Билеты к экзамену по машинному обучению

Тинькофф - Финтех школа. Осень 2019.

Процедура экзамена

Вы получаете 2 билета и 30 минут на подготовку. В процессе подготовки можно использовать любые материалы (собственные записи, материалы курса, интернет) кроме посторонней помощи. Далее вы отвечаете уже без использования каких-либо материалов. Ваш ответ начинается со случайного опроса по вопросам теоретического минимума, далее по билетам, и вам могут задаваться дополнительные вопросы.

Билеты

1. Виды обучения: с учителем (supervised), без учителя (unsupervized), частичное (semi-supervised), трансдуктивное. Типы признаков и типы откликов. Принцип минимизации эмпирического риска. Переобучение, ее зависимость от размера обучающей выборки и сложности модели. Кросс-валидация и A/B тестирование. Дискриминантные функции.

2. Метод ближайших центроидов и K ближайших соседей. Проклятие размерности. Взвешенный учет объектов. Пример весов. Отступ и классификация объектов на выбросы, пограничные, типичные, эталонные. Методы фильтрации обучающей выборки для ускорения работы метода.

3. Методы KD-деревьев и ball-деревьев для ускорения метода K ближайших соседей.

4. Нормализация признаков. Преобразование категориальных признаков в бинарные и вещественные. Метод обычного и стохастического градиентного спуска.

5. Оценка классификаторов: точность, полнота, F-мера, матрица ошибок, ROC-кривая, AUC. Оценка качества предсказания вероятностей классов.

6. Регрессия. Вывод решения для гребневой регрессии. Алгоритм робастной регрессии. Регрессия опорных векторов. L_1 и L_2 регуляризация.

7. Методы многоклассовой классификации бинарными классификаторами. Отступ (margin) для многоклассового случая и случая 2х классов. Оптимизационная задача по настройке весов бинарного классификатора. Основные функции потерь. L_1 и L_2 регуляризация.

8. Определение логистической регрессии через вероятности классов. Какой функции потерь она соответствует? Многомерная логистическая регрессия. Функция soft-max.

9. Метод опорных векторов в линейно разделимом и линейно неразделимом случае. Его вывод из максимизации расстояния между классами. Какой функции потерь и регуляризации он соответствует? Классификация типов объектов в методе опорных векторов.

10. Обобщение методов машинного обучения через ядра. Теорема Мерсера. Операции, не выводящие из класса ядер. Линейное, полиномиальное и RBF ядра. Определение расстояния через ядра.

11. Решение для метода опорных векторов через двойственную задачу и его обобщение через ядра.

12. Определение решающего дерева. Выбор решающего правила в каждом узле для случая классификации/регрессии (для деревьев CART), назначение прогнозов узлам дерева в случае регрессии/классификации, симметричных/несимметричных потерь.

13. Определение решающего дерева. Правиловые критерии остановки наращивания дерева. Обрезка (pruning) для решающих деревьев CART.

14. Фиксированные схемы агрегации классификаторов, выдающих метки, рейтинги и вероятности классов. Стэкинг, бэггинг, метод случайных подпространств. Ошибка прогнозирования в зависимолсти от неопределенности ансамбля (ambiguity decomposition).

15. Разложение ожидаемых ошибок на смещение и дисперсию (bias-variance decomposition). Методы случайного леса (RandomForest) и особо случайных деревьев (ExtraRandomTrees).

16. Бустинг - алгоритм последовательного наращивания ансамбля моделей (FSAM). Алгоритм градиентного бустинга. Использование shrinkage, subsampling.

17. Алгоритм AdaBoost - предположения и аналитический вывод решения.

18. Метод xgBoost.

19. Методы lightGBM и DART. Отличие бустинга, использующего разложение Тейлора 1го порядка и 2го порядка.

20. Отбор признаков по корреляции, взаимной информации и relief-критерию.

21. Алгоритм последовательного отбора признаков и его модификации. Алгоритм генетического отбора признаков.

22. Задача снижения размерности. Метод главных компонент - 2 определения (через проекции и отклонения), их эквивалентность. Оценка качества аппроксимации отдельной компонентой и первыми K компонентами.

23. Метод главных компонент - определение и итеративный алгоритм их построения. Доказательство, что итеративный алгоритм действительно дает главные компоненты (полученные компоненты удовлетворяют определению).

24. Сингулярное разложение, его основные свойства.

25. Baseline-алгоритм для коллаборативной фильтрации. Алгоритмы user-user и item-item. Какой из них применим в онлайн режиме?

26. Алгоритм разреженного сингулярного разложения для коллаборативной фильтрации.

Теоретический минимум

1. Принцип минимизации эмпирического риска.
2. Классификация с помощью дискриминантных функций.
3. Определение линейного классификатора (бинарный/многоклассовый).
4. Отступ для классификатора (бинарного/многоклассового).
5. Типичные функции потерь для регрессии и бинарной классификации.
6. Матрица ошибок. Точность, полнота, F-мера, ROC кривая, AUC.
7. Обобщение методов через ядра. Типичные ядра: линеное, полиномиальное, Гауссово (RBF). Почему они являются ядрами Мерсера?
8. Решающие правила в дереве CART и алгоритм их выбора. Возможные критерии информативности для регрессии и классификации.
9. Типичные регуляризаторы, какой из может отбирать признаки и почему?
10. Логистическая регрессия (бинарная, многоклассовая).
11. Метод опорных векторов.
12. Регрессия опорных векторов.
13. Гребневая регрессия и вывод оптимальных весов для нее.
14. Алгоритм случайного леса (RandomForest) и особо случайных деревьев (ExtraRandomTrees)
15. Алгоритм градиентного бустинга.
16. Идеи xgBoost, lightGBM, DART.
17. Метод последовательного отбора признаков.
18. Метод главных компонент - определение.
19. Сингулярное разложение.
20. Методы user-user и item-item коллаборативной фильтрации.