МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРИМЕНЕНИЕ АЛГОРИТМОВ КЛАССИФИКАЦИИ К АНАЛИЗУ ДАННЫХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Максимкиной Анастасии Эдуардовны

Научный руководитель
д. ф.-м. н., доцент _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы исследования. Машинное обучение представляет собой обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов для анализа данных и получения выводов и выноса решения или предсказания в отношении чего-либо. Подход, при котором прошлые данные или примеры используются для первоначального формирования и совершенствования схемы предсказания, называется методом машинного обучения. Общая задача машинного обучения заключается в восстановлении зависимости между входными и выходными элементами с целью предсказания будущего выхода по заданному входу. Целью машинного обучения является построение максимально точной модели на основе данных и затем использования этой модели для предсказаний в будущем.

В зависимости от наличия или отсутствия прецедентной информации различают ряд категорий машинного обучения: контролируемое обучение или “обучение с учителем”, неконтролируемое обучение, обучение с подкреплением. Большая вариативность позволяет применять методы машинного обучения для данных различных типов в самых разных областях: биоинформатике, медицинской диагностике, технике. Широкий спектр приложений методы машинного обучения получили в экономике. Так, они используются для обнаружения мошенничества, кредитного скоринга, биржевого технического анализа. В современных условиях функционирования социально-экономических систем проблема получения приемлемого прогноза может быть решена за счет комбинирования традиционных классических методов совместно с методами интеллектуального прогнозирования.

Для получения релевантных результатов необходимы подходящие инструменты и корректные алгоритмы, в связи с чем, методы машинного обучения и data mining получили широкое применение при анализе медицинских данных и кредитного скоринга. Существует множество методов машинного обучения эффективно применяемых для данного класса задач: искусственные нейронные сети, деревья принятия решений, логистическая регрессия, генетический алгоритм.

Актуальность определила выбор темы данной работы: "Применение ал-

горитмов классификации к анализу данных".

Цель работы: получение теоретических и практических навыков построения логистической регрессии и деревьев решений.

Для достижения поставленных целей в работе необходимо решить следующие задачи:

— Применение методов классификации на практике.
— Проведение классификации методом:
  – логистической регрессии;
  – дерева решений.
— Автоматизация решения задачи классификации на языке R.

Практическая значимость работы заключается в разработке программных продуктов для построения логистической регрессии и дерева решений.

Основное содержание работы

Бакалаврская работа состоит из: введения, трёх теоретических и двух практических глав, заключения, списка использованных источников, приложения.

Введение содержит основные положения: актуальность темы исследования (цель, объект, предмет, задачи исследования); практическую значимость исследования.

Первый раздел «Задача классификации» описывает теоретические основы классификации.

Проведение классификации в R. В задаче классификации зависимая переменная является категориальной, то есть может принимать конечное число значений. Классификация относится к классическим задачам машинного обучения. Состоит в прогнозировании класса входного вектора на основе одной зависимой переменной.

Также будет рассмотрен вопрос, связанный с оценкой качества модели классификации. Алгоритм классификации вычисляет вероятность принадлежности к одному классу, затем входному вектору присваивается тот класс, вероятность принадлежности к которому вектора больше.

Практическое применение методов классификации. Проблемы, при решении которых возникает задача классификации:

— классификация, как необходимый предварительный этап статистической обработки данных;

— классификация в задачах прогнозирования экономико-социологических ситуаций для отдельных показателей.

Постановка задачи классификации. Задача заключается в том, чтобы построить такую программу, которая, используя обучающую последовательность, вырабатывала бы правило, позволяющее классифицировать вновь предъявляемые «незнакомые» ситуации (вообще говоря, отличные от входивших в обучающую последовательность).

Способность к обучению характеризуется двумя понятиями:

— качеством полученного решающего правила (вероятностью неправильных ответов — чем меньше эта вероятность, тем выше качество);

— надежностью получения решающего правила с заданным качеством (вероятностью получения заданного качества — чем выше эта вероятность, тем выше надежность успешного обучения).

Задача сводится к созданию такого обучающего устройства, которое по обучающей последовательности строило бы решающее правило, качество которого с заданной надежностью было бы не ниже требуемого.

Математическая постановка задачи обучения. В среде, которая характеризуется распределением вероятностей $P(x)$, случайно и независимо появляются ситуации x. Существует «учитель», который классифицирует их, то есть относит к одному из k классов (для простоты $k = 2$). Пусть он делает это согласно условной вероятности $P(t|x)$, где $t = 1$ означает, что вектор x отнесен к первому классу, а $t = 0$ — ко второму. Ни характеристика среды $P(x)$, ни правило классификации $P(t|x)$ нам не известны. Однако известно, что обе функции существуют, то есть существует совместное распределение вероятностей

$$P(x, t) = P(x) \cdot P(t|x).$$

Пусть теперь определено множество Ω решающих правил $F(x, \alpha)$. В этом множестве каждое правило определяется заданием параметра α (обычно это вектор). Все правила F(x,α) — характеристические функции, то есть

могут принимать только одно из двух значений — нуль или единицу:

$$F(x,\alpha) = \begin{cases} 1, x - \text{принадлежит первому классу}, \\ 0, x - \text{принадлежит второму классу}, \end{cases}$$

Для каждой функции $F(x,\alpha) \in \Omega$ может быть определено качество $Q(\alpha)$ как вероятность различных классификаций ситуаций x с учителем.

1. В случае, когда пространство X дискретно и состоит из точек $x^1, ..., x^N$

$$Q(\alpha) = \sum_{t=0}^{1} \sum_{t=1}^{N} (t - F(x^i,\alpha))^2 P(x^i) P(t|x^i),$$

где $P(x^i)$ — вероятность возникновения ситуации x^i.

2. В случае, когда в пространстве X существует плотность распределения $p(x)$,

$$Q(\alpha) = \sum_{t=0}^{1} \int (t - F(x,\alpha))^2 p(x) P(t|x) dx.$$

3. В общем случае можно сказать, что в пространстве X задана вероятностная мера $P(x,t)$, тогда

$$Q(\alpha) = \int\limits_{x,t} (t - F(x,\alpha))^2 dP(x,t).$$

Среди всех функций $F(x,\alpha)$ есть такая $F(x,\alpha^0)$, которая минимизирует вероятность ошибок. Эту функцию (или близкую к ней) и следует найти. Так как совместное распределение вероятностей $P(x,t)$ неизвестно, поиск ведется с использованием обучающей последовательности

$$(x^1,t_1), (x^2,t_2), ..., (x^N,t_N),$$

то есть случайной и независимой выборки примеров фиксированной длины N. Нельзя найти алгоритм, который по конечной выборке безусловно гарантировал успех поиска. Успех можно гарантировать лишь с некоторой вероятностью $1 - \eta$.

Таким образом, задача заключается в том, чтобы для любой функции

$P(x,t)$ среди характеристических функций $F(x,\alpha)$ найти по обучающей последовательности фиксированной длины N такую функцию $F(x,\alpha^*)$, о которой с надежностью, не меньшей $1-\eta$, можно было бы утверждать, что ее качество отличается от качества лучшей функции $F(x,\alpha^0)$ на величину, не превышающую ϵ.

Второй раздел «Классификация в R методом логистической регрессии» описывает теоретические и практические основы построения метода логистической регрессии.

Логистическая регрессия относиться к семейству обобщенных линейных моделей (GLM), и является расширением хорошо известной модели линейной регрессии. Другими названиями модели являются двоичная логистическая регрессия, биномиальная логистическая регрессия и логит-модель.

Заметим, что логистическая регрессия не возвращает непосредственно класс наблюдений, а оценивает вероятность принадлежности к классам, которая находится в диапазоне от 0 до 1. Исследователю необходимо определить пороговую вероятность, которая отделяет один класс от другого. По умолчанию это пороговое значение равно $p = 0.5$, но для большинства практических задач эту величину нужно выбирать на основе цели анализа.

Функция логистической регрессии, которая используется для прогнозирования класса наблюдения с учетом предикторной переменной x, определяемую как

$$p = \frac{e^y}{1+e^y}$$

и представляет собой s-образную кривую.

Проведение классификации на медицинских данных. Для классификации используется набор медицинских данных.Он взят из Национального института диабета, болезней пищеварения и почек. Цель набора данных - предсказать, есть ли у пациента диабет, основываясь на определенных диагностических измерениях, включенных в набор данных. На выбор этих экземпляров из более крупной базы данных было наложено несколько ограничений. В частности, все пациенты здесь - женщины в возрасте не менее 21 года.

Данные проверяются на наличие корреляции между переменными, так

как коррелирующие данные могут снизить точность модели.

Затем набор делится на 2 выборки: обучающую и тестовую.

Для вычисления логистической регрессии используется функция glm(). Для начала укажем опцию family = binomial, которая сообщает R, что модель является биномиальной и является логистической регрессией.

Этап оценки точности предсказания модели и ошибки прогнозирования на новом наборе тестовых данных является крайне важным. Фактический результат каждого примера из набора тестовых данных известен, поэтому оценку эффективности предсказанной силы модели можно проводить на основе сравнения предсказанных моделью значений с известными значениями.

Для этого находятся предсказанные моделью вероятности принадлежности классам примеров тестовой выборки на основе значений предикаторных переменных для этих примеров. Далее нужно рассчитать, что пример принадлежит классу 1, если значение вероятности выше некоторого порогового значения (по умолчанию 0,5). Функция предсказания R predict() может быть использована для прогнозирования вероятности наличия диабета на основе значений предикатора.

Далее сравниваются фактические результаты с прогнозируемыми. Для этого используется матрица ошибок (Confusion Matrix). Confusion Matrix представляет собой таблицу, которая описывает эффективность классификации для каждой модели на основе тестовых данных.

Коэффициент точности классификации (Accuracy) определяет как доля примеров, которые были правильно классифицированы, среди общего количества примеров.

Далее строится кривая ROC и вычисляется AUC (площадь под кривой), которые являются типичными показателями производительности для двоичного классификатора. Значения AUC лежат между 0 и 1. Чем выше значение, тем выше точность.

Проведение классификации на кредитных данных. В качестве кредитных данных взят набор по кредитным рискам, в котором содержится информация о заемщиках.

Для дальшешего проведения классификации данные делятся на обучающую и тестовую выборки.

Далее проводится оценка коэффициентов логистической регрессии. Построение модели производится несколько раз, чтобы удалить из модели незначимые переменные.

Как говорилось выше, чтобы оценить работоспособность модели, необходимо проверить, на сколько она точна на тестовых данных. Для этого находятся предсказанные моделью вероятности пренадлежности классам примеров тестовой выборки на основе значений предикаторных переменных для этих примеров. Далее нужно рассчитать, что пример принадлежит классу 1, если значение вероятности выше некоторого порогового значения.

Затем сравниваются фактические результаты с прогнозируемыми.

Затем строится кривая ROC и вычисляется AUC.

Третий раздел «Классификация в R методом дерева решений» описывает теоретические и практические основы построения метода дерева решений.

Дерево классификации используется для предсказания отклика y, который является категориальной переменной (меткой класса). В бакалаврской работе рассматривается процедура построения дерева решений для классификации:

1. Среди предикторов $X_1, X_2, \ldots, X_k$ для правила разбиения выбирается X_p, значения которого позволяют разделить наблюдения по отклику y наилучшим образом (чтобы каждое из двух подмножеств было однородным по значению y внутри подмножества и различным по отношению друг к другу). Для категориальной переменной X_p в качестве условия разделения выбирается равенство одному из возможных значений X_p в выборке. Для количественной переменной условием является неравенство $X_p \leq l$, где l – некоторое значение на интервале от минимального значения X_p до максимального значения в выборке наблюдений.

2. Наблюдения в левом узле такие, для которых X_p условие – истина (true), а в правом – все остальные, т.е. наблюдения, для которых условие разбиения ложно (false).

3. Для каждого, из получившихся при разбиении узлов, вычисляются процентные доли значений зависимой переменной и выносится решение: к какой категории будет принадлежать попавшее в нее наблюдение.

4. Если в узле достигнута однородность (т.е. все наблюдения, оказавшиеся в узле, из одного класса), то для этого узла процедура закончена.

5. Если в узле оказалось множество наблюдений из разных классов, то процесс бинарного разбиения для этого узла может быть продолжен с пункта 1 (в дальнейшем разбиении для текущего узла учитываются только наблюдения, попавшие в этот узел).

В качестве статистического критерия, оценивающего качество разбиения, используется индекс Джини.

Индекс Джини (Gini impurity):

$$I_G = 1 - \sum_{i=1}^{m} p_i^2,$$

где p_i^2 - частоты представителей разных классов в узле дерева; m - число классов для отклика y.

Деревья решений (если не ограничивать глубину и не вводить иные ограничения) могут давать излишне детализированную картину, когда в листьях сосредоточено мало наблюдений.

Такая ситуация называется переобучением. Дерево решений строится с целью определять по значениям предикатов прогноз отклика для нового наблюдения. Переобученное дерево дает минимальную ошибку на обучающей (тренировочной) выборке и очень часто ошибается на проверочной (тестовой). Соответственно, прогноз на переобученном дереве проигрывает прогнозу на основе дерева, в котором решена проблема переобучения тем или иным способом: за счет ограничений на глубину дерева, на количество наблюдений в узле, для которого возможно расщепление и т.п., либо строится максимально возможное дерево, а затем на основании определенных критериев отсекаются лишние ветки (убирается несущественная детализация).

Проведение классификации на медицинских данных. Классификация с помощью алгоритма дерева решений проводится на медицинских данных из второго раздела.

Данные разделены на тестовую и обучающую выборки.

Дерево решений построено с помощью алгоритма rpart. Функция rpart.plot использована для построения окончательного дерева решений.

Для данного дерева решений используется сокращение, чтобы не возникло переобучение.

Теперь, чтобы протестировать модель дерева решений, будет применён набор тестовых данных к данной модели. В программе для оценки модели type=class указывает на то, что решается задача классификации. Для сравнения фактических результатов с прогнозируемыми используется матрица ошибок. Как говорилось выше, матрица ошибок показывает сколько объектов класса i были распознаны как объект класса j. Диагональные элементы матрицы ошибок указывают долю правильных предсказаний модели, а вне диагонали расположены доли неправильных предсказаний. Коэффициент точности классификации (Accuracy) определяет как доля примеров, которые были правильно классифицированы, среди общего количества примеров.

Затем строится ROC-кривая. У высокоэффективного классификатора будет ROC-кривая, которая круто поднимается в верхний левый угол, то есть он будет правильно идентифицировать множество примеров одного класса без ошибочной классификации на множестве примеров другого класса как примеров из первого класса.

Область под кривой суммирует общую эффективность классификатора по всем возможным значениям вероятности, и представляет собой способность алгоритма классификации различать примеры одного класса от примеров другого.

Проведение классификации на кредитных данных.

Классификация с помощью алгоритма дерева решений проводится на кредитных данных из второго раздела.

Для проведения классификации данные делятся на тестовую и обучающую выборки.

Построение и визуализация была проведена аналогично, как и для медицинских данных.

В данном дереве также было использовано сокращение, т.е. для улучшения модели отсекаются лишние ветви.

Чтобы протестировать модель дерева решений, будет применён набор тестовых данных к данной модели. Оценивание модели проводится таким же способом, как в предыдущей модели.

Затем строится ROC-кривая.

Основные результаты

1. Рассмотрены теоретические основы задачи классификации, проведение классификации на языке программирования R, практическое применение данного класса задач и постановка задачи.

2. Изучены основы построения метода логистической регрессии, разработан программный код на языке R, позволяющий провести классификацию на медицинских и экономических данных.

3. Описаны результаты построенных моделей логистической регрессии.

4. Рассмотрены основы построения метода дерева решений, был построен программный код на языке программирования R, позволяющий легко воспроизвести все расчёты.

5. Описаны результаты построенных моделей деревьев решений для медицинских и кредитных данных.

Программный код приводится в **приложении А**.