Программа экзамена по курсу ML

1. Постановка основных задач (07.09 - 14.09)
    - 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 (14.09 - 21.09)
    - 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 (28.09 - 05.10, семинар 28.09)
    - 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 (05.10-12.10)
    - 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 (12.10 - 19.10)
    - 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 (19.10 - 26.10, семинар 19.10)
   - 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 (02.11 - 09.11, семинар 09.11)
   - 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 (16.11 - 23.11)
   - 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 (23.11 - 30.11)
   - 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 (30.11)
    - 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 (07.12)
    - Случайный лес
    - Настройка параметров методов
    - Extreme Random Trees

12. Градиентный бустинг (21.12)
    - 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 (10.02)
    - Проблема обобщения алгоритмов
    - Bias-variance decomposition для задачи регрессии и квадратичного функционала
    - Способы борьбы с переобучением

14. Функции ошибки / функционалы качества (17.02 - 24.03)
    - 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.