

Вопросы к контрольной работе
Машинное обучение 1
ПМИ ФКН ВШЭ

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. Что такое градиент? Какое его свойство используется при минимизации функций?
5. Запишите формулу для одного шага градиентного спуска. Какие способы оценивания градиента вы знаете? Почему не всегда можно использовать полный градиентный спуск?
6. Что такое кросс-валидация? На что влияет количество блоков в кросс-валидации?
7. Чем гиперпараметры отличаются от параметров? Что являются параметрами и гиперпараметрами в линейных моделях и в решающих деревьях?
8. Что такое регуляризация? Запишите L1- и L2-регуляризаторы. Почему L1-регуляризация отбирает признаки?
9. Запишите формулу для линейной модели классификации. Что такое отступ? Как обучаются линейные классификаторы и для чего нужны верхние оценки пороговой функции потерь?
10. Что такое точность, полнота и F-мера?
11. Что такое AUC-ROC? Опишите алгоритм построения ROC-кривой.
12. Запишите функционал логистической регрессии. Как он связан с методом максимума правдоподобия?
13. Запишите задачу метода опорных векторов для линейно неразделимого случая. Как функционал этой задачи связан с отступом классификатора?
14. В чём заключаются one-vs-all и all-vs-all подходы в многоклассовой классификации?
15. В чём заключается подход с независимой классификацией в задаче классификации с пересекающимися классами?
16. В чём заключается преобразование категориальных признаков в вещественные с помощью счётчиков? Почему использование счётчиков может привести к переобучению? Какие методы борьбы с этой проблемой счётчиков вам известны?
17. Какие способы калибровки вероятностей вы знаете? В чём они заключаются?
18. Как определить для линейной модели, какие признаки являются самыми важными?
19. Опишите жадный алгоритм обучения решающего дерева.
20. Почему с помощью бинарного решающего дерева можно достичь нулевой ошибки на обучающей выборке без повторяющихся объектов?

21. Как в общем случае выглядит критерий информативности? Как он используется для выбора предиката во внутренней вершине решающего дерева? Как вывести критерий Джини и энтропийный критерий?
22. Для какой ошибки строится разложение на шум, смещение и разброс?
Запишите формулу этой ошибки.
23. Запишите формулы для шума, смещения и разброса метода обучения.
24. Приведите пример семейства алгоритмов с низким смещением и большим разбросом; семейства алгоритмов с большим смещением и низким разбросом.
Поясните примеры.
25. Что такое бэггинг?
26. Что такое случайный лес? Чем он отличается от бэггинга над решающими деревьями?
27. Запишите вид композиции, которая обучается в градиентном бустинге. Как выбирают количество базовых алгоритмов в ней?
28. Что такое сдвиги в градиентном бустинге? Как они вычисляются и для чего используются?
29. Как обучается очередной базовый алгоритм в градиентном бустинге? Что такое сокращение шага?
30. В чём заключается переподбор прогнозов в листьях решающих деревьев в градиентном бустинге?
31. Как в xgboost выводится функционал ошибки с помощью разложения в ряд Тейлора?
32. Какие регуляризации используются в xgboost?
33. Задача кластеризации. Метрики качества.
34. Метод K-Means, вывод его шагов.
35. Задача визуализации и метод t-SNE.
36. Понижение размерности. Метод главных компонент: постановка задачи и решение.
37. Постановка задачи построения рекомендаций. Коллаборативные методы: memory-based и на основе моделей со скрытыми переменными. Контентные методы.