Экзамен

Экзамен будет проходить 15 июля. Экзамен сдается устно.
Времени на подготовку ответа не будет, вы должны отвечать сразу после получения вопроса.

К экзамену допускаются только те участники, кто в курсе решил **все проекты** на 8–10 баллов.

На экзамене три трека. Два обычных – машинное обучение и анализ данных, и один дополнительный усложненный – оптимизация. Можно выбирать любые треки.
Полученные баллы на всех выбранных треках будут суммироваться.

На экзамене вам будут задаваться вопросы из списка. На вопросы нужно ответить развернуто. Ответ на каждый вопрос должен содержать все перечисленные в вопросе определения. Если где-то можно написать формулу, то её стоит написать. Будьте готовы к дополнительным вопросам и уточнениям. В списке под каждым вопросом курсивом написаны **примеры** дополнительных вопросов. На них тоже стоит опираться при подготовке.

В скобках указана сложность (и одновременно стоимость) вопроса. За правильный ответ на вопрос вы получите его стоимость. Если вы не ответите на дополнительные вопросы, но правильно дадите определения и напишете формулы, вы можете получить половину стоимости. За неправильный ответ вы потеряете стоимость вопроса.

В отрицательной зоне можно оказаться лишь единожды, при этом опускаться ниже -5 нельзя. А так можно попытаться отвечать хоть на все вопросы.

По результатам экзамена будет составлен топ людей. Люди с верха этого топа получат призы.

Машинное обучение

1. Общая постановка задачи машинного обучения. Признаки (факторы) и типы признаков. (1)
   Что такое модель машинного обучения?
   Вероятностная интерпретация задачи машинного обучения.
   Гиперпамататры и параметры модели.

2. Обучение с учителем и без учителя. Задачи классификации, регрессии и кластеризации. (2)
   В чем разница между классификацией и кластеризацией?
   Как можно оценить качество кластеризации, если истинные метки кластеров неизвестны?

3. Оценка качества модели. Обучающие, валидационные и тестовые выборки. (1)
   В чем разница между значением функции потерь и метрикой?
   Почему нельзя обучать и тестировать модель на одних и тех же данных?
   Как можно выявить недообучение модели?

4. Кросс-валидация: отложенная, полная, k-fold и t×k-fold. (2)
   Перечислите достоинства и недостатки отложенной валидации.
   Зачем была придумана модификация t×k-fold? Какую статистическую проблему она решает?

5. Модель линейной регрессии. Вероятностная интерпретация. Множественная регрессия. (2)
   Как линейная регрессия связана с нормальным распределением?
   Когда линейная регрессия не применима?
   Что такое мультиколлинеарность и как её можно обнаружить?

6. Метрики MSE, MAE, Хьюбера. (2)
   Какое предположение о распределении ошибок минимизирует MSE?
   Какая функция ошибок позволит лучше бороться с выбросами?
   Нужно ли нормализовывать данные перед использованием всех этих метрик?

7. Бинарная классификация. Функция потерь и обучение методом градиентного спуска. (3)
   Верно ли, что в модели бинарной классификации $y = sign(w_1 \cdot x_1 + w_2 \cdot x_2)$ снизу прямой будут находиться только точки, принадлежащие отрицательному классу?
   Гиперпараметры learning rate, batch size и количество итераций. Что зачем нужно?
   Как обучать модель при сильном дисбалансе классов?

8. Оценка качества моделей классификации. Матрица ошибок. (2)
   Что такое accuracy, precision, recall и F-мера? Зачем нужна взвешенная F-мера?
   Применима ли Accuracy к задачам с неравными классами?
   Как сравнить две модели, если одна имеет higher precision, а другая – higher recall?

9. Логистическая регрессия для бинарной классификации. Функция потерь. (3)
   Получите логистическую функцию потерь из принципа максимума правдоподобия.
   Чем кросс-энтропия лучше accuracy?
   Как интерпретировать веса логистической регрессии?

10. Переобучение. Кривые обучения и показатели качества. Дилемма bias-variance. (1)
    Как выбросы связаны с переобучением?
    Страдает ли дерево решений от выбросов?
    Как можно визуально определить переобучение и недообучение по графику зависимости ошибки от сложности модели?

11. Регуляризация. Борьба с переобучением. (2)
L1 и L2 регуляризация. Когда что применять?
Как правильно подбирать коэффициент регуляризации?

12. Кластеризация. Метрики расстояний и сходства. (1)
Евклидово, манхэттенское и косинусное расстояние. Когда какое лучше использовать?
Как нормализация влияет на результаты кластеризации?
Какие проблемы возникают при кластеризации данных с высокой размерностью?

13. Алгоритм k-means и его модификации. (2)
Почему k-means чувствителен к выбросам? Как можно это исправить?
Почему k-means может сходиться к локальному минимуму? Как это избежать?
Как обработать категориальные признаки в k-means?

14. Иерархическая кластеризация. (2)
Как можно ускорить иерархическую кластеризацию?
Когда иерархическая кластеризация предпочтительнее k-means?
В чем разница между агломеративной и дивизивной иерархической кластеризацией?

15. DBSCAN и кластеризация на основе плотности. (2)
Почему DBSCAN плохо работает для данных с разной плотностью кластеров?
Почему DBSCAN лучше k-means для данных с кластерами произвольной формы?
Как ускорить DBSCAN для больших данных?

16. Оценка качества кластеризации. Индексы и кластерные оценки. (3)
Как оценить устойчивость кластеризации?
Как работает silhouette score? Как интерпретировать его значения?
Что такое индекс Davies-Bouldin? Когда он дает некорректные оценки?

Анализ данных

1. Грязные данные. Методы обработки пропущенных значений. (1)
   Какие типы пропусков есть и как их распознать?
   В чем разница между удалением и импутацией? Когда какой метод применять?
   Как оценить качество импутации?

2. Кодирование категориальных и численных признаков. (2)
   Какие есть различия между номинальными и порядковыми категориальными признаками?
   Какие алгоритмы могут привести к слишком большой размерности данных?
   Какие методы используются для преобразований скошенных распределений?

3. Значимость признака, отбор и фильтрация признаков. (2)
   В каких случаях можно удалить признак без вреда для модели?
   Почему корреляционный анализ не всегда достаточен для отбора признаков?
   Как проверить устойчивость отобранных признаков?

4. Нормализация признаков. Масштабирование, стандартизация. (2)
   Нужно ли масштабировать закодированные категориальные признаки?
   Напишите формулы для min-max масштабирования, стандартизации и Robust масштабирования.
   Когда что используется?

5. Выбросы. Обнаружение и борьба с выбросами. (2)
   Я хочу измерить время выполнение кода. Для этого я запускаю его 50 раз на случайных входных данных и считаю среднее. Насколько объективен этот показатель?
   При борьбе с выбросами, когда среднее нужно заменять медианой, а когда модой?
   Какие графики наиболее эффективны для визуального обнаружения выбросов?
   Как обнаружить выбросы в многомерных данных?

6. Формула Байеса, формула полной вероятности. (1)
   Какой смысл несет формула Байеса?
   В чём разница между априорной и апостериорной вероятностью?
   Применение к дереву решений.

7. Среднее арифметическое, геометрическое, гармоническое. Взвешенные средние. (1)
   Свойства средних значений.
   Неравенства, связанные со средними. Неравенство Коши.

8. Статистики и статистические оценки. Среднее, медиана, мода. Их свойства. (2)
   Дан график распределения. Покажите среднее, медиану и моду величины.
   Что означает отсутствие моды в данных?
   В каких случаях мода является более информативной характеристикой, чем среднее?

9. Меры изменчивости. Размах, межквартильный размах, дисперсия. (2)
   Зачем нужно среднеквадратичное отклонение, если есть дисперсия?
   Какие меры изменчивости есть для категориальных данных?

10. Математическое ожидание и дисперсия случайной величины. Их свойства. (2)
    Дано распределение вероятностей. Найдите дисперсию.
    Неравенство Чебышева и его смысл. Как оно позволяет оценить разброс?

11. Квантили, квартили и перцентили. (1)
Нарисован график распределения. Выделите квартили на этом графике.
Что такое Q-Q график? Что означает пересечение Q-Q графика с линией y=x?

12. Корреляции между величинами. Ковариация и ковариационная матрица. (2)
Нарисован корреляционный график, надо определить знак корреляции.
Как выбросы влияют на расчет корреляции?
Как показать зависимости трех и более величин?

13. Центральная предельная теорема. Стандартная ошибка среднего. (3)
Что такое доверительный интервал и как его считать через стандартную ошибку среднего?
Работает ли ЦПТ при маленьком количестве данных? Когда она неприменима?
Можем ли мы использовать ЦПТ не для среднего, а для медианы?

14. Статистические критерии и ключевые распределения (Z и χ^2). Проверка гипотез. (4)
В чем заключается принцип максимума правдоподобия?
Есть 4 разных вида мороженого. Продавец выдвигает гипотезу: все 4 вида пользуются одинаковым спросом. Он записывал за день количество купленного мороженного. Получилось 30, 45, 25, 20. Прав ли продавец?

15. Закон больших чисел. (3)
Когда закон больших чисел не применим?
Применим ли закон больших чисел для слабокоррелирующих величин?
Как применяется ЗБЧ в машинном обучении?

Оптимизация

1. Задача оптимизации. Свойства функций: выпуклость, гладкость. (2)
   Что такое глобальный и локальный минимум. Как они связаны с обучением нейросетей?
   Какие существуют типы ограничений в задачах оптимизации? Приведите примеры из ML.

2. Функции потерь для моделей машинного обучения. Ограничения для функций потерь. (2)
   В чем разница между функциями потерь и функциями стоимости?
   Как можно сгладить функции модуля, знака, ReLU?
   Как можно проверить выпуклость функции?

3. Производная и градиент. Теорема Ферма. Градиентный спуск. (3)
   Как градиент помогает для минимизации функции?
   Градиентный спуск и градиентный спуск с импульсом.
   Какими условия гарантируют сходимость обычного градиентного спуска?
   Зачем нужно уменьшать длину шага при градиентном спуске?

4. Метод наискорейшего спуска. (1)
   В чем различие его с градиентным спуском?
   Какие условия гарантируют сходимость МНС для выпуклых функций?

5. Стохастический градиентный спуск. (3)
   В чем принципиальное отличие от обычного градиентного спуска?
   Как связан размер батча с дисперсией градиента?
   Почему SGD может сходиться к неглобальному минимуму в невыпуклых задачах?

6. Метод имитации отжига. Применимость к задаче обучения моделей. (2)
   Гарантированно ли метод имитации отжига найдет минимум функции?
   Как правильно подбирать функцию остывания?

7. Функция правдоподобия и связь с функциями потерь. (3)
   Почему удобно логарифмировать функцию правдоподобия?
   Покажите, что минимизация log loss в логистической регрессии эквивалентна MLE для биномиального распределения.

8. Задача линейного программирования. Симплекс-метод. (1)
   Геометрическая интерпретация симплекс-метода.
   Какая вычислительная сложность симплекс-метода?

9. Метод сопряженных градиентов. (2)
   Опишите применение метода к решениям СЛАУ.
   Устойчив ли этот метод?

10. Adaptive gradient, RMSprop, adaptive movement estimation. (3)
    Когда какой алгоритм применять?
    Запишите формулы обновления параметров AdaGrad. В чем главный недостаток?
    Adam используется в ML/DL очень часто. Для каких задач он неприменим?

11. Оптимизация гиперпараметров. Валидация и регуляризация. (2)
    Как регуляризация влияет на процесс оптимизации?
    Как построить стратегию обучения с регуляризацией? Для чего нужна валидация?