Вопросы к контрольной работе
Машинное обучение
Магистерская программа ФТиАД, 2018

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 квантильная функция потерь? В каких случаях ее используют?
4. Запишите вероятностную модель, оптимизация правдоподобия в которой равносильна минимизации среднеквадратичной ошибки с L2-регуляризацией. Покажите, почему это так.
5. Что такое градиент? Какое его свойство используется при минимизации функций?
6. Запишите формулу для одного шага градиентного спуска. Какие способы оценивания градиента вы знаете? Почему не всегда можно использовать полный градиентный спуск?
7. Что такое кросс-валидация и для чего она используется? Чем применение кросс-валидации лучше, чем разбиение выборки на обучение и контроль?
8. Чем гиперпараметры отличаются от параметров? Что является параметрами и гиперпараметрами в линейных моделях и в решающих деревьях?
9. Для чего нужно нормировать данные при обучении линейных моделей? Какие способы нормировки вы знаете?
10. Что такое регуляризация? Запишите L1- и L2-регуляризаторы. Почему L1-регуляризация отбирает признаки?
11. Запишите формулу для линейной модели классификации. Что такое отступ? Как обучаются линейные классификаторы и для чего нужны верхние оценки пороговой функции потерь?
12. Что такое точность, полнота и F-мера?
13. Что такое AUC-ROC? Опишите алгоритм построения ROC-кривой.
14. Запишите функционал логистической регрессии. Как он связан с методом максимума правдоподобия?
15. Запишите задачу метода опорных векторов для линейно неразделимого случая. Как функционал этой задачи связан с отступом классификатора?
16. В чём заключаются one-vs-all и all-vs-all подходы в многоклассовой классификации?
17. В чём заключается подход с независимой классификацией в задаче классификации с пересекающимися классами?
18. Опишите жадный алгоритм обучения решающего дерева.
19. Почему с помощью бинарного решающего дерева можно достичь нулевой ошибки на обучающей выборке без повторяющихся объектов?
20. 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 (записать вывод)?

21. Запишите решающее правило, по которому во взвешенном kNN делают предсказания в задаче классификации и регрессии. Как выбирают веса в методе окна Парзена?
22. Что такое проклятие размерности?
23. Запишите формулы для расстояния Минковского, косинусного расстояния, расстояния Джакарда.
24. Что такое бэггинг?
25. Что такое случайный лес? Чем он отличается от бэггинга над решающими деревьями?
26. Запишите вид композиции, которая обучается в градиентном бустинге. Как выбирают количество базовых алгоритмов в ней?
27. Что такое сдвиги в градиентном бустинге? Как они вычисляются и для чего используются?
28. Как обучается очередной базовый алгоритм в градиентном бустинге? Что такое сокращение шага?
29. В чём заключается переподбор прогнозов в листьях решающих деревьев в градиентном бустинге?
30. Что такое стекинг? Для чего они используется?
31. Для какой ошибки строится разложение на шум, смещение и разброс? Запишите формулу этой ошибки.
32. Запишите формулы для шума, смещения и разброса метода обучения.
33. Приведите пример семейства алгоритмов с низким смещением и большим разбросом; семейства алгоритмов с большим смещением и низким разбросом. Поясните примеры.
34. Что такое задача кластеризации? На какие две группы делятся критерии качества кластризации? Чем эти две группы отличаются? Запишите формулы для внутрикластерного и межкластерного расстояний.
35. Как работает метод K-Means? Какой критерий он оптимизирует?
36. Как работает метод DBSCAN?
37. Что такое задача понижения размерности? По какому правилу вычисляются новые признаки в методе главных компонент (PCA)? Какой критерий оптимизируют? В чем состоит решение этой задачи оптимизации?
38. Запишите две формулировки задачи PCA в терминах проецирования выборки на маломерное подпространство.
39. Что такое задача визуализации? Какой критерий оптимизируют в методе MDS?
40. Каким образом в случайных лесах можно оценить важности признаков (2 метода)?