Курс “Основы машинного обучения”

Вопросы к контрольной работе

1. **Что такое объекты и признаки в машинном обучении? Для чего нужен функционал ошибки? Что такое алгоритм (модель)?**

**Объекты** это данные какого-то датасета или выборки, то есть то, на чем будет модель обучаться и тестироваться. Например, если у нас есть датасет людей, то каждый человек это объект. То есть в нашем датасете строки это объекты. Обычно объект обозначают за x, либо x\_i если объектов несколько.

Также объекты имеют **ответы**, то есть целевую переменную. Если у нас задача определить уровень образования человека в предыдущем датасете, то ответом будет являться как раз этот уровень образования (в данном случае это многоклассовая классификация)

**Признаки** это характеристики объектов. Например, в предыдущем примере это рост, вес, семейное положение и тд. В нашем датасете это столбцы. Признаки могут быть категориальными, бинарными, вещественными и тд.

**Моделью (или алгоритмом)** называется такая функция, которая для любого объекта выдает ее ответ. Чтобы измерить качество работы алгоритма существует понятие **функционал качества.** Если функционал качество нужно минимизировать, то логичнее называть его **функционалом ошибки.** Чем более маленькое значение этого функционала дает алгоритм, тем он лучше.

1. **Чем задача классификации отличается от задачи регрессии? Приведите примеры задач классификации и регрессии.**
2. **Что такое вещественные (числовые), бинарные, категориальные признаки? Приведите примеры.**
3. **В чём заключается обобщающая способность алгоритма машинного обучения? К чему приводит её отсутствие? Что такое переобучение?**

Обобщающая способность алгоритма это способность обобщить данные, способность понять как именно устроены данные. Отсутствие обобщающей способности приводит к переобучению, так как алгоритм просто запоминает обучающую выборку.

1. **Что такое гиперпараметр? Чем гиперпараметры отличаются от обычных параметров алгоритмов? Приведите примеры параметров и гиперпараметров в линейных моделях.**

**Параметр модели** - это переменная выбранной модели, которая может быть оценена путем подгонки заданных данных к модели. **Гиперпараметр модели** - это параметр, значение которого устанавливается перед началом обучения модели. Их нельзя узнать, подгоняя модель к данным.

В линейных моделях параметрами называются коэффициенты w0, w1, w2, которые изменяются в процессе обучения. Гиперпараметром у линейной модели Lasso является гиперпараметр alpha

(1 / (2 \* n\_samples)) \* ||y - Xw||^2\_2 + alpha \* ||w||\_1

1. **Что такое отложенная выборка? Что такое кросс-валидация (скользящий контроль)? Как ими пользоваться для выбора гиперпараметров?**

Отложенная выборка это какая-то часть датасета с ответами, на которой измеряется качество предсказаний. но проблема отложенной выборки в том, как данные разделить и действительно ли отложенная выборка будет репрезентативной. Поэтому используют кросс валидацию это такой метод оценивания, в которой датасет делится на несколько фолдов, на n-1 частях обучается, а на 1 оценивается (которая не участвовала в обучении). так обучается k моделей, в которой каждый блок играет роль отложенной выборки. далее результаты усредняются. Чтобы выбрать лучшие гиперпараметры для модели, обычно используют отложенную выборку и выбирают те гиперпараметры, для которых качество предсказаний было наивысшим.

1. **Как метод k ближайших соседей определяет класс для нового объекта?**

измеряет расстояние между новым объектом и объектами, которые модель запомнила в процессе обучения и выбирает k ближайших. дальше выбирает самый популярный класс среди соседей.

1. **Опишите метод k ближайших соседей с парзеновским окном. Какие в нём есть параметры?**
2. Запишите формулу метода kNN для регрессии.
3. Что такое градиент? Какое его свойство используется в машинном обучении?
4. Опишите алгоритм градиентного спуска.
5. Что такое стохастический градиентный спуск? В чём его отличия от обычного градиентного спуска? Какие у него плюсы и минусы?
6. Как обучается линейная регрессия?
7. Что такое регуляризация? Как она помогает бороться с переобучением?
8. Чем L1-регуляризация отличается от L2-регуляризации?
9. Что такое масштабирование признаков? Как его проводить? Зачем это нужно?
10. Чем функционал MSE отличается от MAE? Что такое функция потерь Хубера и для чего она нужна?
11. Как выглядит модель линейной классификации в случае двух классов?
12. Что такое отступ? Для чего он нужен?
13. Как обучаются линейные классификаторы (общая схема с верхними оценками)?
14. Для чего может понадобиться оценивать вероятности классов?
15. Как обучается логистическая регрессия? Запишите функционал и объясните, откуда он берётся.,/
16. Как в логистической регрессии строится прогноз для нового объекта?
17. Что такое метод опорных векторов? Опишите его основную идею.
18. Как устроены метрики accuracy, precision, recall? Что такое F-мера? Чем она лучше арифметического среднего точности и полноты?
19. Для чего нужны ROC- и PR-кривые? Как они строятся? Что такое AUC-ROC и AUC-PRC?
20. Какие метрики можно использовать в случае сильного дисбаланса классов?
21. Как можно свести задачу многоклассовой классификации к серии задач бинарной классификации?
22. Что такое решающее дерево? Как оно строит прогноз для объекта? Как обучаются решающие деревья в задачах классификации и регрессии (и что такое критерии информативности)?
23. Какие вы знаете критерии останова и способы выбора значений в листьях? Какие гиперпараметры имеются у деревьев?
24. Что такое бэггинг и метод случайных подмножеств? Что такое случайный лес, как он обучается и как он строит прогнозы?
25. В чём идея разложения ошибки на смещение и разброс? Как бэггинг меняет смещение и разброс одной модели?
26. Опишите идею градиентного бустинга для среднеквадратичной ошибки. Запишите задачу для обучения очередной базовой модели.
27. Опишите идею градиентного бустинга для произвольной функции потерь. Запишите задачу для обучения очередной базовой модели.
28. Какими способами можно побороть переобучение модели:
    1. в общем случае (независимо от модели);
    2. если модель линейная;
    3. если модель — градиентный бустинг.