Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Механико-математический факультет

УТВЕРЖДАЮ:
Декан ММФ ТГУ
Л. В. Гензе

Оценочные материалы по дисциплине

Методы машинного обучения с использованием Python

по направлению подготовки

01.04.01 Математика

Направленность (профиль) подготовки :
Математический анализ и моделирование (Mathematical Analysis and Modelling)

Форма обучения
Очная

Квалификация
Магистр

Год приема
2023

СОГЛАСОВАНО:
Руководитель ОП
А.В.Старченко

Председатель УМК
Е.А.Тарасов

Томск – 2023

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен формулировать и решать актуальные и значимые проблемы математики.

ПК-1 Способен самостоятельно решать исследовательские задачи в рамках реализации научного (научно-технического, инновационного) проекта.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК 1.1 Формулирует поставленную задачу, пользуется языком предметной области, обоснованно выбирает метод решения задачи.

ИПК 1.1 Проводит исследования, направленные на решение отдельных исследовательских задач

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:
– индивидуальные задания;

Индивидуальное задание (ИОПК-1.1, ИПК 1.1)

В ходе разведки месторождений нефти специалисты производят пробные бурения скважин и осуществляют анализ получаемых в ходе этого технических, геологических и геофизических данных. Целью этого является обнаружение нефтенасыщенных пластов, то есть пластов, содержащих в себе нефть и способных ее отдавать.

Перед вами стоит задача разработать алгоритм интеллектуального анализа реальных данных, позволяющий наиболее качественно определять наличие или отсутствие нефтяных пластов на тех или иных глубинах залегания скважин.

Метрикой качества выступает точность нахождения нефтенасыщенного пласта

$$Accuracy = \frac{samples\_true}{sampels\_all},$$

Где $samples\_true$ - количество правильных предсказаний наличия/отсутствия нефтяного пласта,

$sampels\_all$ - общее количество записей в таблице

Формат ввода

data_train.csv — файл с обучающими табличными данными

X_data_predict.csv — файл с данными, для которых необходимо предсказать целевую переменную

Файл с тренировочными табличными данными содержит информацию по 600 скважинам, для каждой из которых имеется различная техническая, геологическая и геофизическая информация в виде следующих полей:

- **MD** — относительная глубина скважины (относительно поверхности бурения), всегда является положительной величиной, используется для привязки глубин внутри скважины, но не может выступать в роли какого-то признака при прогнозе (по крайней мере с физической точки зрения).
- **TVDSS** — глубина скважины относительно уровня моря, всегда является положительной величиной, может отражать поверхность геологического пласта или уровень водонефтяного контакта.

- **Layer** — название пласта, геологическая принадлежность интервала, качественная характеристика, выдаваемая геологом на основе его понимания геометрических характеристик целевого пласта, служащая для сопоставления пластов из различных скважин между собой.
- **GK** — гамма-каротаж, измеряет естественную радиоактивность пород, различные минералы имеют разное содержание радиоактивных материалов, как правило, чем выше — тем больше глинистая составляющая и меньше песчанистая, может измеряться в единицах API или мкр/ч.
- **NNKT_big** — нейтронный каротаж, регистрирует относительное водородосодержание, что может говорить о количестве пор в горных породах (они не могут быть пустыми и всегда содержат какой-то флюид, который в значительном объеме содержит в себе водород). Меньшие значения отвечают за более высокое флюидосодержание.
- **PS** — каротаж естественной поляризации, последняя возникает при фильтрации флюида через породу, уменьшение значений говорит о наличии проницаемого интервала. Единица измерения — милливольты, может иметь совершенно разный масштаб в разных скважинах.
- **IK** — индукционный каротаж, отражает электрическую проводимость горных пород, величину, обратную сопротивлению. Поскольку нефть является диэлектриком, а вода проводником, высокие показания отражают водонасыщенные пласты, а низкие — интервалы, вмещающие нефть. С другой стороны, плотные породы, не содержащие в себе пор, также имеют высокое сопротивление, поскольку не имеют в себе флюида, который способен проводить ток.
- **BK** — боковой зонд, отражает сопротивление горной породы, интерпретируется схожим образом с кривой индукционного каротажа, но уже наоборот, повышенные значения — нефть или плотные породы, пониженные — вода или глина.
- **PZ** — потенциал-зонд, отражает сопротивление горной породы, интерпретируется схожим образом с кривой индукционного каротажа, но уже наоборот, повышенные значения — нефть или плотные породы, пониженные — вода или глина. Схож с боковым зондом (BK), но имеет другую глубинность исследования.
- **Grad_zond** — другая группа зондов, отвечающих за сопротивление горных пород, в зависимости от числа в названии определяется глубинность метода. При бурении буровой раствор попадает в пласт и может изменить содержание того или иного флюида, поэтому, в теории, пониженые сопротивления в затронутой части пласта и повышенные в глубинной могут быть признаком наличия углеводородов.
- **target_collector** — бинарная характеристика, выдаваемая специалистом по интерпретации каротажных данных, отвечающая за то, является ли тот или иной интервал коллекторским пластом, то есть пластом, способным принимать и отдавать флюид.
- **target_oil** — бинарная характеристика, выдаваемая специалистом по интерпретации каротажных данных, отвечающая за то, является ли тот или иной интервал коллекторским нефтенасыщенным пластом.
- **Well** — номер скважины.

В качестве целевой переменной выступает **target_oil**, которая при значении 1 говорит о наличии нефтенасыщенного пласта, а при значении 0 — о его отсутствии.

Формат вывода

В файл submission.csv необходимо записать одну колонку, в которой для каждой скважины из тестовой выборки стоит классифицирующая ее метка.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Примерный перечень теоретических вопросов

1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
2. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.
3. Метрики качества алгоритмов регрессии и классификации.
4. Линейная регрессия. Регрессия с полиномиальными признаками. Методы регуляризации: Ridge, Lasso, ElasticNet.
5. Деревья решений. Методы построения деревьев. Их регуляризация. Случайный лес, его особенности.
6. Дилллема: смещение/разброс. Переобучение, недобоучение.
7. Структура нейрона. Нейронные сети прямого распространения. Прямое распространение ошибки. Обратное распространение ошибки
8. Минимизация ошибки в НС. Почему при оптимизации используются частные производные. Как борются с переобучением

При оценке выполнения индивидуальных заданий учитывается правильность, оригинальность и сроки выполнения.

Ближе к концу семестра каждый студент получает итоговое индивидуальное задание. Оно включает в себя некий результирующий итог по освоению материала курса. Работа оформляется в виде отчёта, который студенту необходимо защитить: рассказать о ходе выполнения работы и ответить на дополнительные вопросы по теории.

По результатам защиты итогового индивидуального задания и по результатам сданной практики определяется оценка.

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Тест (ИОПК-1.1, ИПК 1.1)

Что такое задача классификации в машинном обучении?
a) Прогнозирование непрерывного значения
b) Прогнозирование категориальной метки или класса
c) Анализ временных рядов

Какой алгоритм машинного обучения используется для задачи регрессии?
a) Логистическая регрессия
b) Метод опорных векторов (SVM)
c) Линейная регрессия

Что такое обучение с учителем и обучение без учителя?
a) Обучение с учителем - это задача классификации, а обучение без учителя - задача регрессии
b) Обучение с учителем - это задача регрессии, а обучение без учителя - задача кластеризации
c) Обучение с учителем - это задача кластеризации, а обучение без учителя - задача классификации

Что такое задача регрессии в машинном обучении?
a) Прогнозирование непрерывного значения
b) Прогнозирование категориальной метки или класса
c) Анализ временных рядов

Какая метрика обычно используется для оценки качества модели регрессии?

a) Accuracy

b) Precision

c) Mean Squared Error (MSE)

Какая метрика может быть использована для оценки качества модели регрессии, если данные имеют выбросы?

a) Mean Absolute Error (MAE)

b) Root Mean Squared Error (RMSE)

c) R^2

Что такое переобучение (overfitting) в контексте моделей регрессии?

a) Модель слишком простая и не способна улавливать сложные зависимости в данных

b) Модель слишком сложная и "запоминает" тренировочные данные, не обобщаясь на новые данные

c) Модель не может быть обучена из-за недостаточного количества данных

Теоретические вопросы (ИОПК-1.1, ИПК 1.1):

1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
2. Метрики качества алгоритмов регрессии и классификации.
3. Что такое ансамбль моделей и какие принципы используются для их построения.
4. Структура нейронных сетей прямого распространения.
5. Роль функции активации в нейронной сети.

Информация о разработчиках

Стребкова Екатерина Александровна, ст. преподаватель кафедры вычислительной математики и компьютерного моделирования ММФ ТГУ;