

Министерство образования и науки Российской Федерации
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

И.А. ЦИЛЬКОВСКИЙ,
В.М. ВОЛКОВА

МЕТОДЫ АНАЛИЗА ЗНАНИЙ И ДАННЫХ

Утверждено Редакционно-издательским советом университета
в качестве конспекта лекций

НОВОСИБИРСК
2010

УДК 519.23(075.8)

Ц 609

Рецензенты:

д-р техн. наук, проф. *А.А. Попов*, канд. техн. наук, доц. *Е.В. Чимитова*

Работа подготовлена на кафедре программных систем и баз данных
и предназначена для студентов V курса ФГМИ

Цильковский И.А.

Ц 609 Методы анализа знаний и данных : конспект лекций / И.А. Цильковский, В.М. Волкова. – Новосибирск : Изд-во НГТУ, 2010. – 68 с.

ISBN 978-57782-1377-7

Приведен системный подход к определению типа и шкал переменных, произведена классификация методов по целям решаемых задач и типам используемых переменных. Рассмотрены различные подходы к решению задачи распознавания образов и классификации наблюдений. Среди них представлены как классические (дискриминантный анализ, логистическая регрессия, кластерный анализ), так и относящиеся к области Data Mining исследования данных (деревья решений и нейронные сети). Кроме того, изложены важные в практических приложениях методы факторного анализа, которые могут использоваться как для снижения размерности, так и для структуризации множества исходных переменных. Рассмотрены возможности прогнозирования для различных типов зависимой переменной с использованием целого спектра методов. Проведен их сравнительный анализ, рассмотрены особенности, указаны способы верификации качества полученных результатов.

Цильковский Игорь Александрович, Волкова Виктория Михайловна

МЕТОДЫ АНАЛИЗА ЗНАНИЙ И ДАННЫХ

Конспект лекций

Редактор *И.Л. Кескевич*

Выпускающий редактор *И.П. Брованова*

Корректор *И.Е. Семенова*

Дизайн обложки *А.В. Ладыжская*

Компьютерная верстка *В.Н. Зенина*

Подписано в печать 14.04.2010. Формат 60×84 1/16. Бумага офсетная. Тираж 120 экз.
Уч.-изд. л. 3.95. Печ. л. 4,25. Изд. № 358. Заказ № Цена договорная

Отпечатано в типографии
Новосибирского государственного технического университета
630092, г. Новосибирск, пр. К. Маркса, 20

УДК 519.23(075.8)

ISBN 978-57782-1377-7

© Цильковский И.А., Волкова В.М., 2010

© Новосибирский государственный
технический университет, 2010

ОГЛАВЛЕНИЕ

1. МЕТОДЫ АНАЛИЗА ДАННЫХ (DATA MINING)	5
1.1. Определение понятия Data Mining.....	5
1.2. Элементы теории измерений	5
1.3. Общая структура задач.....	8
1.4. Основные трудности при проведении Data Mining	11
2. ДИСКРИМИНАНТНЫЙ АНАЛИЗ.....	12
2.1. Введение	12
2.2. Формулирование проблемы.....	13
2.3. Каноническая дискриминантная функция.....	14
2.4. Коэффициенты канонической дискриминантной функции	15
2.5. Нестандартизованные и стандартизованные коэффициенты	18
2.6. Число дискриминантных функций.....	19
2.7. Определение значимости дискриминантной функции.....	19
2.8. Интерпретация результатов	20
3. КЛАСТЕРНЫЙ АНАЛИЗ	21
3.1. Общая характеристика методов кластерного анализа.....	21
3.2. Выбор переменных. Преобразование данных.....	23
3.3. Меры сходства	23
3.4. Иерархический кластерный анализ.....	25
3.5. Метод k -средних	26
3.6. Метод поиска сгущений.....	28
3.7. Критерии качества классификации	29
4. ФАКТОРНЫЙ АНАЛИЗ	31
4.1. Введение	31
4.2. Метод главных компонент.....	32
4.3. Другие методы выделения факторов	34
4.4. Результаты применения методов выделения факторов.....	35
4.5. Методы вращения	37

4.6. Критерии определения числа факторов.....	39
4.7. Критерий качества модели (подтверждающий факторный анализ)	40
4.8. Интерпретация факторов.	41
5. ДЕРЕВЬЯ РЕШЕНИЙ	42
5.1. Определение и свойства деревьев решений	42
5.2. Построение деревьев решений	44
6. НЕЙРОННЫЕ СЕТИ	51
6.1. Введение	51
6.2. Понятия и свойства нейронных сетей.....	51
6.3. Формальный нейрон.....	53
6.4. Виды функций активации	54
6.5. Ограничения модели нейрона	56
6.6. Однослойные искусственные нейронные сети	56
6.7. Многослойные искусственные нейронные сети	57
6.8. Нелинейная активационная функция. Многослойный перцептрон	58
6.9. Обучение искусственных нейронных сетей.....	59
6.10. Выбор количества нейронов и слоев	60
6.11. Линейный алгоритм обучения многослойного перцептрона с учителем	61
6.12. Сеть Кохонена.....	62
6.13. Сети с обратными связями.....	62
7. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ.....	63
7.1. Введение	63
7.2. Модели бинарного выбора.....	63
7.3. Порядковые зависимые переменные. Ранжированный выбор.	65
7.4. Номинальные зависимые переменные. Неранжированный выбор	66
Библиографический список.....	68