

Вопросы к государственному экзамену

Магистерская программа «Интеллектуальный анализ больших данных»

1. Линейная регрессионная модель. Метод наименьших квадратов. Основные свойства оценок метода наименьших квадратов.
2. Проверка линейных гипотез в рамках классической модели регрессии.
3. Дискриминантный анализ: постановка задачи и ее решение в случае известных параметров.
4. Кластерный анализ: постановка задачи и основные понятия.
5. Спектральный анализ временных рядов. Периодограмма выборочная. Свойства периодограммы. Сглаженная периодограмма.
6. Линейная ARMA(p,q) модель временных рядов. Стационарность, обратимость. Автокорреляционная и частная автокорреляционная функции. Оценивание параметров. Прогнозирование.
7. Общие принципы выбора страховых тарифов и тарифных ставок.
8. Асимптотическая формула для страховых тарифов в статической модели страхования: сравнение классической и факторизационной модели.
9. Гарантированные оценки страховых тарифов для факторизационной модели страхования при пуассоновском объеме портфеля.
10. Преобразование Фурье и его свойства. Линейные фильтры. Теорема о свертке.
11. Кратномасштабный анализ. Алгоритмы разложения и реконструкции.
12. Пороговая обработка коэффициентов вейвлет-разложения. Методы выбора порога.
13. Модель вычислений MapReduce. Пример реализации (Google MapReduce либо Hadoop MapReduce). Стадии вычислений, основные компоненты, понятие локальности данных.
14. Модель обработки данных в Apache Spark. Понятие resilient distributed dataset (RDD), примеры преобразований и действий над RDD. Граф вычислений.
15. Распределенная файловая система. Пример распределенной файловой системы (Google File System либо HDFS). Компоненты, основные операции, сценарии чтения и записи данных, механизмы отказоустойчивости.
16. Способы объектно-реляционного отображения для классов и атрибутов, бинарных и N-арных ассоциаций, классов ассоциаций, иерархий наследования. Примеры применения этих способов. Моделирование схемы реляционной базы данных с помощью диаграммы классов.
17. Образцы (паттерны) проектирования, их классификация и способ описания. Примеры образцов: структурного, поведенческого и порождающего.
18. Модели прогнозирования на основе деревьев решений. Алгоритмы CHAID, CART, C4.5: критерии поиска разбиений, параметры ограничения роста и обрубания дерева.
19. Нейронные сети прямого распространения. Архитектуры MLP и RBF: структура сетей, виды функций активации, алгоритмы обучения, борьба с переобучением и с проблемой локальных минимумов.
20. Поиск ассоциативных правил. Алгоритмы Apriori и FP-tree: построение частых эпизодов с ограничением по поддержке и формирование правил с ограничением по достоверности.
21. Векторная модель информационного поиска. Показатели tf и idf. Вычисление близости между запросом и документом. Применение метода "Наивный Байес" для автоматической классификации текстов.

22. Алгоритм ссылочного ранжирования PageRank: способ вычисления, роль в информационном поиске.
23. Шаг обработки данных (data step) в языке программирования SAS Base. Общая схема функционирования, понятие PDV, опции и ограничения наборов данных (where, keep, drop, rename, end). Использование by-группировки и retain переменных для расчета агрегаций.
24. Применение процедуры SAS/STAT GLM для решения задач дисперсионного и регрессионного анализа: синтаксис, основные параметры. Отбор и группировка категориальных предикторов. Интерпретация результатов: основные статистики и графики.
25. Применение процедуры SAS/STAT LOGISTIC для решения задач бинарной и многоклассовой классификации: синтаксис, основные параметры. Пошаговые алгоритмы отбора переменных. Интерпретация результатов: основные статистики и графики.
26. Преобразование изображения на основе эквализации гистограммы яркости.
27. Свёртка функций. Одномерная и двумерная свёртка и её свойства. Дискретная свёртка изображений. Обработка края изображения при свёртке.
28. Последовательная и параллельная сложность алгоритмов, информационный граф и ресурс параллелизма алгоритмов.
29. Архитектурные особенности графических процессоров, направленные на массивно-параллельные вычисления.

Список рекомендованной литературы

1. Андерсон Т. Введение в многомерный статистический анализ. – М.: Наука, 1963.
2. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980.
3. Hardle W., Simar L. Applied Multivariate Statistical Analysis. – Springer-Verlag, 2003.
4. Johnson R.A. and Winchurn D.W. Applied Multivariate Statistical Analysis. – Pearson Prentice Hall, 2007.
5. Факторный, дискриминантный и кластерный анализы. – М.: Финансы и статистика, 1989.
6. Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.
7. Дженкинс Г. Ваттс Д. Спектральный анализ и его приложения. М. Мир, 1972.
8. Бокс Дж., Дженкинс Г. Анализ временных рядов прогноз и управление. М. Мир, 1974.
9. Б.В. Гнеденко, Э.А. Даниелян, Б.Н. Димитров, Г.П. Климов, В.Ф. Матвеев, Приоритетные системы обслуживания. Издательство Московского университета, 1973.
10. В.Ф. Матвеев, В.Г. Ушаков Системы массового обслуживания. Издательство Московского университета, 1984.
11. Захарова Т.В., Шестаков О.В. Вейвлет-анализ и его приложения. 2-е изд., перераб. и доп. Учебное пособие. М.: ИНФРА-М. 2012.
12. Чак Лэм. Надоор в действии. – М: ДМК Пресс. – 2012, 424с.
13. Х. Карау, Э. Конвински, П. Венделл, М. Захария. Изучаем Spark. Молниеносный анализ данных. – М: ДМК Пресс. – 2015, 304с.
14. Рамбо Дж., Блаха М. UML 2.0. Объектно-ориентированное моделирование и разработка. – СПб.: Питер. – 2007. Главы 3 и 19.
15. Гамма Э. и др. Приемы объектно-ориентированного проектирования. Паттерны проектирования.: Пер. с англ. – СПб.: Питер, 2016.

16. Гонсалес Р., Вудс Р. Цифровая обработка изображений.: Пер. с англ. – М.: Техносфера, 2006. – 1070 с.
17. Шапиро Л., Стокман Дж. Компьютерное зрение: Пер. с англ. – М.: БИНОМ. Лаборатория знаний, 2006. – 752 с.
18. Hastie, Tibshirani and Friedman. The Elements of Statistical Learning. – Springer-Verlag, 2009. – 763 pages. <http://statweb.stanford.edu/~tibs/ElemStatLearn/>
19. Саймон Хайкин. Нейронные сети. Полный курс. 2-е изд., испр.: Пер. с англ. – М.: ООО «И. Д. Вильямс», 2006. – 1104 с.
20. Lora D. Delwiche, Susan J. Slaughter. The Little SAS® Book: A Primer, Fifth Edition. - SAS Institute, 2012. - 331 pages.
21. SAS/STAT User's Guide.
<http://support.sas.com/documentation/onlinedoc/stat/index.html>
22. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
23. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011
24. Ингерсолл Г., Мортон Т. Обработка неструктурированных текстов. Поиск, организация и манипулирование. ДМК-Пресс, 2015.