УДК 004.85

ПРИКЛАДНЫЕ ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

А. В. Кушнеров[1)], М. Д. Гершов[2)]

[1), 2)] *Белорусский государственный университет, Беларусь, Минск,*
[1)] *al.v.kushnerov@gmail.com,*[2)] *michaelgershov@yandex.by*

Материал посвящён особенностям изложения материала по одной из дисциплин компьютерной математики – машинного обучения. Выделены основные этапы предполагаемого освоения материала. Также внимание уделено основным проблемам методического и технического характера, с которыми могут столкнуться преподаватели и обучающиеся. В работе также приведён пример построения конкретной модели классификатора, решающей задачу из области прикладной медицины.

Ключевые слова: машинное обучение; прикладное программирование; математические специальности; компьютерная математика.

APPLIED MACHINE LEARNING

A. V. Kushnerov[1)], M. D. Gershov[2)]

[1), 2)] *Belarussian state university, Belarus, Minsk,*
[1)] *al.v.kushnerov@gmail.com,*[2)] *michaelgershov@yandex.by*

The article is devoted to the material presentation problem on one of the computer mathematics disciplines - machine learning. It contains a description of the main steps, which may arise during machine learning studying. In addition, attention was paid to methodological and technical problems that teachers and students may encounter. In addition, the paper provides an example of solving a specific problem.

Keywords: machine learning; applied programming; mathematical specialties; computer mathematics.

Введение

Компьютерная математика представляет собой достаточно обширную область прикладного программирования. Отличительной особенностью типичных для неё задач является прочная математическая основа компьютерной модели. Идея заключается в изучении специфических задач, которые требуют от специалиста как математических знаний, так и навыков разработки программных продуктов. Несмотря на наличие в современных средах разработки готовых решений типичных математиче-

ских задач, предполагается, что обучающийся должен владеть навыками для их реализации с нуля.

Типичный пример дисциплины компьютерной математики – машинное обучение (МО). Машинное обучение представляет собой большой класс методов для анализа и обработки данных в подавляющем большинстве математическими методами. Актуальность задач МО в современном мире очевидна и неоспорима. Алгоритмы МО нашли своё применение в обработке изображений, текстов, числовых данных, медицине, спорте, построении рекомендательных систем и других сферах. Изучение студентами подобных дисциплин отличная возможность одновременно получить навыки, востребованные на современном рынке труда, а также применить и развить свои математические знания.

Математические модели для задач машинного обучения

Практически все методы и алгоритмы машинного обучения опираются на материал из классических математических дисциплин. В общем виде задача машинного обучения предполагает получение некоторой функциональной оценки данных, которые характеризуются набором числовых признаков. Как правило предполагается наличие некоторой зависимости между исходными данными, которую алгоритму и следует выявить.

В частности, задача классификации объектов по некоторому набору из m меток $\{y_1, y_2, \dots y_m\}$ предполагает нахождение некоторой «функции» f такой что $f\left(\bar{x}\right) = y_i$ для конкретного экземпляра исходных данных из n признаков $\bar{x} = \left(x_1, x_2, \dots x_n\right)$.

Первый этап работы с моделью МО — это *предварительная обработка и анализ данных*. Предполагается применение приёмов математической статистики. В частности, обычно выделяют статистики центральной тенденции (среднее значение, медиана, мода), статистики разброса (дисперсия, стандартное отклонение, интерквантильный интервал) и статистики формы, позволяющие сделать предположении о конкретном распределении выборки по тому или иному признаку. В ходе работы с исходными данными можно также провести предварительный отбор признаков.

Следующим этапом является *выбор и построение модели* исходя из поставленной. На сегодняшний день известно множество геометрических, вероятностных и логических моделей МО, таких как логистическая и линейная регрессии, дерево решений и случайный лес, метод главных компонент, Гауссова смесовая модель и многие другие. Для их освоения

респонденту необходимы знания главным образом математического анализа, линейной алгебры, теории оптимизации и математической статистики.

Для примера рассмотрим модель логистической регрессии, которая предполагает применение к вектору исходных данных $\bar{x}$ следующей функции:

$$f(\bar{x}, \bar{w}) = \frac{1}{1+e^{-\bar{w}^T \bar{x}}} \ .$$

Непосредственно обучение в данном случае предполагает нахождение вектора весов $\bar{w}$ с помощью решения следующей задачи оптимизации:

$$L_{\log}(X, \bar{y}, \bar{w}) = \sum_{i=1}^{n} \left(-y_i \ln f\left(\bar{w}^T \bar{x}_i\right) - (1-y_i) \ln\left(1 - f\left(\bar{w}^T \bar{x}_i\right)\right)\right) \to \min \ ,$$

где X - матрица исходных признаков, а $\bar{y}$ - вектор исходных целевых меток. А в совокупности эти данные принято называть обучающей выборкой.

Как правило студенты математических специальностей имеют достаточный уровень подготовки для освоения описанных выше методик, но нужно понимать, что порой требуются дополнительная подготовительная работа в рамках курса МО.

Также стоит отметить, что важным этапом также является *оценка качества полученной* модели. Методы оценки разнятся в зависимости от конкретных моделей, но в целом также базируются на математических вычислениях [1].

Компьютерные модели для задач машинного обучения

Описанные выше математические модели МО безусловно требуют компьютерной реализации. Выбор среды разработки сегодня весьма широкий так как практически все поставщики девелоперского ПО обзавелись инструментами для реализации моделей машинного обучения.

Наиболее популярной средой для решения подобных задача сегодня является *python*. *Python* – бесплатный кроссплатформенный язык программирования с относительно низким порогом входа для новичка. Более того, наличие огромного количества встроенных математических возможностей за последние годы сделало *python* профильным языком для анализа данных и МО, в частности.

В процессе изучения компьютерных моделей МО обязательно необходимо производить обзор возможностей языка в контексте решаемых

задач. Важно учитывать актуальность используемого программного обеспечения, которое хоть и обладает обратной совместимостью, но предполагает оптимизацию проблем в актуальных версиях. Это требует от преподавателя постоянного изучения актуальных тенденций обновлений и плотную работу с документацией соответствующих библиотек.

Также остро стоит проблема готовых решений при компьютерном моделировании. По сути, разработчик имеет возможность реализовать достаточно объёмные модели с помощью готовых решений. При изучении материалов по МО это нужно учитывать. Порой для лучшего освоения материала, стоит отказаться от готовых решений и реализовать ту или иную модель «с нуля». Такое подход позволяет сохранить концепцию неразрывности математической и компьютерной моделей МО. Более того, понимание математической парадигмы позволяет аналитику данных быть более гибким в подборе параметров моделей и оценке их качества [2].

Пример модели классификатора на основе случайного леса

Используя имеющуюся лицензированную базу данных Кливленда, которая содержит в себе 920 записей о наличии или отсутствии ССЗ, нам удалось реализовать качественную модель МО на основе случайного леса.

Работа включала в себя несколько этапов:

1. Анализ и обработка данных;
2. Обучение модели МО и подбор гиперпараметров;
3. Оценка качества модели с применением различного рода метрик.

На первом этапе нами была проведена первичная обработка признаков: удаление дубликатов, выбросов, а также корректировка нереалистичных значений. В связи с особенностями предметной области выбросами мы считали данные, находящиеся на расстоянии 1.7×IQR, где IQR – межквартильный диапазон.

Для заполнения пропусков использовался многомерный метод, основанный на группировке признаков и выведении внутригруппового значения. Стандартизация признаков была необходима для того, чтобы числовые данные были правильно интерпретированы моделью МО.

В первом случае для кодирования категориальных переменных использовался метод One-Hot Encoder, опирающийся на создание бинарных признаков, показывающих принадлежность к уникальному значению, а во втором – метод Target Encoder, который использует метку для кодирования категориальных признаков.

На этапе обучения для одной из модели мы использовали метод главных компонент PCA, взяв соответственно 3, 5, и 7 вычисленных главных компонент.

Подбор гиперпараметров моделей осуществлялся с использованием функции GridSearchCV, которая реализует метод поиска гиперпараметров по сетке.

Для правильного выбора метрики оценки качества следует обратить внимание на предметную область, из которой данные были получены. Нами был сделан вывод о том, что главной характеристикой, определяющей качество нашей модели, стоит считать процент ошибки первого рода, то есть, когда больной человек был классифицирован как здоровый.

Наиболее качественная построенная нами модель имеет следующие оценки точности:

	precision	recall	f1-score	support
Здоровый	0.79	0.81	0.80	118
Больной	0.85	0.83	0.84	149
accuracy			0.82	267
macro avg	0.82	0.82	0.82	267
weighted avg	0.82	0.82	0.82	267

где:

- precision – доля объектов, названных классификатором положительными и при этом действительно являющимися положительными;
- recall показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм;
- f1-score – F-мера;
- support – количество объектов обучающей выборки.

Заключение

Машинное обучение – современное и развивающееся направление, которое сочетает в себе элементы классический математики и программирования. Изучение подобных дисциплин способствует разностороннему развитию обучающихся и делает их востребованными на рынке труда. Множество сфер приложения позволяет развивать способности к решению прикладных математических задач на важных и нужных примерах. Также стоит отметить, что актуальность направления делает его крайне перспективным для научных исследований.

Библиографические ссылки

1. *S. Rashka, V. Mirjalili.* Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. 2nd Edition. Packt Publishing, 2017. ISBN 9781787125933.

2. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014.