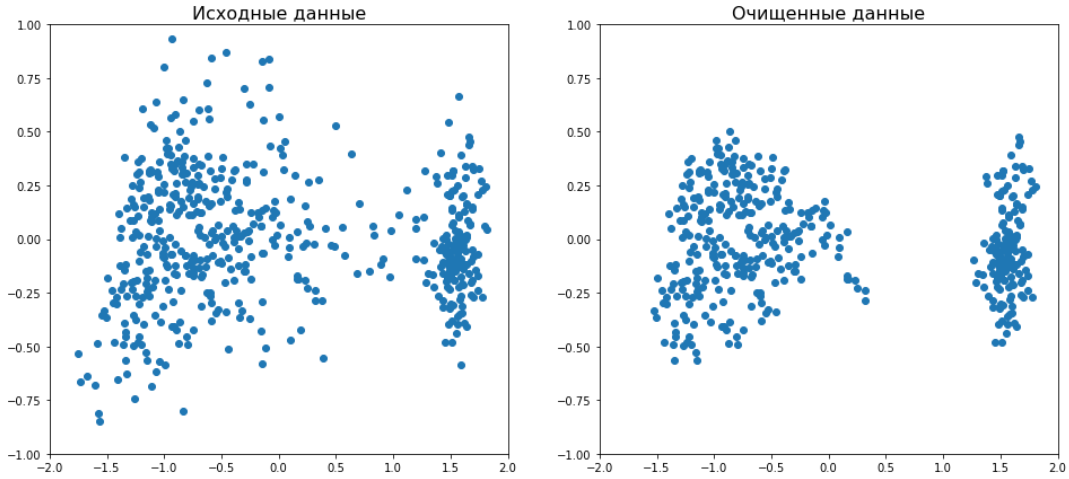

Рисунок 1 – Исходные и очищенные данные

Теперь посмотрим на результаты фильтрации выбросов алгоритмом DBSCAN. В связи с тем, что были убраны выбросы в данных, кластеры приобрели более четкие формы.

При применении метода K-средних на исходном наборе данных группа точек справа на 20% состояла из корпоративных пользователей и на 80% из персональных. То есть в этот кластер попало много объектов, которые изначально должны были попасть в другой кластер. Из-за некоторых аномалий в поведении эти пользователи находятся между двумя кластерами и при кластеризации отнесены к неправильной группе. Таким образом, центр кластера смещен относительно его правильного положения.

При использовании набора очищенных данных группа точек справа теперь только на 7% состоит из корпоративных пользователей, все остальные – персональные. Это значит, что теперь этот кластер более точно описывает группу персональных пользователей, центр кластера находится ближе к его истинному значению.

Таким образом, использование алгоритма кластеризации DBSCAN на практике позволяет очистить и подготовить набор данных для дальнейшего анализа, полученные при этом кластеры выделяются в более отчетливые формы, что несомненно приносит пользу, когда исследователю необходимо наиболее точно охарактеризовать свойства кластеров.

Список использованных источников:

1. Сегаран Т. Программируем коллективный разум. / Сегаран. Т. – Пер. с англ. – СПб: Символ-Плюс, 2008. — 50-51 с.

2. Comparing different clustering algorithms on toy datasets [Электронный ресурс]. — Режим доступа: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html/. — Дата доступа: 05.03.2019.

3. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. / Ester, M., Kriegel, H. P., Sander, J., & Xu, X. // KDD. — 1996. — Vol. 96, №34 — P. 226–231.

БАЙЕСОВСКАЯ ОПТИМИЗАЦИЯ

Козак А. В., Сухов Н. Ю.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Теслюк В. Н. – доцент, кандидат физико-математических наук

В данной работе рассмотрен алгоритм глобальной оптимизации функций без использования производных. Приведен класс функций, для которых может быть использована байесовская оптимизация. Рассмотрены потенциальные проблемы использования байесовской оптимизации. Приведены потенциальные области для улучшения и дальнейшего исследования данного класса оптимизационных алгоритмов.

Байесовская оптимизация – это подход к оптимизации целевых функций, оценка и подсчет которых требует большого числа ресурсов, например, времени (минуты или часы). Она лучше всего подходит для оптимизации непрерывных функций, определенных на многомерном пространстве размерности не более 20 и допускающих стохастический шум в значениях самой функций.

Байесовская оптимизация – это класс алгоритмов оптимизации, основанных на принципах машинного обучения и сфокусированных на решении оптимизационной задачи вида:

$$f(x) .$$

При этом множество допустимых значений (A) и целевая функция ($f(x)$) удовлетворяют следующим свойствам:

1. Целевая функция определена на многомерном пространстве, то есть $f: R^d \to R$. При этом для достижения наибольшей эффективности байесовской оптимизации как правило выполняется следующее неравенство $d \leq 20$.
2. Множество допустимых значений A является простым, то есть для любой точки из R^d легко понять принадлежит она A или нет. Как правило, A является многомерным параллелепипедом (то есть $A \in \{x \in R^d: a_i \leq x_i \leq b_i\}$ или d-мерным симплексом (то есть $A \in \{x \in R^d: \sum_i \quad x_i = 1\}$. Однако данное свойство является лишь опциональным.
3. Целевая функция должна быть непрерывной. Это ограничение необходимо для регрессии гауссовского процесса.
4. Подсчет значения целевой функции в точке является «дорогостоящей» операцией в том смысле, что для нахождения значения необходимо произвести несколько сотен вычислительных операций при этом каждая из таких операций требует также большого количества вычислительных ресурсов, что в свою очередь является «роскошью» и требует большого количества временных ресурсов (например, несколько дней).
5. Целевая функция не имеет известной для нас структуры, то есть мы не знаем ничего об этой функции, например, линейность или выпуклость. Будем считать, что мы знаем только алгоритм вычисления, но не знаем выражение функции в элементарных функциях, что делает невозможным найти производную целевой и, как следствие, использовать стандартные методы оптимизации. Будем говорить, что целевая функция – «черный ящик».
6. Нет возможности вычислить не первую, не вторую производную, то есть знания о целевой функции являются «свободными» от производной, как результат нет возможности использовать метод Ньютона и другие.
7. Целевая функция может быть зашумленной, при этом будем считать, что шум является независимым от значений целевой функции.
8. Наша задача найти глобальный оптимум, нежели локальный.

Таким образом, мы можем сказать, что байесовская оптимизация предназначена для решения задач глобальной оптимизация «черного ящика» без производных.

Способность оптимизации такого класса целевых функций без использования производных делает область применения байесовской оптимизации очень разносторонней. Так стало популярным использовать данный класс оптимизационных алгоритмов в машинном обучение для подборки гиперпараметров. В течение более длительного периода, начиная с 1960-х, байесовская оптимизация использовалась для проектирования инженерных систем. По мимо этого, оптимизацию, основанную на байесовской статистике, использовали для выбора экспериментов, которые дадут наиболее значимые результаты, в области разработки лекарственных средств, для калибровки моделей окружающей среды и в обучении с подкреплением (класс задач в машинном обучении).

Помимо байесовской оптимизации существуют и другие методы позволяющие решать задачу глобальной оптимизации «дорогостоящий черных ящиков». Большинство этих методов схожи на байесовскую оптимизацию: они используют так называемую суррогатную функцию, которая моделирует нашу целевую функцию и которую в дальнейшем используют для выбора точки подсчета значения целевой функции, тем самым уточняя множество точек потенциальных на место оптимума. Такой класс методов оптимизации называется суррогатными методами. Байесовская оптимизация отличается от других суррогатных методов использованием суррогатов, смоделированных с использованием байесовской статистики, и при принятии решения о том, где оценивать целевую функцию, использует байесовскую интерпретацию этих суррогатов.

Алгоритм байесовской оптимизации состоит из двух компонент: модели байесовской статистики для моделирования целевую функцию и дополнительно функции принятия решения того, где производить подсчет целевой функции. После оценки целевой функции согласно первоначальному экспериментальному предположению заполнения пространства возможных значений (то есть выбранных точек), часто состоящему из точек, выбранных равномерно случайным образом, алгоритм

использует итеративную процедуру для распределения оставшейся части «бюджета» подсчета целевой функции N.

Статистическая модель, которая неизменно (так называемое свойство инвариантности) является гауссовским процессом, обеспечивает байесовскую апостериорную вероятность распределение неизменной по своей структуре с апостериорным распределением, что упрощает процесс оптимизации, описывающее потенциальные значения для $f(x)$ в точке-кандидате x. Каждый раз, когда мы наблюдаем данные, новая точка приводит к обновлению и уточнению апостериорного распределения.

В этой увлекательной и необъятной области байесовских оптимизационных алгоритмов существует множество направлений для исследований. Во-первых, область для более глубокого теоретического понимания и освоения методов байесовской статистики. Так например, ничего нельзя сказать о скорости сходимости данного алгоритма к оптимуму и требует отдельного теоритического исследования. Во-вторых, исследование других методов построения байесовской статистики, отличных от гауссовского процесса. В-третьих, исследование и разработка алгоритмов байесовской оптимизации, работающих на целевых функциях любой размерности, тем самым снимая ограничение номер 1 (описанных выше) с оптимизируемой функции.

Несмотря на обширную область неизвестности байесовских алгоритмов, они предоставляют нам обширных класс решения эффективной оптимизации, когда стандартные методы являются нецелесообразными.

Входе работы были получены следующие результаты: действительно, байесовская оптимизация справляется эффективнее в поиске гиперпараметров нейросетевой модели, чем стандартные методы оптимизации, более того, байесовская оптимизация не требует нахождения производной, что является дорогостоящим (в плане вычислительных ресурсов) для поиска оптимальных гиперпараметров модели. Возможность оптимизации и модификации алгоритма байесовской оптимизации под алгоритм обучения моделей машинного обучения (особенно в байесовском нейросетевом моделировании) повышает эффективность использования обученной модели как качественно, так и в плане экономичности использования вычислительных ресурсов в ходе обучении. Все, выше перечисленное, является неотъемлемым преимуществом для дальнейшей работы в моей магистерской диссертации по использованию байесовских методов (в том числе и оптимизационных) в анализе медицинской информации (где получение значения целевой функции является «роскошью»).

Список использованных источников:

1. Berger J. O. – Statistical Decision Theory and Bayesian Analysis, 2013,
2. Blum J. R. – Multidimensional stochastic approximayion methods, 1954,
3. Gelman A., Carlin J. B., Stern H. S. – Bayesian Data Analysis, 2014.

АВТОМАТИЗИРОВАННАЯ СИСТЕМА УПРАВЛЕНИЯ ВЗАИМООТНОШЕНИЯМИ С КЛИЕНТАМИ С ПРИМЕНЕНИЕМ ОБЛАЧНЫХ CRM-СИСТЕМ

Козлова А. А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Калугина М.А. – к. ф.-м. н., доцент

Интерес к CRM-системам непрерывно растёт, поскольку они решают проблемы загрузки ресурсов и эффективности работы бизнеса. Именно поэтому этот класс программного обеспечения заслуживает пристального внимания. В докладе приведены результаты исследования работы CRM-системы SAP Hybris Cloud For Customers, а также предложено значительное улучшение в направлении автоматизации работы менеджеров с целевыми клиентами согласно бизнес-процессу корпоративного предприятия.

Важным аспектом функционирования бизнеса является процесс взаимодействия с клиентами. В силу растущей конкуренции и компьютеризации бизнеса неотъемлемой частью успешных компаний на сегодняшний день считают наличие CRM-систем [1]. Система управления взаимоотношениями с клиентами (Customer Relationship Management, CRM) – это комплексный подход, бизнес-стратегия к выявлению, приобретению и сохранению клиентов [2]. Основная задача CRM – получать на базе собранных данных информацию, которую можно использовать непосредственно для повышения