АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ
«Машинное обучение»

Автор: Александров М. А.

Код и наименование направления подготовки, профиля: 38.04.02 Менеджмент («Финансы и Технологии»)
Квалификация (степень) выпускника: Магистр
Форма обучения: очная

Цель освоения дисциплины:
Сформировать способность применять методы алгоритмы машинного обучения.

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: 48

Формы текущего контроля и промежуточной аттестации: зачет

Основная литература:

1. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — 3-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2009. — 512 с.: ил. + CD-ROM — (Учебная литература для вузов) ISBN 978-5-9775-0368-6.

2. Тель Ж. Введение в распределенные алгоритмы. Пер. с англ. – М.: МЦНМО, 2009. – 616 с.
3. Kshemkalyani A. D., Singhal M. Distributed Computing: Principles, Algorithms, and Systems. Cambridge University Press, 2008.
4. Таненбаум Э. и др. Распределенные системы. Принципы и парадигмы. – СПб.: Питер, 2003.
5. Bollobas B. Modern Graph Theory. – Corrected ed. – Springer, 2013. – 394 p.
6. Handbook of Graph Theory. Edited by Gross J.L., Yellen J., Zhang P. – 2th ed. – CRC Press, 2014. – 1632 p.
7. Handbook of Graph Drawing and Visualization. Edited by Tamassia R. – CRC Press, 2013. – 862 p.
8. Барсегян А. А. и др. Анализ данных и процессов: учеб. пособие. 3-е изд. – 2009.
9. Бизнес-аналитика. От данных к знаниям (+ CD-ROM). Авторы Николай Паклин, Вячеслав Орешков
10. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
11. Натан Марц, Джеймс Уоррен. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени.
12. Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман. Анализ больших наборов данных
13. «Управление мастер-данными». Алекс Берсон, Лоуренс Дубов
14. Gifkins, Mike; Hitchcock, David (1988). The EDI handbook. London: Blenheim Online.
15. Эдвард Тафти. Визуальное представление больших объемов информации.
16. «Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами» Нейтан Яу, «Манн, Иванов и Фербер», 2013 г."
17. Корпоративные хранилища данных. Планирование, разработка и реализация. Эрик Спирли.
18. Интеграция хранилищ данных с открытыми и большими данными для решения задач финансовой организации: проблемы и подходы к решению