Вопросы к экзамену
Машинное обучение

1. Что такое объект, целевая переменная, признак, модель, функционал ошибки и обучение?
2. Запишите формулы для линейной модели регрессии и для среднеквадратичной ошибки. Запишите среднеквадратичную ошибку в матричном виде.
3. Что такое коэффициент детерминации? Как интерпретировать его значения?
4. Чем отличаются функционалы MSE и MAE?
5. Что такое градиент? Какое его свойство используется при минимизации функций?
6. Как устроен градиентный спуск?
7. Почему не всегда можно использовать полный градиентный спуск? Какие способы оценивания градиента вы знаете? Почему в стохастическом градиентном спуске важно менять длину шага по мере итераций? Какие стратегии изменения шага вы знаете?
8. В чём заключаются метод инерции и AdaGrad/RMSProp?
9. Что такое кросс-валидация? На что влияет количество блоков в кросс-валидации? Как построить итоговую модель после того, как по кросс-валидации подобраны оптимальные гиперпараметры?
10. Чем гиперпараметры отличаются от параметров? Что является параметрами и гиперпараметрами в линейных моделях и в решающих деревьях?
11. Что такое регуляризация? Запишите L1- и L2-регуляризаторы.
12. Почему L1-регуляризация отбирает признаки?
13. Почему плохо накладывать регуляризацию на свободный коэффициент?
14. Запишите формулу для линейной модели классификации. Что такое отступ? Как обучаются линейные классификаторы и для чего нужны верхние оценки пороговой функции потерь?
15. Что такое точность, полнота и F-мера? Почему F-мера лучше арифметического среднего и минимума?
16. Для чего нужен порог в линейном классификаторе? Из каких соображений он может выбираться?
17. Что такое AUC-ROC? Опишите алгоритм построения ROC-кривой.
18. Что такое AUC-PRC? Опишите алгоритм построения PR-кривой.
19. Что означает “модель оценивает вероятность положительного класса”?
20. Что такое калибровочная кривая? Какие методы калибровки вероятности вы знаете? Почему важно проводить калибровку не на обучающей выборке?
21. Запишите функционал логистической регрессии. Как он связан с методом максимума правдоподобия?
22. Запишите задачу метода опорных векторов для линейно неразделимого случая. Как функционал этой задачи связан с отступом классификатора? Как выглядит задача безусловной оптимизации в SVM?
23. В чём заключаются one-vs-all и all-vs-all подходы в многоклассовой классификации?
24. Как измеряется качество в задаче многоклассовой классификации? Что такое микро- и макро-усреднение?

25. В чём заключается преобразование категориальных признаков в вещественные с помощью mean-target encoding? Почему использование этого способа кодирования может привести к переобучению? Какие методы борьбы с этой проблемой вам известны?
26. Как определить для линейной модели, какие признаки являются самыми важными?
27. Опишите жадный алгоритм обучения решающего дерева.
28. Почему с помощью бинарного решающего дерева можно достичь нулевой ошибки на обучающей выборке без повторяющихся объектов?
29. Как в общем случае выглядит критерий хаотичности? Как он используется для выбора предиката во внутренней вершине решающего дерева?
30. Для какой ошибки строится разложение на шум, смещение и разброс? Запишите формулу этой ошибки.
31. Запишите формулы для шума, смещения и разброса метода обучения для случая квадратичной функции потерь.
32. Приведите пример семейства алгоритмов с низким смещением и большим разбросом; семейства алгоритмов с большим смещением и низким разбросом. Поясните примеры.
33. Что такое бэггинг? Как его смещение и разброс связаны со смещением и разбросом базовых моделей?
34. Что такое случайный лес? Чем он отличается от бэггинга над решающими деревьями?
35. Что такое out-of-bag оценка в бэггинге?
36. Запишите вид композиции, которая обучается в градиентном бустинге. Как выбирают количество базовых алгоритмов в ней?
37. Что такое сдвиги в градиентном бустинге? Как они вычисляются и для чего используются?
38. Как обучается очередной базовый алгоритм в градиентном бустинге? Что такое сокращение шага?
39. Как в xgboost выводится функционал ошибки с помощью разложения в ряд Тейлора?
40. Какие регуляризации используются в xgboost?
41. Какие деревья используются в реализации catboost, в чем их особенность?
42. Как работает метод k ближайших соседей для классификации и для регрессии?
43. Как влияет увеличение размерности признакового пространства на KNN? Что такое проклятие размерности?
44. Опишите алгоритм LSH. Как устроены хэши для евклидовой метрики и для метры сходства Джакара?
45. Опишите алгоритмы поиска ближайших соседей NSW и HNSW.
46. Опишите метод LIME для интерпретации моделей.
47. Опишите метод SHAPE для интерпретации моделей.
48. Задача кластеризации. Метрики качества.
49. Метод K-Means, вывод его шагов.
50. Описание алгоритма DBSCAN.
51. Что такое иерархическая кластеризация?
52. Алгоритм спектральной кластеризации.
53. User-based и item-based подходы к рекомендациям.

54. Как выглядит модель со скрытыми переменными для рекомендательных систем? Какие данные необходимы для её обучения? Какие методы обучения этой модели вы знаете?
55. Опишите модель factorization machine.
56. В чём заключается идея неявных моделей для рекомендательных систем? Запишите функционал.
57. Как в рекомендательных системах можно учитывать контент?
58. Какие вы знаете метрики качества для рекомендательных систем? Что может измеряться помимо качества предсказания кликов?
59. Какие вы знаете подходы к холодному старту для пользователей и для айтемов?
60. Как свести задачу временных рядов к табличному виду? Рекурсивная стратегия. Прямая стратегия.