Вопросы к контрольной работе
Машинное обучение

На все вопросы оижадются развернутые ответы с формулами. Во всех вопросах необходимо уметь расшифровывать все обозначения.

1. Что такое объект, целевая переменная, признак, модель, функция потреь, функционал ошибки и обучение?
2. Какие бывают типы признаков?
3. Что такое переобучение и недообучение? Как отличить переобучение от недообучения?
4. Что такое кросс-валидация и для чего она используется? Чем применение кросс-валидации лучше, чем разбиение выборки на обучение и контроль?
5. Чем гиперпараметры отличаются от параметров? Что является параметрами и гиперпараметрами в линейных моделях, в решающих деревьях, в методе k ближайших соседей?
6. Запишите формулы для линейной модели регрессии и для среднеквадратичной ошибки. Запишите среднеквадратичную ошибку в матричном виде.
7. Что такое градиент? Какое его свойство используется при минимизации функций?
8. Запишите формулу для одного шага градиентного спуска. Чем отличается один шаг стохастического градиентного спуска? Почему не всегда можно использовать полный градиентный спуск?
9. Для чего нужно нормировать данные при обучении линейных моделей? Какие способы нормировки вы знаете?
10. Что такое регуляризация? Для чего ее используют в линейных моделях? 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. Как в логистической регрессии выполняются предсказания для новых объектов? Запишите функционал логистической регрессии. Как он связан с методом максимума правдоподобия?
13. Что такое точность, полнота и F-мера?
14. Что такое AUC-ROC? Опишите алгоритм построения ROC-кривой.
15. В чём заключаются one-vs-all и all-vs-all подходы к многоклассовой классификации?
16. В чём заключается подход с независимой классификацией в задаче классификации с пересекающимися классами?
17. Что такое микро- и марко-усреднение при оценивании качества многоклассовой классификации?
18. Что такое решающее дерево? Запишите формулу предсказания решающего дерева (через разбиение признакового пространства на области).
19. Опишите жадный алгоритм обучения решающего дерева.

20. Какими основными свойствами должен обладать критерий информативности? Как он используется для выбора предиката во внутренней вершине решающего дерева?
21. Запишите критерий Джини и энтропийный критерий информативности.
22. Как во взвешенном методе k ближайших соседей выполняются предсказания в задачах классификации и регрессии? Приведите примеры выбора весов объектов. Как выполняются предсказания в методе парзеновского окна?
23. Каковы проблемы использования метода k ближайших соседей на практике?
24. Запишите формулы для следующий функций расстояния: расстояние Минковского, евклидово расстояние, манхэттэнское расстояние, расстояние Чебышева, косинусное расстояние, расстояние Хеллингера, симметризованная KL-дивергенция, расстояние Джаккарда, редакторское расстояние. В каком признаковом пространстве используется каждая функция расстояния?
25. Что такое композиция алгоритмов машинного обучения? Покажите, что в предположении некоррелированных ошибок базовых алгоритмов, ошибка композиции будет в N раз меньше, чем средняя ошибка базовых алгоритмов, где N - число базовых алгоритмов.
26. Что такое бэггинг? Что такое случайный лес? Что такое out-of-bag ошибка, для чего она используется?
27. Опишите алгоритм построения композиции методом градиентного бустинга. Что такое сдвиги?
28. Что такое сокращение шага в градиентном спуске и для чего оно используется?