Приложение к рабочей программе дисциплины

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ОРГАНИЗАЦИИ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

Наименование дисциплины	*Машинное обучение*
Направление подготовки / Специальность	*38.03.01 Экономика*
Направленность (профиль) / Специализация	*Экономика и анализ данных* ОП *ВО*
Форма обучения	*очная*

Разработчик Гайдамак И.В., старший преподаватель кафедры фундаментальной математики и механики

1. Темы дисциплины для самостоятельного освоения обучающимися

Отсутствуют.

2. План самостоятельной работы:

№ п/п	Учебные встречи	Виды самостоятельной работы	Форма отчетности / контроля	Количество баллов	Рекомендуемый бюджет времени на выполнение (ак.ч.)
1.	Введение в машинное обучение и анализ данных	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
2.	Линейные модели для задачи регрессии	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
3.	Линейные модели для задачи классификации	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
4.	Решающие деревья	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
5.	Композиции моделей	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
6	Обучение без учителя	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
7	Рекомендательные системы	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	10
8	Зачетная работа	Проектная работа	Защита проектной работы	50	62
		Подготовка к зачетной работе	Контрольная работа	-	10
	Итого			50	142

3. Требования и рекомендации по выполнению самостоятельных работ обучающихся, критерии оценивания

Вид: изучение темы на онлайн-платформе ВШЭ

Краткая характеристика: Студенты проходят курс «Машинное обучение» на онлайн платформе ВШЭ, изучая последовательно темы:

- Основные понятия и задачи в машинном обучении
- Метод k ближайших соседей
- Линейная регрессия
- Обучение моделей градиентными методами
- Линейная классификация: общие принципы
- Линейная классификация: методы
- Решающие деревья

- Композиции: бэггинг, блендинг и стэкинг
- Градиентный бустинг
- Обучение без учителя
- Рекомендательные системы

В каждой теме студенты проходят неоцениваемое тестирование (задания на понимание), а также выполняют практические задания, результаты которых также проверяются тестированием.

Рекомендации по выполнению: просматриваются онлайн-лекции, делается конспект. Далее студент проходит тестирования по теоретической части. Затем студент выполняет практическое задание.

Вид: проектная работа

Краткая характеристика: в работе над проектом с дальнейшей его защитой студент показывает текущий уровень знаний и компетенций. На основе выбранного датасета студент проходит весь путь – от сбора и предобработки данных и получения обученной модели.

Рекомендации по выполнению:

Последовательность работы студента над проектом:

1. Выбрать датасет для проекта
2. Сформулировать цель и задачи исследования
3. Провести анализ и предобработку данных
4. Проверить статистические гипотезы
5. Обучить модели
6. Оценить качество моделей
7. Сделать выводы
8. Подготовить презентацию
9. Защитить свой проект

Отчетные документы по проекту:

1. Файл(ы) с исходными данными
2. Notebook с кодом
3. Презентация в двух форматах: PowerPoint и pdf

Названия всех файлов должны начинаться с фамилии студента

Критерии оценивания работы над проектом

	Оцениваемый параметр работы студента	Максимальный балл
1	Формулировка цели и задач проекта	5
2	Предобработка данных	5
3	Проверка гипотез	5
4	Обучение и сопоставление моделей	10
5	Выводы по работе	5

6	Презентация	5
7	Доклад	5
8	Ответы на вопросы	5
9	Вопросы к другим докладчикам	5
	Итого	50

Вид: подготовка к зачетной работе

Краткая характеристика: задания направлены на проверку знаний, умений и навыков анализа данных и работы с моделями на основе машинного обучения.

Рекомендации по выполнению:

Для подготовки к зачетной работе рекомендуется:

- повторить лекционный материал по пройденным темам;
- решать задачи по пройденным темам;
- провести самопроверку, решив демонстрационный вариант.

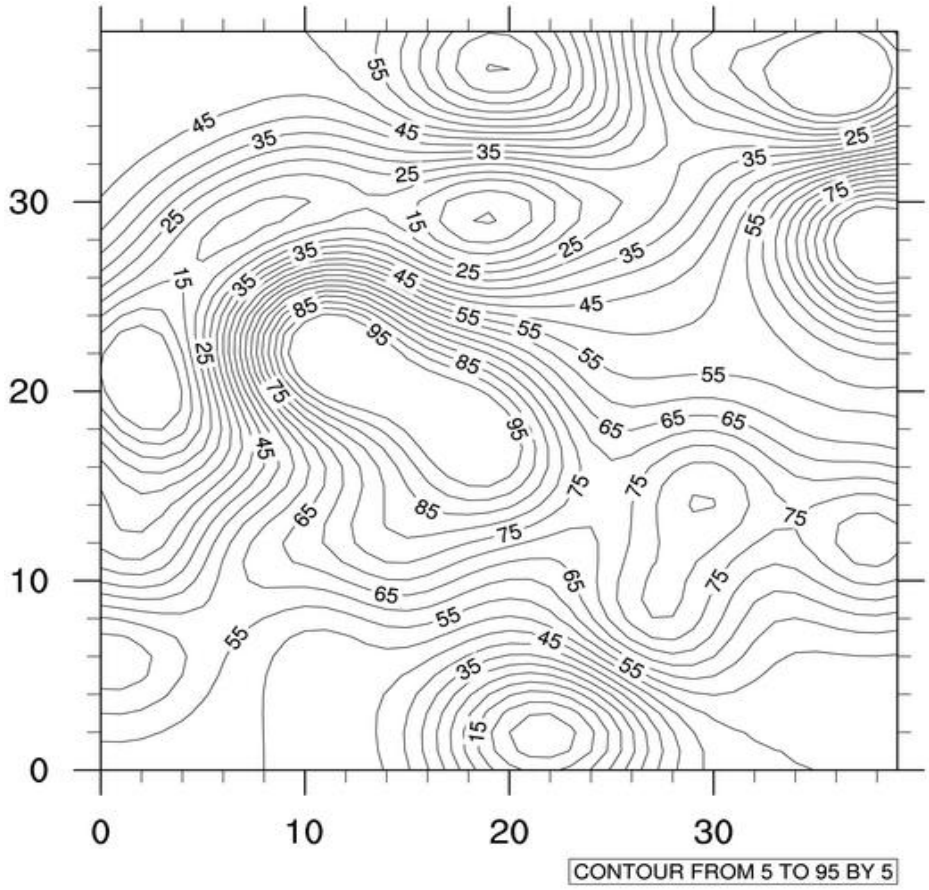
Демонстрационный вариант зачетной работы

А-1. Визуализация.

Иногда приходится иметь дело с трехмерными поверхностями, т.е. результатом табулирования функции от двух переменных $z = f(x, y)$. Сейчас, как правило, с восприятием таких графиков нет никаких проблем - есть много способов создать интерактивный трехмерный график, который можно вертеть и масштабировать как угодно. Но если мы говорим о статичной картинке, то с пониманием графика возникают некоторые трудности. Выход нашли еще до появления компьютеров, при издании карт местности, который заключается в том что строят не трехмерный график а его отображение на плоскости. Это отображение получают по следующим правилам: берут заданное количество плоскостей параллельных плоскости *xy* и выделяют место пересечения поверхности с каждой из этих плоскостей - контуры, затем проецируют данные контуры на плоскость *xy* и все график готов.

Каждому контуру соответствует определенное значение величины *z*, его подписывают рядом с контуром.

Перед нами стоит задача минимизировать некоторую функцию потерь, contour plot которой изображен ниже.

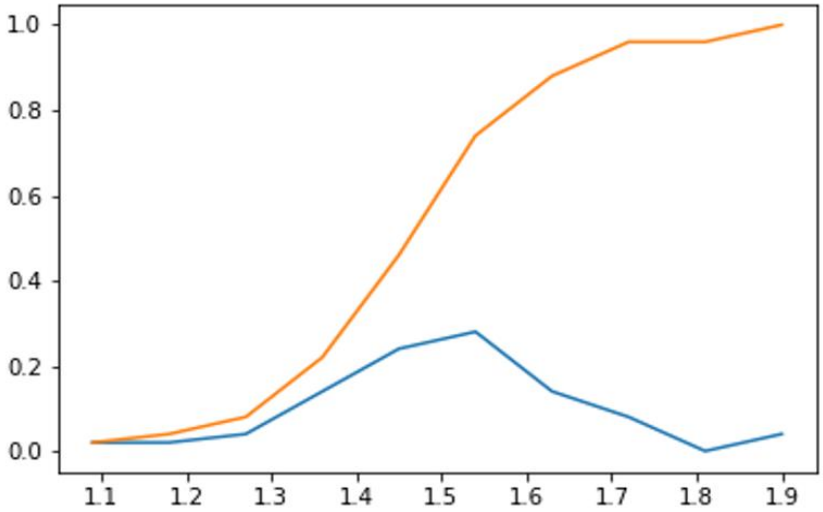


Выберите все верные утверждения:

1) Функция имеет несколько локальных экстремумов
2) На картинке есть как области локальных минимумов, так и области локальных максимумов
3) По графику можно точно определить минимальное значение функции, оно равно 15
4) В точке (x,y)=(16,20) значение функции близко к своему минимальному значению.

A-2 Анализ графика плотности / функции распределения.

На картинке нарисованы графики плотности распределения (pdf) и функции распределения (cdf) некоторой случайной величины.



Выберите все верные варианты ответа:

1) Оранжевый график - это плотность, синий - функция распределения
2) Оранжевый график - это функция распределения, синий - плотность
3) Медиана случайной величины равно примерно 1.55
4) Мода случайной величины равна примерно 1.8

А-3 тервер (мат.ожидание, дисперсия)

Аналитик Вася с утра просыпается с головной болью и попадает по нужной клавише на клавиатуре с вероятностью 0.7. Найдите математическое ожидание количества правильно нажатых клавиш в предложении из 40 символов.

А-4 основные понятия проверки гипотез + коэффициент корреляции

При решении некоторой задачи аналитики выяснили, что корреляция Пирсона между средним уровнем дохода клиентов некоторой компании и частотой невозврата кредитов равна -0.4.

Какие выводы можно сделать из этого утверждения? Выберите все НЕВЕРНЫЕ варианты ответа.

1) Чем больше средний уровень дохода, тем меньше частота невозврата кредитов.
2) Нельзя сказать, что существует какая-либо (линейная или нелинейная) зависимость между средним уровнем дохода и частотой невозврата кредитов.
3) Чем больше средний доход, тем реже возвращают кредит.
4) Коэффициент корреляции Пирсона никак не связан с линейной взаимосвязью величин.

А-5 проверка гипотез

Инженер данных Сергей построил пайплайн для сбора и хранения данных. Сергей утверждает, что в его алгоритме данные теряются с вероятностью 2%. Затем пришел инженер данных Виталий и, проверяя гипотезу Сергея, заявил, что такая маленькая потеря данных невозможна.Сергей обиделся и протестировал свой алгоритм многократными проверками, при этом каждый раз только 2% данных терялось.

Выберите верное утверждение:

1) Виталий совершил ошибку первого рода.
2) Виталий совершил ошибку второго рода.
3) Виталий не совершил ошибку, ошибку совершил Сергей - но какого рода, неизвестно.

А-6 линал для МО

В линейной алгебре и машинном обучении существует алгоритм под названием “метод главных компонент”, с помощью него можно снижать размерность пространства. В нем для снижения размерности нужно искать собственные значения и собственные векторы матрицы $X^T X$, где X - матрица объект-признак.

Вам дана матрица $A = X^T X$. Найдите её наибольшее собственное значение.

$$A = \begin{pmatrix} -1 & -6 \\ 2 & 6 \end{pmatrix}$$

А-7 основные задачи МО

Предположим, вы проводите исследование благосостояния граждан по странам. Имеющиеся у вас данные включают в себя уровень дохода, среднегодовую температуру в месте проживания респондента, степень удовлетворенности жизнью, ВВП на душу населения, возраст респондента, его рост и другие характеристики для 100 тысяч граждан 100 государств. Хотим построить модель, которая будет по данным для человека предсказывать уровень благосостояния - любое число на отрезке от 1 до 100 (где 1 - низкий уровень, а 100 - высокий уровень). Ответ может быть любым числом из отрезка [1; 100], в том числе и нецелым.

Выберите все верные утверждения

1) Имеем дело с задачей классификации
2) Будем работать над решением задачи регрессии
3) Масштабирование (нормализация) признаков при использовании в этой задаче линейной регрессии без регуляризации может изменить среднеквадратичную ошибку в сравнении с обучением модели без масштабирования
4) Объектом в этой задаче является страна
5) Объектом является гражданин страны

A-8 переобучение + регуляризация

Цель банка - найти мошенников при совершении банковских операций. Пусть при обучении классификации для выявления мошеннических транзакций в банке accuracy на тренировочных данных получилось равным 0.9, а на тестовых - 0.85. О чем может говорить эта ситуация?

Для решения задачи использовали логистическую регрессию с Lasso-регуляризацией.

Выберите наиболее подходящий вариант ответа.

1) Модель переобучилась, так как качество на тесте ниже, чем на трейне.
2) Модель недообучилась, так как качество на трейне низкое
3) Неизвестно, переобучилась или недообучилась модель, так как выбрана плохая метрика для данной задачи.
4) Модель обучилась как надо (нет ни недообучения, ни переобучения).

A-9 работа с текстами

Перед data scientistom Алексеем была поставлена задача классифицировать отзывы на отели на положительные и отрицательные. Для преобразования отзывов в числовые признаки Алексей выполнил посимвольную токенизацию и заменил каждый символ на его ASCII-код (пример посимвольной токенизации: “бегать” -> [”б”, ”е”, ”г”, ”а”, ”т”, ”ь”]). Больше никакой обработки данных Алексей не делал.

Что произойдет при обучении модели классификации на обработанных таким образом данных? Выберите наиболее подходящий вариант ответа.

1) Если подобрать подходящую модель и настроить её гиперпараметры должным образом, то мы получим модель, дающую хорошие предсказания на новых данных

2) Даже при подборе хорошей модели и настройке её гиперпараметров мы не получим хорошую предсказательную модель, качество предсказания будет низким
3) Модель не сможет обучиться при такой обработке данных и выдаст ошибку.
4) Линейная модель выдаст плохое качество предсказания на новых данных, а нелинейная гарантированно сработает лучше.

A-10 выбор метрики

Мы решаем задачу классификации - определения по МРТ-снимкам, болен пациент некоторой болезнью или нет. Мы не хотим часто ошибаться, то есть называть больных здоровыми, но и здоровых называть больными мы также не хотим. В обучающих данных 8543 пациента, из которых 4100 больных, а остальные здоровые.

Какие метрики можно использовать для измерения качества модели в этой задаче?

1) Accuracy
2) MAPE
3) AUC-PR (площадь под Precision-Recall кривой)
4) MSLE (mean squared logarithmic error)
5) F1-score

A-11 интерпретация метрики

Для оценки качества модели классификации использовали f-beta score:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision + recall)}.$$

Какое значение β необходимо выбрать, если для нас более важно минимизировать ошибку типа FP (false positive) - неверно назвать объекты отрицательного класса положительными, чем ошибку типа FN (false negative) - неверно назвать объекты положительного класса отрицательными?

1) $\beta = 0$
2) $\beta < 1$
3) $\beta = 1$
4) $\beta > 1$

A-12 линейный алгоритм + градиентный спуск

Будем решать задачу классификации линейно разделимой выборки при помощи метода опорных векторов с линейным ядром.

В общем случае функция потерь метода имеет вид

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{l} \xi_i \to min_{w,\xi_i},$$

где первое слагаемое регулирует ширину разделяющей полосы

($\frac{2}{||w||}$ – ширина разделяющей полосы), второе регулирует сумму штрафов (ξ_i- штраф на i-м объекте за попадание внутрь разделяющей полосы и/или не в свой класс).

C - гиперпараметр, задаваемый человеком при объявлении модели.

Какую ошибку классификации мы получим на тестовых данных при решении задачи этим методом?

1) 0, так как выборка линейно разделима
2) 0 при достаточно большом значении C, а при малых C неизвестно
3) 0 при достаточно маленьком значении C, а при больших C неизвестно
4) нельзя гарантировать нулевую ошибку ни при каком C

A-13 композиции

Существует два популярных алгоритма построения решающих деревьев в задаче классификации - ID3 (минимизируем энтропию) и CART (минимизируем критерий Джини). Выберите верную формулу для критерия информативности в алгоритме CART:

1) $H(R) = -\sum_{k=1}^{K} p_k log(p_k), p_k$ - доля объектов k-го класса в вершине дерева R, K - число классов в задаче.
2) $H(R) = \sum_{k=1}^{K} p_k(1 - p_k), p_k$ - доля объектов k-го класса в вершине дерева R, K - число классов в задаче.
3) $H(R) = -\sum_{i=1}^{n} p_i log(p_i), p_i$ - вероятность принадлежности i-го объекта к положительному классу, n - число объектов в вершине R.
4) $H(R) = 2 \cdot AUC - 1$, где AUC - значение метрики ROC-AUC в вершине R.

A-14 композиции, bias-variance decomposition

Напомним, что существует формула для разложения ошибки модели на шум, смещение и разброс:

$$Err = Bias^2 + Var + \sigma^2,$$

$Bias^2$ - смещение модели, Var - разброс, σ^2 - шум в данных.

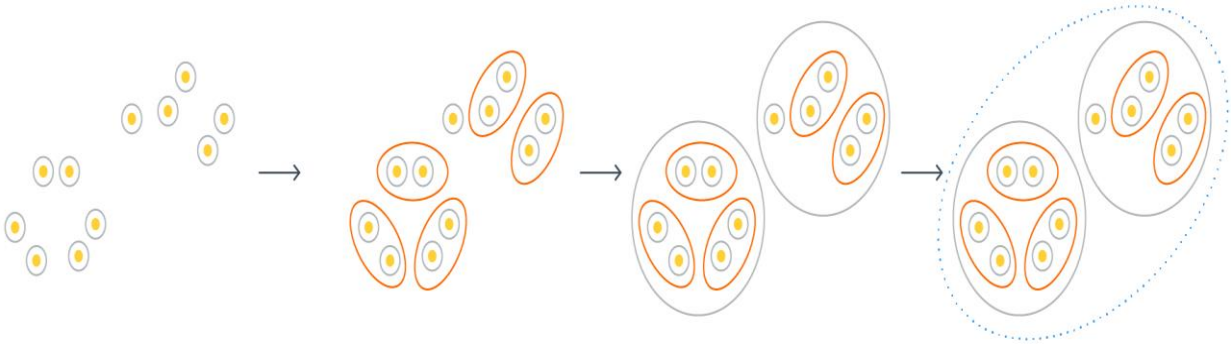
На некоторых данных обучили линейную регрессию. Как изменятся компоненты $Bias^2$ и Var ошибки линейной регрессии, если добавить в данные для обучения модели полиномиальные признаки степени 2?

1) $Bias^2$ увеличится

2) $Bias^2$ уменьшится
3) $Bias^2$ не изменится
4) Var увеличится
5) Var уменьшится
6) Var не изменится

А-15 кластеризация

На рисунке ниже изображен некоторый алгоритм кластеризации. А какой?



1) K-means
2) DBSCAN
3) Аггломеративная (иерархическая) кластеризация
4) Спектральная кластеризация

А-16 нейронные сети

На скрытом слое нейронной сети используется функция активации a. Выходное значение некоторого нейрона после применения функции активации получилось равным 1.05. Какая из перечисленных функций активации a могла быть использована в этой сети?

1) ReLU
2) Tanh
3) Sigmoid
4) Никакая из перечисленных

Задачи (часть В)

В-1 метрики

b_i	y_i
0.7	+1
0.6	-1
0.3	-1

0.45	+1
0.92	-1

Алгоритм бинарной классификации для каждого объекта x_i выдает оценку b_i его принадлежности к положительному классу. Ниже в таблице даны предсказания b_i и правильные ответы y_i.

Студент, которому необходимо вычислить PR-AUC алгоритма, решил, что вероятности ему не нужны, поэтому он перевел предсказания b_i в классы (если $b_i \geq 0.5$, то предсказанный класс +1, а иначе -1). И по полученным предсказанным классам посчитал PR-AUC. На сколько полученное студентом значение PR-AUC меньше значения, которое получилось бы, если бы студент использовал b_i вместо классов для вычисления PR-AUC? Ответ округлите до сотых.

PR-AUC (площадь под кривой precision-recall) вычисляется следующим образом. Предсказания модели упорядочиваются по убыванию вероятностей положительного класса. Затем для каждого объекта в упорядоченной выборке вычисляется precision и recall, и точка с этими координатами ставится на координатную плоскость с осями x (precision) и y (recall). Затем полученные точки последовательно соединяются, и вычисляется площадь под построенной кривой (PR-кривая).

B-2 алгоритм

Мы решаем задачу регрессии. Для этой задачи было решено использовать квантильную регрессию, при обучении которой минимизируется квантильная функция потерь с параметром $\tau = \frac{1}{2}$:

$$Q(w, X) = \frac{1}{l}\sum_{i=1}^{l} \quad \rho_\tau(y_i - a(w, x_i)), \text{где } \rho_\tau(z) = (\tau - \frac{1}{2})z + \frac{1}{2}|z|,$$

$a(w, x_i) = w_0 + w_1 x_i$ - предсказание модели на объекте x_i, l - количество объектов в обучающей выборке.

Данные для задачи:

x_i	y_i
5	9
2.3	5
7.7	-15

Что предскажет эта модель на объекте $x_i = -1$?

В-3 алгоритм

В вершине дерева, решающего задачу бинарной классификации, находилось 40 объектов класса 1 и 60 объектов класса 0. После разбиения вершины на две группы по некоторому условию:

- в левой вершине оказалось 20 объектов класса 1 и 50 объектов класса 0
- в правую вершину попали все остальные объекты.

Вычислите Information Gain:

$$Q = H(R) - \frac{|R_l|}{|R|}H(R_l) - \frac{|R_r|}{|R|}H(R_r),$$

Где |A| - количество объектов в вершине A, а

$H(R) = \sum_{k=1}^{2} \quad p_k(1 - p_k)$ - значение критерия Джини в вершине R.

В-4 линал

В машинном обучении есть подход, позволяющий при помощи линейных моделей решать линейно неразделимые задачи классификации: в этом подходе мы переходим в новое пространство признаков и в этом пространстве решаем задачу при помощи линейной модели. Скалярное произведение векторов в новом пространстве задается функцией, называющейся ядром.

Дано ядро $K(a, b) = exp(-||a - b||^2)$, где $||a - b|| -$
евклидова норма (длина) вектора $a - b$.

Вычислите косинус угла между векторами $a = (1,1,1)$ и $b = (1,2,0)$ в новом признаковом пространстве, в котором скалярное произведение задается функцией $K(a, b)$.

Ответ округлите до сотых.

В-5 тервер

За круглый стол на 201 стул в случайном порядке рассаживаются 199 разработчиков и 2 аналитика. Найдите вероятность того, что между аналитиками будет сидеть один разработчик.

В-6 нейронные сети

Для того, чтобы градиентный спуск при оптимизации сложных функций потерь в при обучении нейронных сетей не застревал в локальных минимумах, используют метод моментов. Одна из форм записи метода такая:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

где α - величина градиентного шага, β - константа, отвечающая за "память" модели.

α, β - гиперпараметры.

Укажите, в какую точку x приведёт нас метод моментов с $\alpha = 0.1, \beta = 0.5$ *после двух итераций* при поиске минимума функции

$f(x) = x^3 - 3x,$

если стартовая точка $x_0 = 3$ (считаем на первой итерации, что $x_0 - x_{-1} = 0$).

Ответ округлите до сотых.

Работа с данными (часть С)

В файлах TrainData.csv и TestData.csv находятся данные о клиентах некоторой компании.

В этом задании предлагается изучить анонимизированные характеристики клиентов и на их основе решить задачу оттока: понять, покинет клиент компанию или нет.

Описание столбцов:

* столбцы 0-13 - анонимизированная информация о клиентах

* столбец target - целевая переменная: 1 - клиент покинет компанию, 0 - не покинет

Далее за задания можно получить максимум 4 балла.

Считайте данные в два pandas dataframe: df_train и df_test.

Задание 1 (0.15 балла). Проверьте, есть ли в тренировочных и в тестовых данных пропуски? Укажите количество столбцов тренировочной выборки, имеющих пропуски.

Задание 2 (0.15 балла).

a) (0.05 балла). В столбце с наибольшим количеством пропусков заполните пропуски средним значением по столбцу. В ответ запишите значение вычисленного среднего. Ответ округлите до десятых.

b) (0.1 балла). Найдите строки в тренировочных данных, где пропуски стоят в столбце с наименьшим количеством пропусков. Удалите эти строки. Сколько строк вы удалили?

Задание 3 (0.25 баллов). Переведите столбец с целевой переменной в бинарные значения по правилу: Churn - 1, Not churn - 0. Исправьте опечатки в названиях категорий целевой переменной (до того, как переводить их в бинарные значения).

Сколько опечаток вы исправили?

Задание 4 (каждый пункт - 0.3 балла, 1.2 балла максимум).

В этом задании все пункты выполняйте только по таблице df_train.

a) Найдите числовой признак, имеющий наибольшую по модулю корреляцию Пирсона с колонкой target. В ответ напишите модуль корреляции Пирсона с целевой переменной для этого признака. Ответ округлите до сотых.

b) Сколько столбцов в таблице (не считая target) содержат меньше 5 различных значений?

c) Верно ли, что если значение единственного категориального признака в таблице равно A, то клиент уйдет из компании? (target = 1) В ответ запишите “да” или “нет”.

d) Вычислите долю ушедших из компании клиентов, для которых значение признака 2 больше среднего значения по столбцу, а значение признака 13 меньше медианы по столбцу. Ответ округлите до сотых.

Задание 5 (0.75 баллов).

a) (0.25 балла) Закодируйте категориальные столбцы в тренировочных и тестовых данных при помощи label encoding (категории кодируйте подряд идущими числами, начинающимися с 0).

Сколько столбцов после кодировки стало в тестовых данных?

b) (0.5 балла) Разбейте тренировочные данные на целевой вектор y, содержащий значения из столбца target, и матрицу объект-признак X, содержащую остальные признаки. Обучите

на этих данных логистическую регрессию из sklearn (LogisticRegression) с параметрами по умолчанию. Выведите среднее значение метрики f1-score алгоритма на кросс-валидации с тремя фолдами. Ответ округлите до сотых.

Комментарий: параметры по умолчанию предполагаются следующими

penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None

Задание 6 (1.5 балла максимум).

6. a) Подберите значение константы регуляризации C в логистической регрессии, перебирая гиперпараметр от 0.001 до 100 включительно, проходя по степеням 10. Для выбора C примените перебор по сетке по тренировочной выборке (GridSearchCV из библиотеки sklearn.model_selection) с тремя фолдами и метрикой качества - f1-score. Остальные параметры оставьте по умолчанию. В ответ запишите наилучшее среди искомых значение C. (0.25 балла)

Комментарий: параметры по умолчанию предполагаются следующими

penalty='l2', dual=False, tol=0.0001, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None

b) Добавьте в тренировочные и тестовые данные новый признак ‘NEW’, равный произведению признаков ‘7’ и ‘11’.

На тренировочных данных с новым признаком заново с помощью GridSearchCV (с тремя фолдами и метрикой качества - f1-score) подберите оптимальное значение C (перебирайте те же значения C, что и в предыдущих заданиях), в ответ напишите наилучшее качество алгоритма (по метрике f1-score), ответ округлите до сотых. (0.5 балла).

Комментарий: параметры по умолчанию предполагаются следующими

penalty='l2', dual=False, tol=0.0001, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None

c) Теперь вы можете использовать любую модель машинного обучения для решения задачи. Также можете делать любую другую обработку признаков. Ваша задача - получить наилучшее качество по метрике accuracy на тестовых данных. (0.75 балла)

Качество проверяется на представленных тестовых данных.

* accuracy >= 0.88 - 0.25 балла

* accuracy >= 0.9 - 0.75 балла

Пример файла для отправки результатов: result.txt

4. Рекомендации по самоподготовке к промежуточной аттестации по дисциплине

Вопросы для самопроверки к дифференцированному зачету

1. Что такое объекты и признаки в машинном обучении? Для чего нужен функционал качества? Что такое алгоритм (модель)?
2. Чем задача классификации отличается от задачи регрессии? Приведите примеры задач классификации и регрессии.
3. Что такое вещественные (числовые), бинарные, категориальные признаки? Приведите примеры.
4. В чём заключается обобщающая способность алгоритма машинного обучения? К чему приводит её отсутствие? Что такое переобучение?
5. Что такое отложенная выборка? Что такое кросс-валидация (скользящий контроль)? Как ими пользоваться для выбора гиперпараметров?
6. В чём заключается гипотеза компактности?
7. Как метод k ближайших соседей определяет класс для нового объекта?
8. Опишите метод k ближайших соседей с парзеновским окном. Какие в нём есть параметры?
9. Запишите формулу метода kNN для регрессии.
10. Что такое градиент? Какое его свойство активно используется в машинном обучении?
11. Опишите алгоритм градиентного спуска.
12. Как обучается линейная регрессия?
13. Почему наличие линейно зависимых признаков представляет проблему при обучении линейной регрессии?
14. Что такое регуляризация? Как она помогает бороться с переобучением?
15. Чем L1-регуляризация отличается от L2-регуляризации?
16. Что такое масштабирование (шкалирование) признаков? Как его проводить? Зачем это нужно?
17. Как выглядит модель линейной классификации в случае двух классов?
18. Что такое отступ? Для чего он нужен?
19. Как обучаются линейные классификаторы (общая схема с верхними оценками)?
20. Для чего может понадобиться оценивать вероятности классов?
21. Как обучается логистическая регрессия? Запишите функционал и объясните, откуда он берётся.
22. Как в логистической регрессии строится прогноз для нового объекта?
23. Как можно использовать категориальные признаки в линейных моделях?
24. В чём заключается использование квадратичных признаков в линейных моделях? Для чего это нужно?
25. Что такое доля правильных ответов? В чём заключаются её проблемы?

26. Что такое точность и полнота? Что такое F-мера?
27. В чём заключается разница между метриками Accuracy и Precision?
28. Что такое ROC-кривая? Что такое AUC-ROC? Для чего он используется?
29. Что такое PR-кривая? Что такое AUC-PRC? Для чего он используется?
30. Как можно свести задачу многоклассовой классификации к серии задач бинарной классификации?
31. Что такое гиперпараметр? Чем гиперпараметры отличаются от обычных параметров алгоритмов? Приведите примеры параметров и гиперпараметров в линейных моделях.
32. Что такое решающее дерево? Как оно строит прогноз для объекта? Как обучаются решающие деревья в задачах классификации и регрессии (критерии информативности)?
33. Какие вы знаете критерии останова и способы выбора значений в листьях? Какие гиперпараметры имеются у деревьев?
34. Что такое бэггинг и метод случайных подмножеств? Что такое случайный лес, как он обучается и как он строит прогнозы?
35. Опишите идею градиентного бустинга для среднеквадратичной ошибки. Как устроен градиентный бустинг для произвольной дифференцируемой функции потерь? Запишите задачу для обучения очередной базовой модели.
36. Опишите одномерные подходы к отбору признаков с помощью дисперсии, корреляции и t-score. Какие у них недостатки?
37. Как можно отбирать признаки с помощью линейных моделей, решающих деревьев, случайных лесов?
38. Как устроен метод главных компонент? Как он формирует новые признаки? Как ищется решение, как оно связано с собственными векторами?
39. Дайте определение задачи кластеризации. Чем она отличается от задач классификации и регрессии? Опишите графовые алгоритмы кластеризации, основанные на компонентах связности и минимальном остовном дереве.
40. Опишите алгоритм кластеризации K-Means. Как в нём можно выбирать количество кластеров?
41. В чём заключается идея алгоритма word2vec? Что такое контекст слова? Как в word2vec оценивается близость слов по смыслу?
42. Опишите подход user-based collaborative filtering к построению рекомендательной системы.
43. Опишите подход к построению рекомендательных систем, основанный на моделях со скрытыми переменными.
44. Чем задача ранжирования отличается от задач классификации и регрессии? Запишите формальную постановку задачи ранжирования. Как выглядит метрика DCG?
45. Опишите поточечный и попарный подходы к решению задачи ранжирования.