Machine Learning and Data Mining

Игнатов Дмитрий Игоревич

Национальный исследовательский университет Высшая школа экономики
Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

2016

План

План лекции

Примерная программа курса

1. Введение ✓
2. Кластеризация ✓
3. Классификация ✓
4. Частые множества признаков (frequent itemsets) и ассоциативные правила ✓
5. Рекомендательные системы и алгоритмы ✓
6. Анализ формальных понятий и его приложения. Мультимодальная кластеризация ✓
7. Машины опорных векторов (SVM) ✓
8. Регрессия и регуляризация ✓
9. Тематическое моделирование и EM-алгоритм*
10. Ансамблевые методы кластеризации*
11. Ансамблевые методы классификации*
12. Нейронные сети и генетические алгоритмы*
13. Отбор признаков. Снижение размерности. Семплирование. Аномалии в данных.*
14. Технологии и методы работы с Big Data*
15. Статистический взгляд на машинное обучение*

Итоговая оценка

Сценарий 1

Домашние задания+зачет (экзамен)

Сценарий 2

Домашние задания + проект (индивидуальный или групповой)+зачет (экзамен)

План лекции

О терминологии. KDD и Data Mining

Knowledge discovery in Databases (KDD)

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Fayyad, Piatetsky-Shapiro, and Smyth 1996

Data Mining

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

Там же

О терминологии. KDD и Data Mining

Схема процесса обнаружения знаний в данных

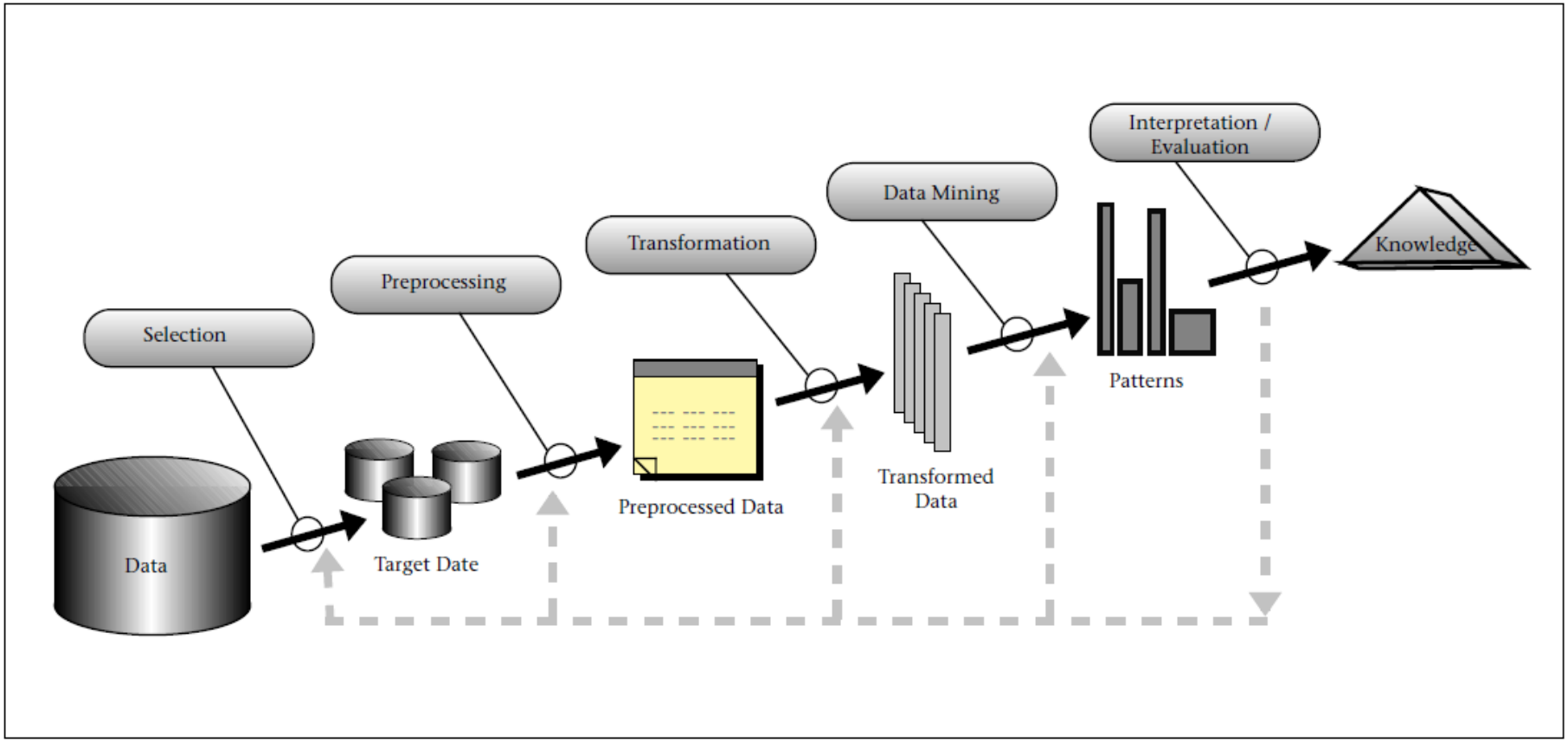


Figure 1. An Overview of the Steps That Compose the KDD Process.

(*Fayyad, Piatetsky-Shapiro, and Smyth 1996*)

О терминологии. KDD и Data Mining

[J. Han et al., Data Mining. Concepts and Techniques, 3rd Ed., 2012]

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation
7. Knowledge presentation

Data Mining

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

О терминологии. Машинное обучение

[T. Mitchell. The Discipline of Machine Learning,2006]

Основной вопрос в машинном обучении

How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

Более точно

To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

О межпредметных связях

Гипотеза

Data Mining $\stackrel{?}{=}$ Machine Learning

Связанные дисциплины

- Computer Science (Информатика)
- Artificial Intelligence (Искусственный интеллект)
- Pattern Recognition (Распознавание образов)
- Information Retrieval (Информационный поиск)
- Social Network Analysis (Анализ социальных сетей)
- Теория вероятностей и математическая статистика
- Дискретная математика (в т.ч. порядки и графы)
- Optimization (Методы оптимизации)

Области применения DM&ML

Области применения

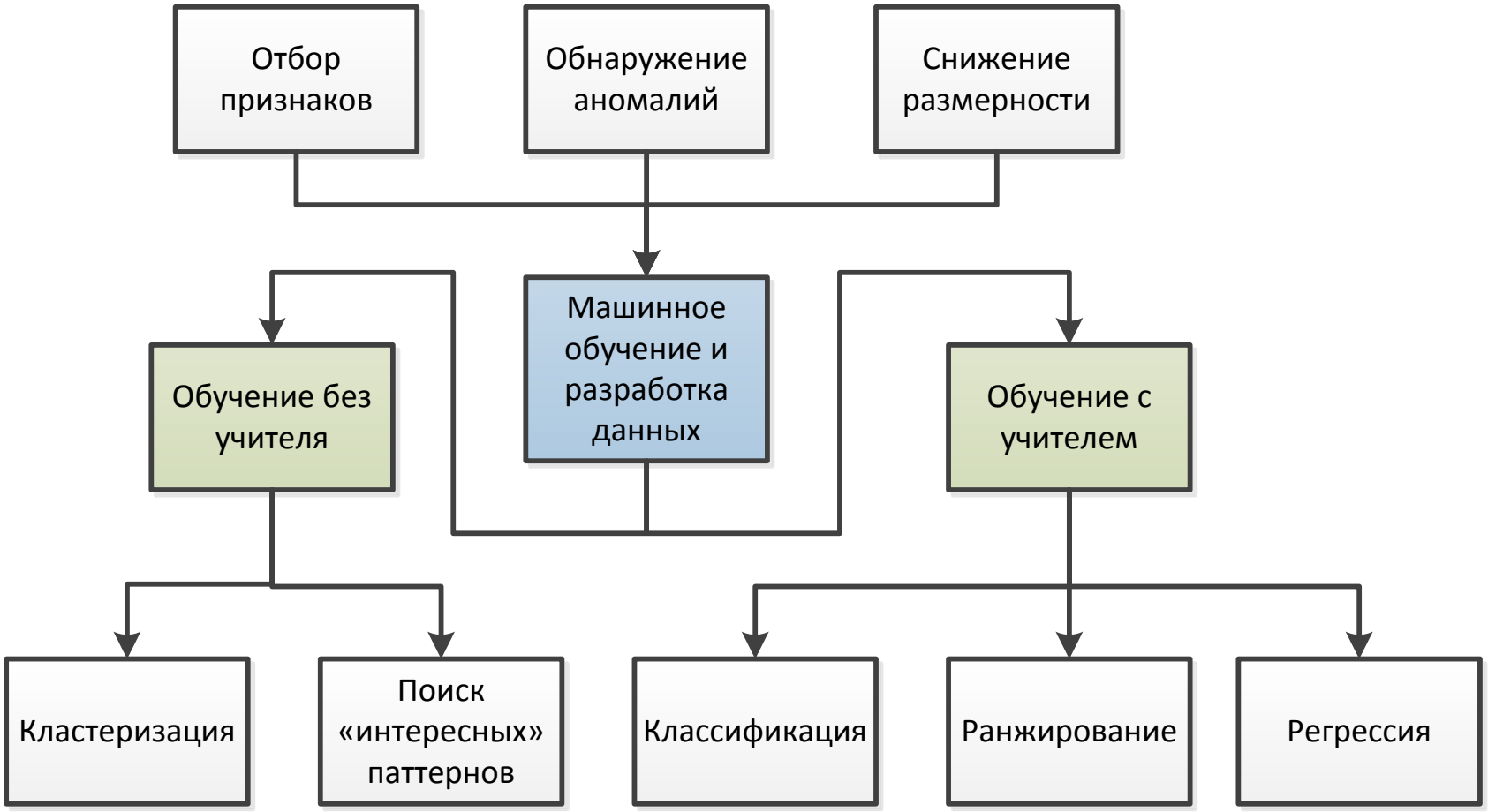
- Бизнес
- Медицина
- Образование
- Науки о жизни
- Интернет-данные
- Банковское дело и финансы
- ...

Тренды в областях применения DM&ML

[J. Han et al., 2012]

- Application exploration: e.g., counter-terrorism and mobile (wireless) data mining
- Scalable and interactive data mining methods
- Integration of data mining with search engines, database systems, data warehouse systems, and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving-objects, and cyber-physical system
- Mining multimedia, text, and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

Таксономия методов DM&ML



Кластеризация

Постановка задачи

- Найти разбиение исходного множества объектов на группы (кластеры).
- Объекты внутри одного кластера обладают высоким сходством.
- Объекты из разных кластеров сильно различаются.

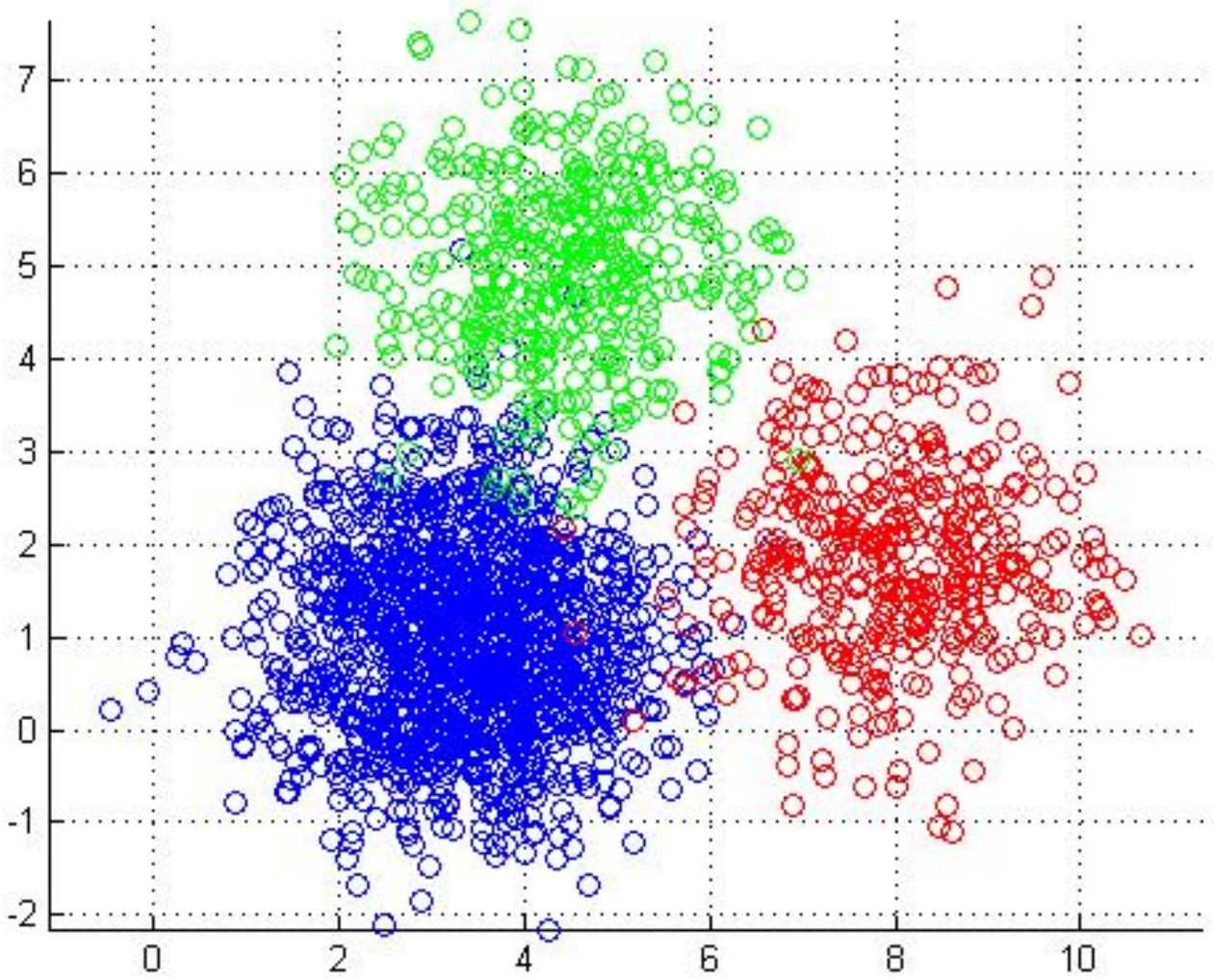
Кластеризация

Методы кластеризации

- Метод k-средних
- Иерархическая кластеризация (агломеративный и дивизимный подходы)
- Спектральная кластеризация
- Мультимодальная кластеризация: бикластеризация и трикластеризация.

Кластеризация

Метод k-средних



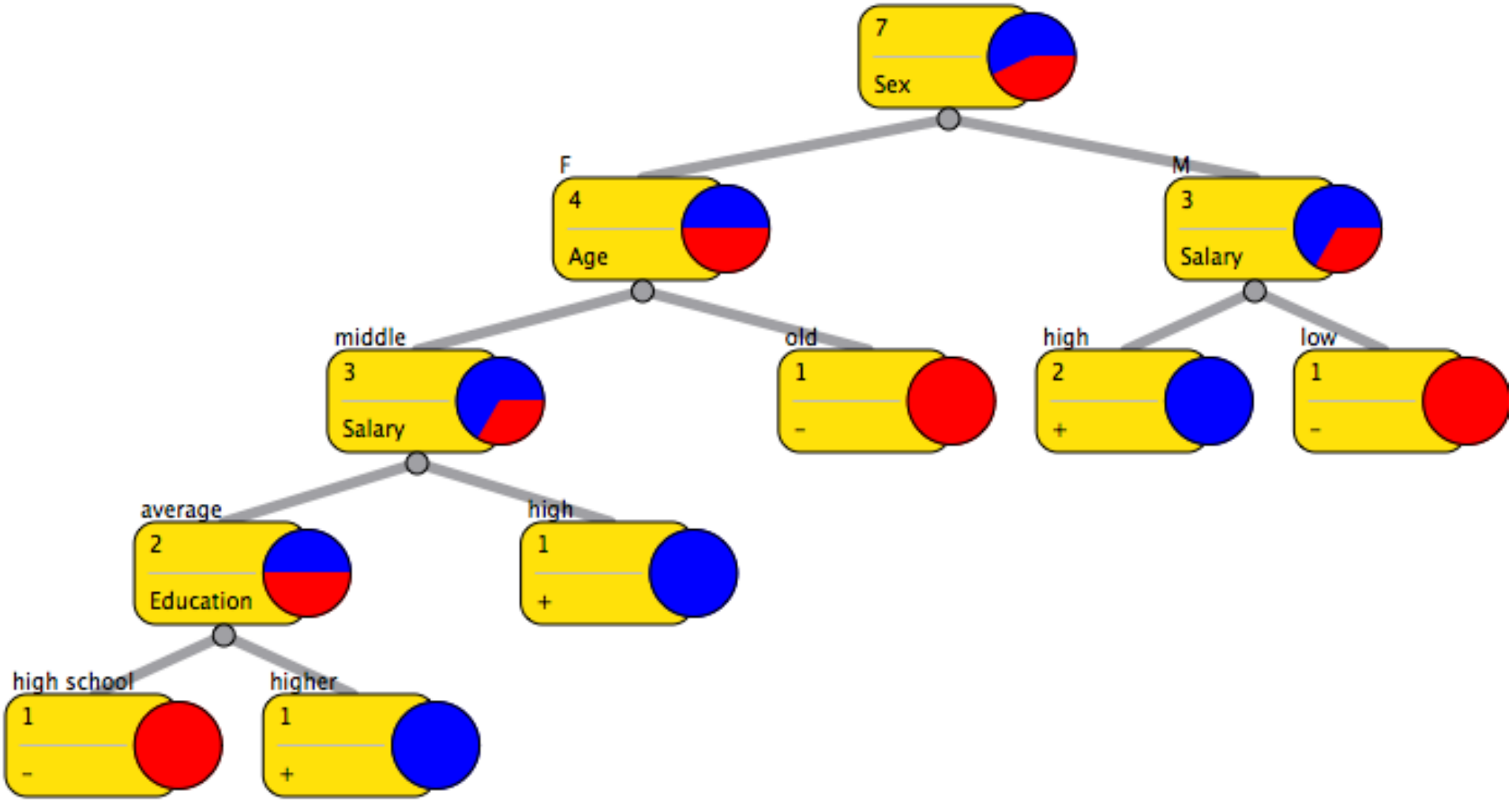
Классификация

Постановка задачи

- По описанию объектов некоторого множества с известными метками классов определить класс объектов той же природы (в том же признаковом пространстве) с неизвестными метками.

Классификация

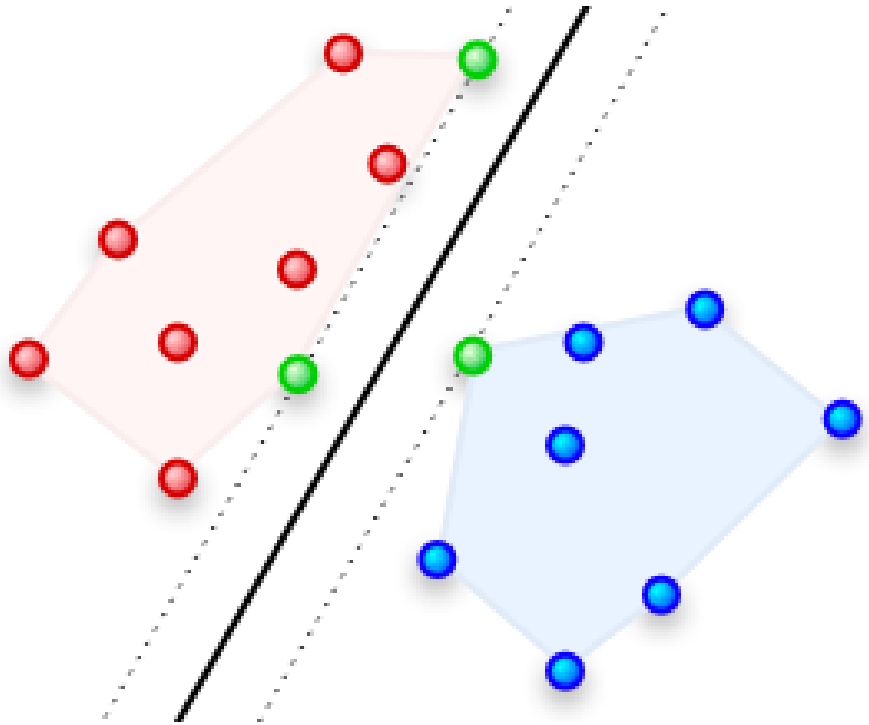
Деревья решений в оценке кредитного риска



- Алгоритм 1-Rule
- kNN классификатор (k ближайших соседей)
- Наивный баейсовский классификатор (Naïve Bayes classifier)
- Деревья решений (decision trees)
- Машины опорных векторов (Support Vector Machines (SVM))
- ДСМ-метод (в честь Джона Стюарта Милля)

Классификация

Машины опорных векторов (SVM)



Регрессия

- Линейная парная и множественная регрессия (Эконометрика и математическая статистика)
- Лассо-регуляризация. Логистическая регрессия как метод классификации. (Этот курс)

Поиск паттернов/зависимостей

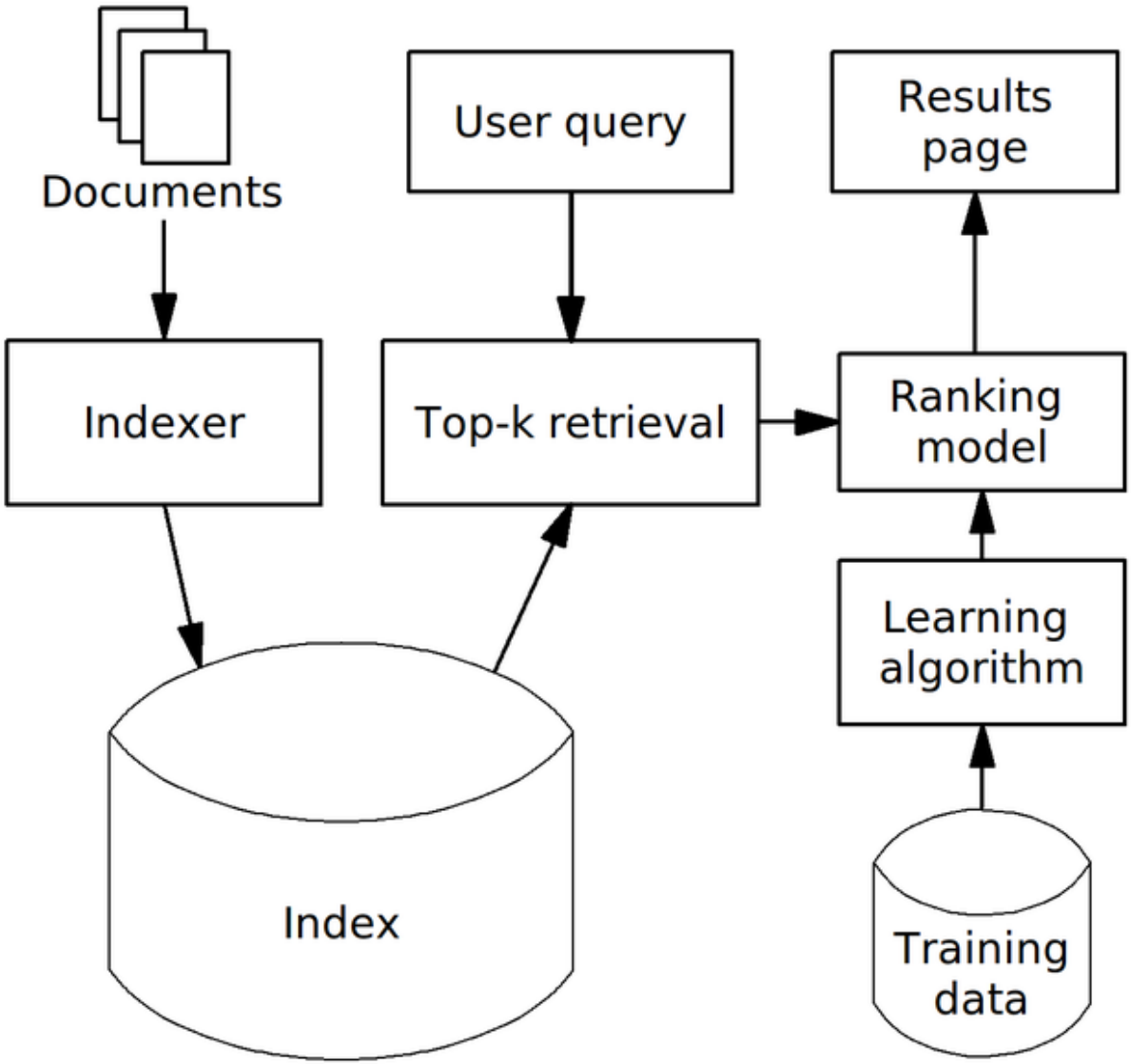
Постановка задачи

- Поиск закономерностей в данных об использовании каких-либо ресурсов. Например, часто используемых вместе ресурсов.
- Пример. $support(\{\text{хлеб, молоко}\}) = 0.7$
- Часто такие закономерности записываются в виде правил $A \longrightarrow B$
- Пример. $\{\text{Студент, Возраст от 16 до 25}\} \longrightarrow \{iPhone, iPad\}$

Поиск паттернов/зависимостей

The FIMI'03 best implementation award was granted to Gosta Grahne and Jianfei Zhu (on the left). The award consisted of the most frequent itemset: $\{diapers, beer\}$.

Ранжирование



Рекомендательные системы

http://Amazon.com

Frequently Bought Together

Price For All Three: $86.01

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

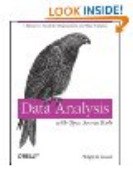
- **This item:** Machine Learning for Hackers by Drew Conway Paperback $33.87
- Machine Learning in Action by Peter Harrington Paperback $25.75
- Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback $26.39

Customers Who Bought This Item Also Bought

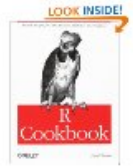
Programming Collective Intelligence: Building ...
› Toby Segaran
(84)
Paperback
$26.39

Machine Learning in Action
› Peter Harrington
(10)
Paperback
$25.75

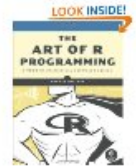
Mining the Social Web: Analyzing Data from ...
› Matthew A. Russell
(19)
Paperback
$26.36



Data Analysis with Open Source Tools
› Philipp K. Janert
(29)
Paperback
$24.05



R Cookbook (O'Reilly Cookbooks)
› Paul Teetor
(18)
Paperback
$32.43



The Art of R Programming: Tour of Statistical ...
Norman Matloff
(29)
Paperback
$25.06

Are any of these items inappropriate for this page? Let us know

Рекомендательные системы

http://Imhonet.ru

Оценки фильма Любопытное стечение обстоятельств

Een Bizarre Samenloop Van Omstandigheden, A Curious Conjunction of Coincidences

Фильмы / Комедии / обстоятельств» /

Да, Вам стоит смотреть фильм «Любопытное стечение обстоятельств»

Людям, с оценками, похожими на Ваши, этот фильм нравится

А ещё они рекомендуют Вам 31 фильм

Ваша прогнозируемая оценка фильма после его просмотра
8.2

Смотрели? Оцените

Не рекомендовать

Про фильм | Онлайн | Скачать | Отзывы | Персоны | Кадры | Оценки | Похожие

Распределение оценок

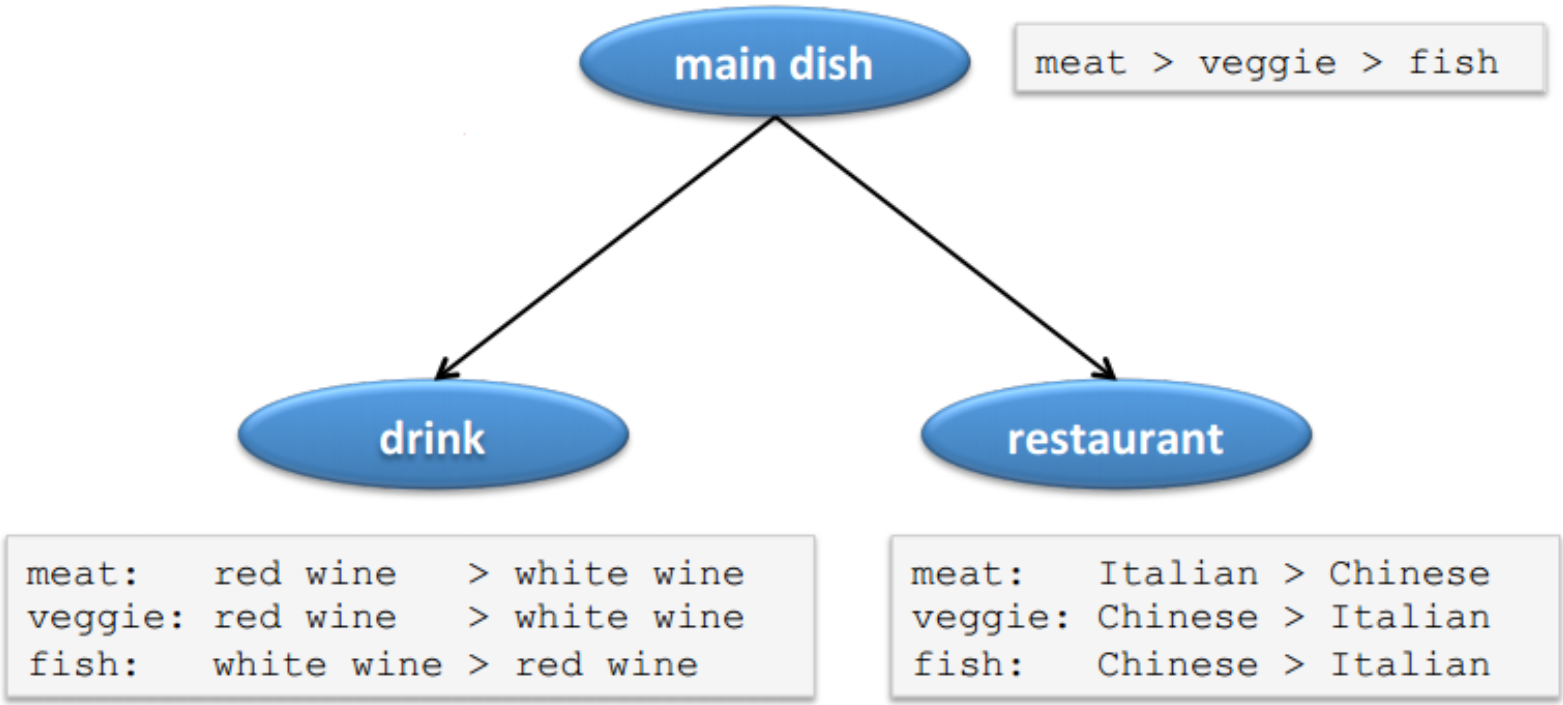
1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	1	4	4

Кому больше нравится

Кому нравится: 99% 1%

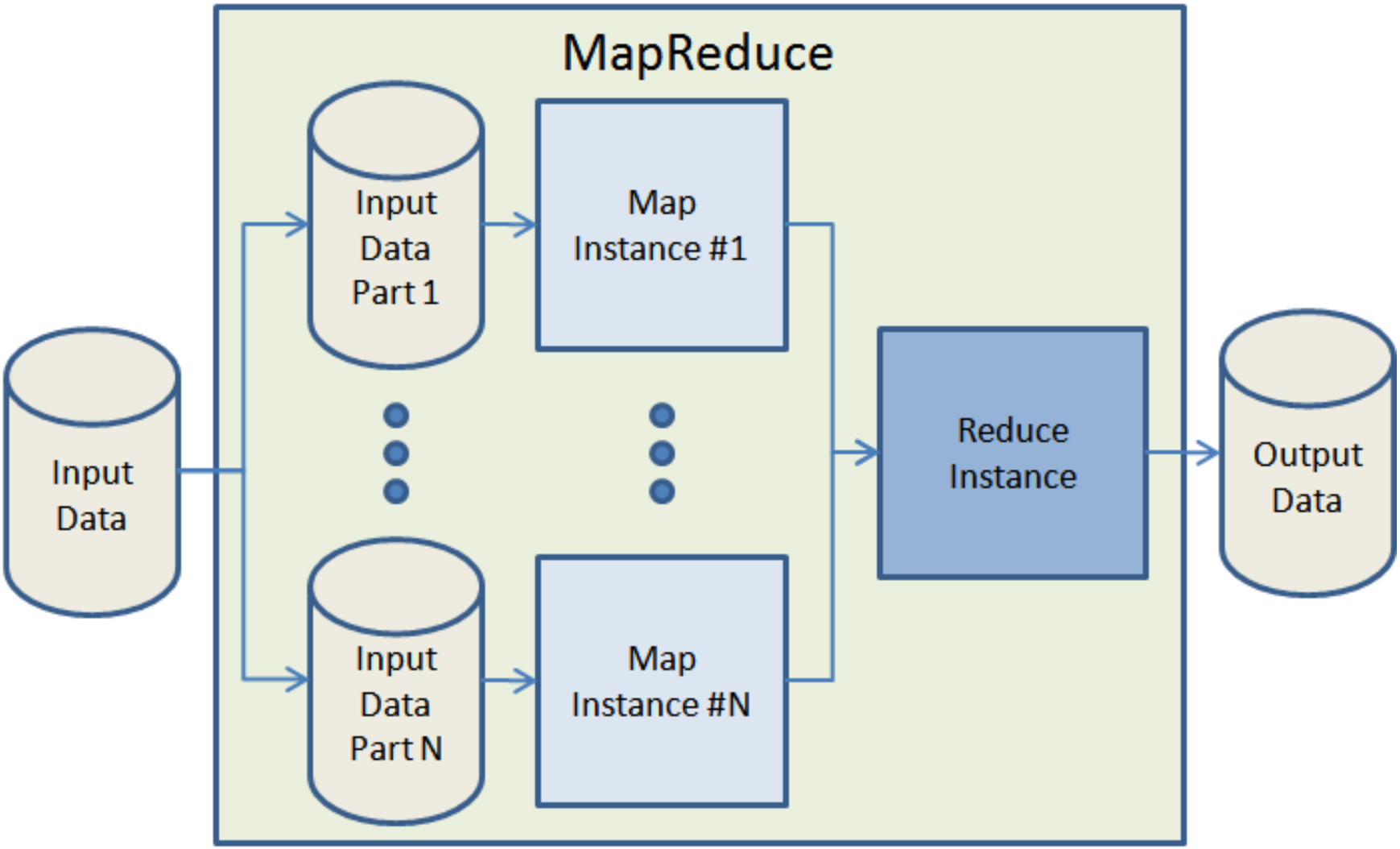
Обучение предпочтениям

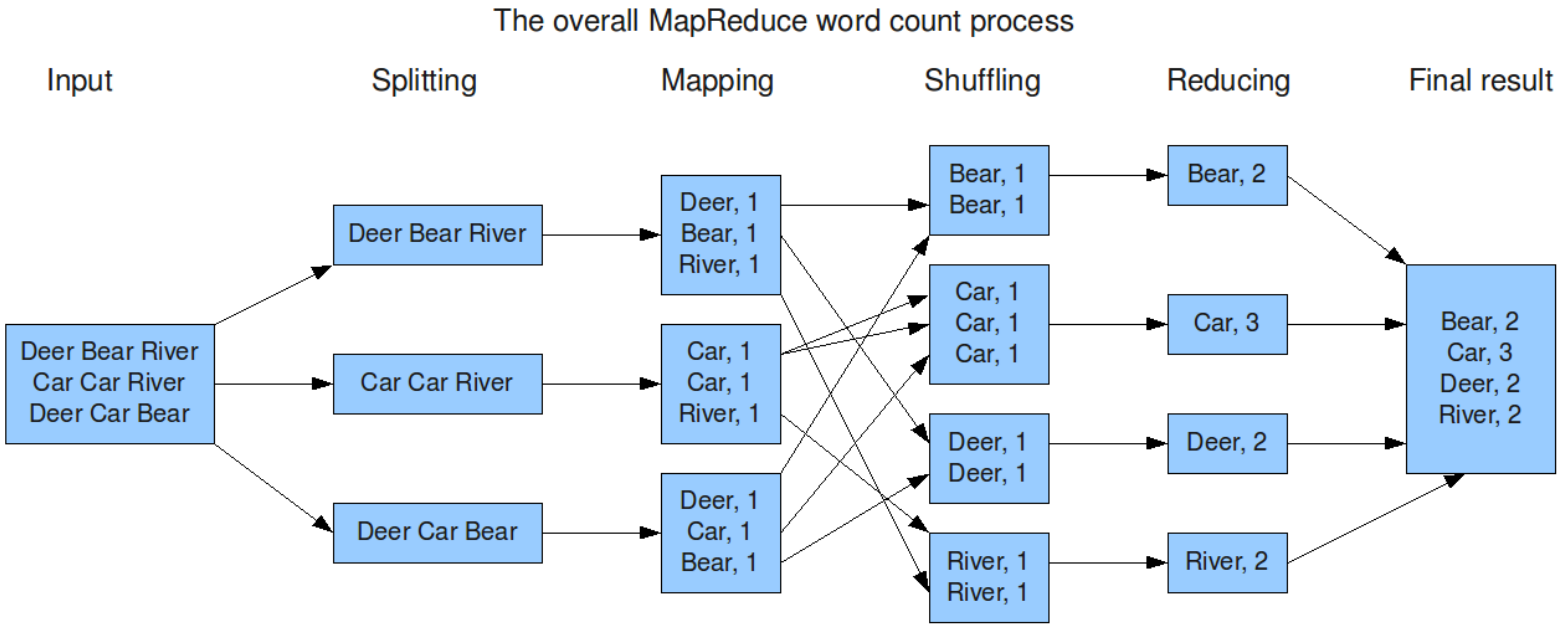
http://www.preference-learning.org/



Big Data

Технология MapReduce




The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Что такое Apache Mahout?

Apache Mahout^{TM} – библиотека масштабируемых методов машинного обучения в основном по технологии MapReduce.

Что такое Apache Spark?

- "Apache Spark™ is a fast and general engine for large-scale data processing."
- Включает библиотеку методов машинного обучения MLlib.
- Работает как с Hadoop, так и без.

План лекции

Системы машинного обучения и анализа данных

1. Orange (freely available)
2. Weka (freely available)
3. Knime (community edition for free)
4. RapidMiner (community edition for free)
5. Deductor (бесплатная версия для обучения)
6. QuDA (freely available)

Библиотеки машинного обучения и анализа данных

1. scikit-learn (freely available Machine Learning in Python)
2. MALLET — MAchine Learning for LanguagE Toolkit (freely available)
3. Accord.NET Framework (.NET machine learning framework combined with audio and image processing libraries completely written in C#)
4. Infer.NET (framework for running Bayesian inference in graphical models)
5. R (free software environment for statistical computing and graphics+many packages for ML&DM)

Стандарты в ML&DM

http://www.dmg.org

PMML

Язык разметки для прогнозного моделирования (Predictive Model Markup Language — PMML) разработан Data Mining Group (DMG) на основе XML, обеспечивает приложениям способ определения моделей машинного обучения и Data Mining, а также обмен такими моделями между PMML-совместимыми приложениями.

План лекции

Книги

- P. Flach Machine Learning: The Art and Science of Algorithms that Make Sense of Data, 2012
- M. Zaki et al. *Data Mining and Analysis: Fundamental Concepts and Algorithms*, 2014 (free)
- J. Leskovec et al. Mining of Massive Datasets, 2014 (free)
- C.M. Bishop *Pattern Recognition and Machine Learning*, 2006
- D. Barber Bayesian Reasoning and Machine Learning, 2012 (free)
- K.P. Murphy Machine Learning: a Probabilistic Perspective, 2012
- T. Hastie et al. *Elements of Statistical Learning*, 2009 (free)
- G. James et al. *An Introduction to Statistical Learning with Applications in R*, 2013 (free)
- J. Han et al. *Data Mining. Concepts and Techniques*, 2012
- Т. Митчелл *Machine Learning*, 1997
- Т. Сегаран *Программируем коллективный разум*, 2007 (на английском)
- Барсегян А. и др. *Анализ данных и процессов*, 2009

Портал machinelearning.ru

http://machinelearning.ru

- Лекции К.В. Воронцова. *Математические методы обучения по прецедентам (машинное обучение)*
- Лекции Д.П. Ветрова, Д.А. Кропотова Байесовские методы машинного обучения, 2014
- Учебник А.Г. Дьяконова. *Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab*, 2010

Лекции и книга С. Николенко

http://logic.pdmi.ras.ru/~sergey/

- Игрок Что?Где?Когда?
- С.Николенко, А. Тулупьев. *Самообучающиеся системы* 2009

Coursera: курсы и специализации

http://www.coursera.org/

- Andrew Ng. *Machine Learning*
- Jiawei Han Pattern Discovery in Data Mining
- Jure Leskovec et al. Mining Massive Datasets
- Hastie & Tibshirani Statistical Learning

Специализации (платные сертификаты) — состоят из отдельных курсов (участие бесплатно)

- Data Mining
- Data Science

Deep Learning (Глубинное обучение или глубокое обучение)

- Deep Learning by Udacity
- Deep Learning Course by NVIDIA
- Geoffrey Hinton. *Neural Networks for Machine Learning* (2012)

ИНТУИТ

http://intuit.ru

- Интернет-университет информационных технологий
- К.В. Воронцов Машинное обучение, 2015 (Видео к курсу на сайте ШАД)
- И.А. Чубукова. Data Mining, 2006

Сообщество

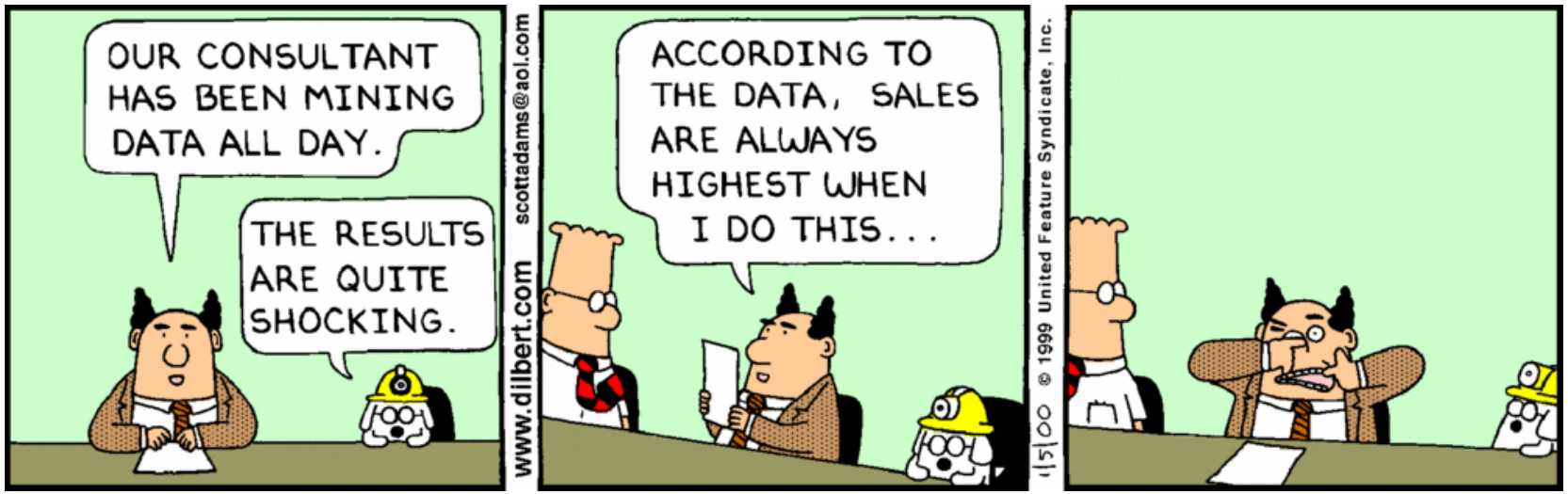
- IMLS – The International Machine Learning Society
- Kaggle – платформа для соревнований по анализу данных
- KDD Nuggets – Data Mining Community Top Resource
- Open ML – Machine Learning community portal
- UCI Machine Learning Repository – Репозиторий данных

Конференции

- ICML – International Conference on Machine Learning
- IEEE ICDM – IEEE International Conference on Data Mining
- KDD – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- ECML & PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- NIPS – Neural Information Processing Systems
- RecSys – The ACM conference series on Recommender Systems
- ИОИ & ММРО – Серия конференций «Интеллектуализация обработки информации»/«Математические методы распознавания образов»
- АИСТ – International conference on Analysis of Images, Social Networks, and Texts

Just for fun или шутки ради

http://dilbert.com



Вопросы и контакты

www.hse.ru/staff/dima

Спасибо!

dmitrii.ignatov[at]gmail.com