АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ
МАШИННОЕ ОБУЧЕНИЕ

Автор: Стефановский Дмитрий Владимирович
Код и наименование направления подготовки, профиля: 38.04.01 Экономика («Системы больших данных в экономике»)
Квалификация (степень) выпускника: Магистр
Форма обучения: очная

Цель освоения дисциплины:

Сформировать компетенции в сфере аналитической работы в области экономики и финансов, обработки и анализов данных.

План курса:

Тема 1. Введение в машинное обучение. Цели и основная проблематика машинного обучения.

Существующие, наборы данных, визуализация модели классификации. Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, прогнозирование, ранжирование.

Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль.

Линейные модели регрессии и классификации. Метод наименьших квадратов. Полиномиальная регрессия.

Тема 2. Методы оценки точности полученных решений, включая ROC анализ.

Линейный регрессионный анализ, чувствительность, специфичность и точность. Корреляционный анализ. Анализ выживаемости и многомерная статистика. Таблицы дожития (mortality table) и метод Каплана-Мейера (Kaplan-Meier method). Лог-ранк тест. Модель Кокса.

Тема 3. Современные регрессионные методы, включая эластичные сети, регрессионные деревья и леса. Стандартный метод наименьших квадратов. Методы распознавания.

Логистическая регрессия. Автокорреляционная функция. Алгоритм Левенберга-Марквардта. Алгоритмы выбора линейных регрессионных моделей. Вспомогательные функции. Анализ регрессионных остатков. Аппроксимация Лапласа.

Регрессионные деревья и леса. Методы распознавания.

Тема 4. Байесовские методы и другие статистические модели, включая логистическую регрессию и др.

Понятие о случайном процессе. Байесовский подход к статистическому оцениванию. Априорные распределения, сопряженные с наблюдаемой генеральной совокупностью. Байесовский прогноз зависимой переменной, основанный на нормальной линейной модели множественной регрессии. Проверка статистических гипотез при байесовском подходе.

Тема 5. Нейросетевые методы. Современные подходы и идеи.

Биологический нейрон, модель МакКаллока-Питтса как линейный классификатор. Функции активации. Проблема полноты. Задача исключающего или. Полнота двухслойных сетей в пространстве булевых функций. Теоремы Колмогорова, Стоуна, Горбаня (без

доказательства). Алгоритм обратного распространения ошибок. Эвристики: формирование начального приближения, ускорение сходимости, диагональный метод Левенберга-Марквардта. Проблема «паралича» сети. Метод послойной настройки сети. Подбор структуры сети: методы постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage). Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM.

Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена.

Тема 6. Метод опорных векторов.

Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin).

Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно-линейная функция потерь. Задача квадратичного программирования и двойственная задача. Понятие опорных векторов. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера.

Способы конструктивного построения ядер. Примеры ядер.

SVM-регрессия.

Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

Метод релевантных векторов RVM.

Тема 7. Решающие деревья и леса.

Понятие логической закономерности.

Параметрические семейства закономерностей: конъюнкции пороговых правил, синдромные правила, шары, гиперплоскости.

Переборные алгоритмы синтеза конъюнкций: стохастический локальный поиск, стабилизация, редукция. Двухкритериальный отбор информативных закономерностей, парето-оптимальный фронт в (p,n)-пространстве. Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения. Вывод критериев ветвления. Мера нечистоты (impurity) распределения. Энтропийный критерий, критерий Джини. Редукция решающих деревьев: предредукция и постредукция. Алгоритм C4.5. Деревья регрессии. Алгоритм CART. Небрежные решающие деревья (oblivious decision tree). Решающий лес. Случайный лес (Random Forest).

Тема 8. Комбинаторно-логические методы, АВО. Представление о графических моделях (Байесовские сети)

Аппарат графических моделей (байесовские и марковские сети). Аппарат байесовского вывода. Некоторые методы дискретной оптимизации. Методы структурного обучения. Факторизация байесовских сетей. Потенциалы и энергия клик, связь с байесовскими сетями.

Аудиторные часы: 180

Формы текущего контроля и промежуточной аттестации: опросы, домашнее задание, контрольная работа, экзамен.

Основная литература:

1. Уэс Маккинли Python и анализ данных / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 482 с.: ил."
2. Райан Митчелл: Скрапинг веб-сайтов с помощью Python. Сбор данных из современного интернета

3. Spark для профессионалов: современные паттерны обработки больших данных. Риза С., Лезерсон У., Оуэн Ш., Уиллс Д.