7. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [Электронный ресурс] – Режим доступа: http://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf

ОЦЕНКА АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ. ДИЛЕММА СМЕЩЕНИЯ-ДИСПЕРСИИ

Тишковский М.А., Лимонтов А.С., Подвальников Д.С.

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

Чернявский Ю.А. – к.т.н.,доцент

В данной работе рассмотрена оценка алгоритмов машинного обучения, а также дилемма смещения-дисперсии. Были проведены эксперименты по оценке различных алгоритмов машинного обучения через аппроксимацию функции косинуса, результаты которых представлены в работе.

В настоящее время с помощью алгоритмов машинного обучения решают все большее количество задач. В связи с этим возникает проблема, как правильно подобрать алгоритм машинного обучения под конкретную задачу, а также как оценить полученную натренированную с помощью этого алгоритма модель. Вторая задача решается с помощью введения правильных метрик оценки качества, что само по себе тоже является непростой задачей.

Задача выбора алгоритма машинного обучения является достаточно сложной и требует анализа исходных данных, понимания доменной области, знаний особенностей семейств алгоритмов. В связи с этим вводится понятие дилеммы смещения-дисперсии [1], по которому модели с меньшим смещением в параметре оценки имеют более высокую дисперсию, и наоборот, что в терминах машинного обучения является проблемой недообученности или переобученности алгоритма.

Переобучение [2] – явление, когда построенная модель хорошо объясняет примеры из обучающей выборки, но относительно плохо работает на примерах, не участвовавших в обучении.

Недообучение случается, когда алгоритм машинного обучения не может найти закономерности между исходными данными и целевой зависимостью.

Рассмотрим задачу аппроксимации функции косинуса на промежутке от 0 до 2π с помощью различных алгоритмов. На рисунке 1 представлена аппроксимация с помощью линейной регрессии [3], на рисунке 2 с помощью метода опорных векторов [4] с ядром радиальной базисной функции с параметром γ=1, на рисунке 3 с помощью метода опорных векторов с ядром радиальной базисной функции с параметром γ=100.

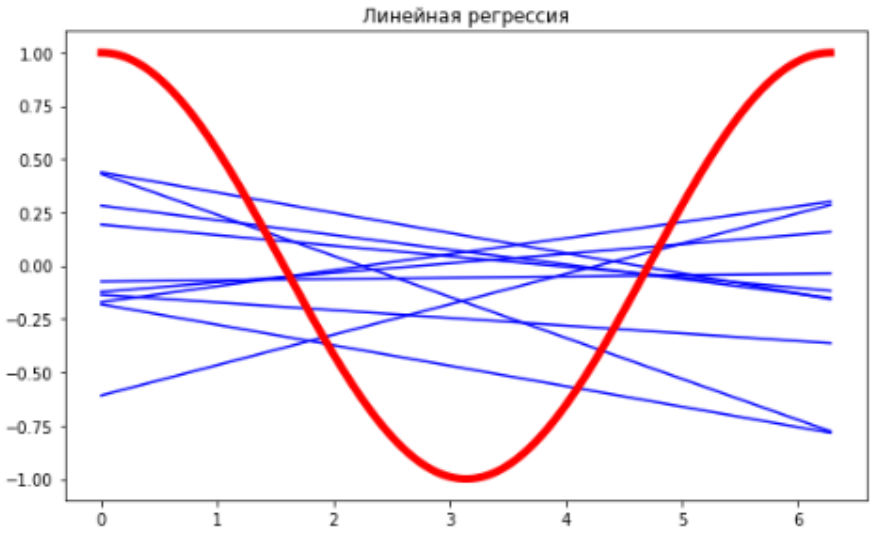

Рисунок 1 – Аппроксимация функции cos(x) с помощью линейной регрессии

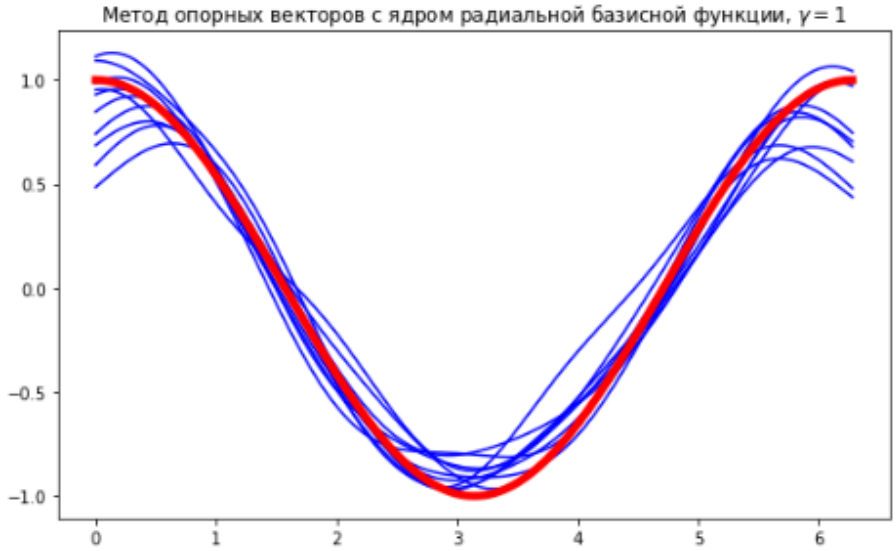

Рисунок 2 – Аппроксимация функции cos(x) с помощью метода опорных векторов с ядром радиальной базисной функции с параметром γ=1

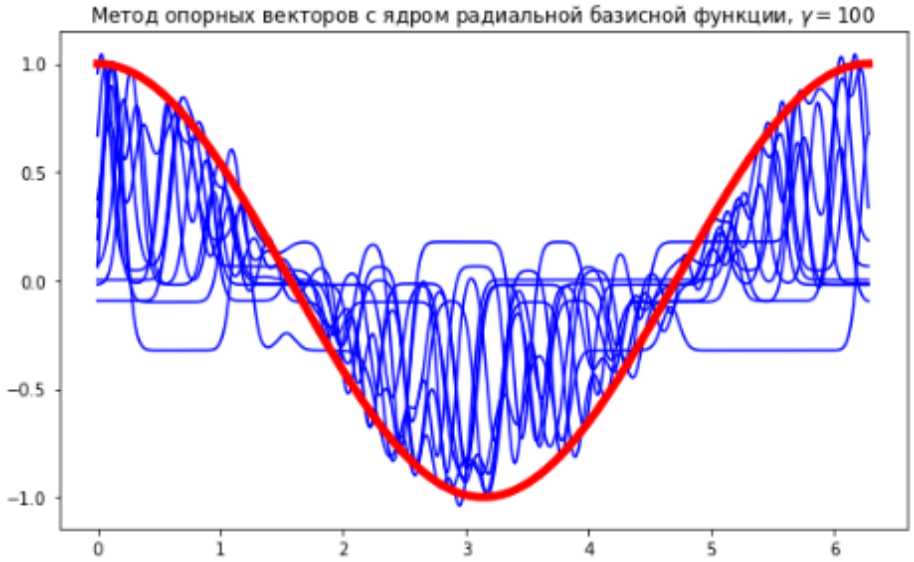

Рисунок 3 – Аппроксимация функции cos(x) с помощью метода опорных векторов с ядром радиальной базисной функции с параметром γ=100

Проанализируем результат аппроксимации с помощью линейной регрессии. Результаты имеют достаточно средний разброс, что скорее всего обусловлено количеством точек, на которых тренировался алгоритм. Также можно заметить, что результаты имеют очень большое смещение. В связи с этим можно сделать вывод о плохой аппроксимации функции косинуса с помощью алгоритма линейной регрессии.

Проанализируем результат аппроксимации с помощью метода опорных векторов с ядром радиальной базисной функции с параметром γ=1. Результаты имеют очень маленькую дисперсию, что говорит об устойчивости алгоритма. Также можно наблюдать очень маленькое смещение аппроксимации относительно графика настоящей функции, что говорит о том, что алгоритм является хорошим приближением исходной функции.

Проанализируем результат аппроксимации с помощью метода опорных векторов с ядром радиальной базисной функции с параметром γ=100. Результаты имеют большую дисперсию, что говорит о неустойчивости алгоритма. Также мы видим достаточно большое смещение, что говорит о том, что приближение с помощью данного алгоритма является плохим. Причиной этого эффекта является излишне сильная настройка весов модели на тренировочные данные, что ведет к плохой аппроксимации на данных, которые алгоритм еще не видел. В терминах машинного обучения данный эффект называется переобучением.

Список использованных источников:

1. Bias-variance tradeoff [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

2. Overfitting [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Overfitting
3. Linear regression [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Linear_regression
4. Support-vector machine [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Support-vector_machine

АНАЛИЗ ЭФФЕКТИВНОСТИ РЕКЛАМЫ КАК ETL ПРОЦЕСС

Харитонов Н.В., Хоронеко М.П., Медунецкий М.А.

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

Стержанов М.В. – к.т.н.,доцент

В данной работе рассматривается механизм оценки эффективности интернет рекламы при помощи стека технологий Big Data. Речь пойдет о ETL процессе (реализованном на планировщике задач Luigi) с помощью которого мы собираем, храним и обрабатываем большие объемы данных. В качестве результата будет представлена архитектурная схема платформы, осуществляющей данный процесс, и пример возвращаемых данных - еженедельный отчет, включающий в себя статистические данные рекламной кампании. Эволюция вычислительных систем обусловлена развитием задач бизнеса, а именно задач сбора данных, хранения и обработки полученных результатов. Конкурирующими показателями для данных систем становятся такие характеристики как скорость работы, отказоустойчивость, безопасность и защищенность от внешних воздействий. Глобальная информатизация привела к тому, что централизованные информационные системы становятся достаточно уязвимыми. Не менее уязвимым становится бизнес, работа которого зависит от качественной работы информационных систем.

В настоящее время наблюдается устойчивый рост интереса к практическому применению технологий Big Data в сфере маркетинга. Среди решаемых проблем можно выделить задачи повышения конкурентоспособности, создания новых услуг, совершенствования управления взаимоотношениями с клиентами. В результате развития Интернет, социальных сетей, и иных сетевых сервисов непрерывно растут потребности в информационных продуктах и услугах. Чтобы предлагать клиентам такие услуги, предприятиям приходится анализировать большие объемы данных из различных источников. Поэтому накопленная информация становится стратегически важным активом, от эффективности управления которым существенно зависят результаты деятельности предприятий.

В рамках данной работы мы представляем платформу для анализа эффективности интернет рекламы с использованием стека технологий Big Data.

Big Data [1] – обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

Нами предлагается использование следующих параметров для оценки эффективности интернет-рекламы: количество просмотров; количество переходов; количество нажатий мыши по рекламной области; местоположения; время просмотра. Мы логируем все доступные акты показа рекламы, агрегируем их, проводим обработку, а затем создаем отчет за заданный период времени (неделя, месяц, год).

Отчет включает в себя: данные о рекламодателе; рассмотренный период; количество показов; количество переходов; отношение переходов к показам; количество уникальных местоположений; количество уникальных пользователей; количество уникальных устройств и т. д.

Для более детального описания, представим архитектуру платформы в виде диаграммы на рисунке 1:

Представление архитектуры и стека используемых технологий

Архитектура системы представляет собой ETL[2] процесс, организованный на кластере компьютеров с использованием технологий Hadoop и HDFS. Рассмотрим каждый из шагов данного процесса.

Extract. На этом шаге мы собираем данные из различных источников (массива социального Веба, логов действий пользователей, корпоративных баз данных, разнообразных датчиков, внешнего набора данных и т.д.), а затем передаем их на следующий этап для выполнения преобразований. В данном случае для извлечения данных нами используются RTB[3] аукционы и подход “1x1 pixel”. Каждый показ рекламы записывается в HDFS[4] базу данных (см. hdfs://logs на Рис. 1). Каждая запись