Программа по курсу
«Машинное обучение и интеллектуальный анализ данных»

1. Основные определения: прецедент, обучающая выборка, признаки объектов, виды признаков, матрица объектов-признаков. Модель алгоритмов, метод обучения, функционал качества алгоритма.
2. Вероятностная постановка задачи обучения. Принцип максимума правдоподобия. Связь максимизации правдоподобия и минимизации эмпирического риска.
3. Проблема переобучения и обобщающая способность алгоритма. Состоятельные методы обучения. Эмпирические оценки обобщающей способности.
4. Выбор алгоритма для вероятностной постановки задачи. Функционал среднего риска.
5. Метрические алгоритмы классификации. Обобщенный метрический классификатор. Виды и особенности частных случаев: методы ближайшего соседа, k ближайших соседей, взвешенных соседей, парзеновского окна постоянной и переменной ширины.
6. Классификация объектов по значению отступа. Алгоритм STOLP отбора эталонных объектов. Выбор метрики и проклятие размерности.
7. Приближенное вычисление плотности распределения. Парамерический и непараметрический подходы. Наивный байесовский классификатор. Одномерный случай. Многомерный случай. Проблемы мультиколлинеарности и выбросов.
8. Логистическая регрессия. Случайные величины с экспонентным законом распределения. Теорема о линейности байесовского классификатора (с доказательством). Бинаризация признаков. Скоринг.
9. Смеси распределений. EM-алгоритм разделения смеси. Смеси многомерных нормальных распределений.
10. Линейные алгоритмы классификации. Модель Мак Каллока-Питтса, алгоритм стохастического градиента для минимизации функционала среднего риска. Частные случаи. Сходимость метода СГ с правилом Хэбба с доказательством. Эвристики для улучшения сходимости и обобщающей способности.
11. Кривая ошибок ROC и AUC. Формула вычисления AUC. Градиентная максимизация AUC.
12. Метод опорных векторов (SVM). Случай линейно разделимой выборки. Случай линейно неразделимой выборки. Функция Лагранжа. Классификация объектов в зависимости от значений множителей Лагранжа. Двойственная задача. Обучение SVM. Нелинейные обобщение SVM. SVM-регрессия. Lasso SVM.
13. Алгоритмы восстановления регрессии. Метод наименьших квадратов. Многомерная линейная регрессия. Подход с использованием SVD-разложения матрицы. Гребневая регрессия. Метод главных компонент PCA. Непараметрическая регрессия. Проблема выбросов. Алгоритм LOWESS
14. Логические методы классификации. Понятие информативности предиката: эвристическое, вероятностное, энтропийное. Поиск информативных закономерностей. Построение решающего списка и решающего дерева. Редукция деревьев. Применение деревьев для решения задачи регрессии. Небрежные решающие деревья.
15. Композиции алгоритмов. AdaBoost. AnyBoost. Градиентный бустинг. Бэггинг, метод случайных подпространств. Случайные лес.
16. Ранжирование и рекомендательные системы. Постановка задачи. Оценки качества. Алгоритмы построения ранжирующих систем: поточечный, попарный и списочный. Их сильные и слабые стороны.
17. Тематическое моделирование. Векторная модель текста, TF-IDF. Недостатки векторной модели. Тематические модели: LSA, PLSA. Распределение Дирихле. Тематическая модель LDA.
18. Кластеризация. Близость и связанность. EM-алгоритм, метод k-средних. DBSCAN. Выбор Eps и MinPts.
19. Задачи компьютерного зрения. Признаки изображений: глобальные, локальные. Применение сверточных нейронных сетей для построения признаков.
20. Прогнозирование временных рядов. Модель авторегрессии. Модель скользящего среднего. Модель ARMA. ARIMA - интегрированная ARMA. Подбор параметров модели. Авторегрессионный спектр