Программа экзамена по курсу ML

1. Постановка основных задач (13.09)

● Целевая функция, объект, метка, пространство объектов, признаковое пространство, функция ошибки, эмпирический риск, обучающая выборка
● Обучение с учителем, типы задач обучения с учителем
● Алгоритм, модель алгоритмов, обобщая способность
● Схема решения задачи машинного обучения

2. Математика в машинном обучении (20.09)

● Основы теории вероятностей и матстатистики: распределения, формулы пересчёта вероятностей, математическое ожидание, дисперсия
● Точечное оценивание, оценка максимального правдоподобия
● Оценка плотности: непараметрический и параметрический подходы
● Сингулярное разложение матриц

3. Метрические алгоритмы (25.10 - 01.11)

● Понятие метрического алгоритма (distance-based)
● Метод ближайшего центроида (Nearest centroid algorithm)
● Метод k ближайших соседей (kNN) для классификации и регрессии
● Весовые обобщения kNN
● Примеры функций расстояния в методе kNN
● Регрессия Надарая-Ватсона
● LSH для быстрого поиска ближайших соседей

4. Контроль качества и выбор модели (18.10)

● Проблема контроля качества
● Общие правила разбиения выборки
● Отложенный контроль (held-out data, hold-out set)
● Скользящий контроль / перекрёстная проверка (cross-validation)
● Бутстреп (bootstrap)

5. Оптимизация в машинном обучении (27.09)

● Типы методов оптимизации: нулевого, первого, второго порядков
● Градиентный спуск (GD) и стохастический градиентный спуск (SGD)
● Критерии останова
● Обучение: пакетное, онлайн, по минибатчам

6. Линейная и логистическая регрессии (27.09 - 11.10)

● Линейная регрессия, прямой метод нахождения решения
● Проблема вырожденности матрицы
● Регуляризация. Гребневая регрессия (Ridge Regression). LASSO. Elastic Net.
● Селекция признаков при использовании LASSO
● Устойчивая регрессия (Robust Regression), RANSAC (RANdom SAmple Consensus)
● Логистическая регрессия, нахождение решения через SGD
● Многоклассовая логистическая регрессия

7. Линейные модели классификации (11.10)

● Линейный классификатор, использование суррогатных функций (surrogate loss functions)
● Персептронный алгоритм (perceptron)
● Hinge Loss
● Метод опорных векторов (SVM), постановка задачи
● Решения задач условной оптимизации. Условия Каруша-Куна-Такеера.
● SVM: решение прямой задачи
● SVM: решение обратной задачи
● Soft-Margin SVM: разделение допуская ошибки

8. Нелинейные методы (13.12)

● Проблема линейности
● Полиномиальная модель, базисные функции, радиально-базисная функция (RBF)
● Ядерные методы (Kernel Tricks), опредление ядра, примеры ядер
● Ядерный SVM
● Ядерная Rige регрессия
● Операции над ядрами

9. Деревья решений (07.02)

● Деревья решений (CART). Построение дерева.
● Критерии расщепления в задачах классификации (missclassification criteria, энтропийный, Джини) и регрессии
● Критерии остановки при построении деревьев
● Проблема переобучения для деревьев. Подрезка (post-pruning).
● Подсчёт важности признаков на основе решающего дерева
● Учёт пропусков (Missing Values)

10. Ансамбли алгоритмов (07.02 - 21.02)

● Ансамбли алгоритмов: примеры и обоснование
● Способы повышения разнообразия в ансамбле
● Бэгинг (bootstrap aggregating)
● OOB-prediction и OOB-estimation
● ~~Стекинг (stacking) и блендинг~~
● Бустинг
~~● AdaBoost (алгоритм, вывод формул)~~

11. Случайный лес (14.02)

● Случайный лес
● Настройка параметров методов
● Extreme Random Trees

12. Градиентный бустинг (21.02)

● Градиентный бустинг над решающими деревьями
● Настройка параметров методов
● Продвинутые методы оптимизации бустинга
● Современные реализации градиентного бустинга (XGBoost, LightGBM, CatBoost) и их особенности
● Способы работы с категориальными признаками

13. Сложность алгоритмов, переобучение, смещение и разброс (24.04)

● Проблема обобщения алгоритмов
● Bias-variance decomposition для задачи регрессии и квадратичного функционала
● Способы борьбы с переобучением

14. Функции ошибки / функционалы качества (03.04 - 17.04)

● Базовые функции ошибки в задаче регрессии (средний модуль отклонения (MAE), средний квадрат отклонения (MSE) и его производные, вероятностное и невероятностное обоснование RMSE)
● Базовые функционалы качества в задаче классификации (матрица ошибок

(Confusion Matrix), точность (Accuracy, MCE), ошибки 1 и 2 рода, полнота (Recall, TPR), специфичность (TNR), точность (Precision), FPR(False Positive Rate), F1-мера)
● Базовые скоринговые ошибки (Log Loss, AUROC)
● Качество в многоклассовых задачах. Разные виды усреднения качества: макро, микро, весовое, по объектам.

~~15. Отбор признаков (?)~~

~~● Причины отбора признаков. Классификация методов отбора: фильтры, обёртки, встроенные методы.~~
~~● Оценка стабильности и зависимости признака.~~
~~● Отбор как задача глобальной оптимизации. Классы алгоритмов метаоптимизации: перебор, направленный поиск (градиентный алгоритм, симуляция отжига, метод луча (beam search), локальный поиск), стохастическая оптимизация (генетический алгоритм).~~
~~● Проблема «исследования-использования».~~

17. Визуализация (28.02 - 13.03)

● Одномерный анализ, описательные статистики: среднее, характерные элементы, разброс значений, абсолютные вариации, относительные вариации, моменты, стандартизованные моменты.
● Многомерный анализ, визуализация пары признаков.

18. Кластеризация данных (29.11 - 06.12)

● Задача кластеризации, типы кластеризации
● k-средних (Lloyd’s algorithm), обобщения k-means, soft k-Means
● Affinity propagation: кластеризация сообщениями между точками
● Сдвиг среднего (Mean Shift): обнаружение мод плотности
● Иерархическая кластеризация (Hierarchical clustering)
● Разные виды связности - Linkage, графовые методы, Minimum Spanning Tree (MST) – на основе минимального остовного дерева
● Spectral Clustering – Спектральная кластеризация
● DBSCAN = Density-Based Spatial Clustering of Applications with Noise
● BIRCH = Balanced Iterative Reducing Clustering using Hierarchies
● EM-алгоритм, обоснование EM-алгоритма

~~19. Байесовский подход (?)~~

~~● Формула Байеса~~
~~● Оптимальное решение задач классификации – байесовский алгоритм, частный случай онормальных распределений.~~
~~● Минимизация среднего риска, Наивный байес (naive Bayes)~~
~~● Байесовский подход в машинном обучении, Метод максимального правдоподобия, MAP~~
~~● Байесовская теория для линейной регрессии~~
~~● Иерархические модели, RVM, что такое случайность~~

20. Подготовка данных, генерация признаков (15.11, 27.03)

● Фундаментальные свойства данных. Виды данных. Предобработка данных.
● Очистка данных (Data Cleaning): аномалии/выбросы, пропуски, шум, некорректные значения.
● Сокращение данных (Data Reduction): сэмплирование, сокращение размерности, отбор признаков, отбор объектов.
● Трансформация данных (Data Transformation): переименование признаков, объектов, значений признаков, преобразование типов; кодирование значений категориальных переменных; дискретизация; нормализация; сглаживание; создание признаков; агрегирование; обобщение; деформация значений.
● Интеграция данных.
● Типы числовых признаков. Контекстные признаки. Служебные признаки. Утечка в данных. Странности в данных.
● Географические (пространственные) признаки: Spatial Variables. (проекции на разные оси, кластеризация, идентификация, привязка, характеристики окрестности, анализ траекторий, деанонимизация данных, использование контекста и исследование странностей, генерация расстояний и использование для других признаков).
● Финансовые и временные признаки

21. Обучение на неразмеченных данных

● Задачи USL
● Сокращение размерности: SVD, PCA, kernel PCA, LLE, t-SNE, IsoMap, MDS, Maximum Variance Unfolding, Spectral Embedding

22. Специальные задачи

● Классификация на несколько классов
● Задачи с дисбалансом классов