Введение в машинное обучение

Сергей Лисицын

lisitsyn.s.o@gmail.com

15 марта 2011 г.

План курса

1. Введение в машинное обучение, 15 марта 2011
2. Классификация, 22 марта 2011
3. Классификация, 29 марта 2011
4. Кластеризация, 5 апреля 2011
5. Поиск ассоциативных правил, 12 апреля 2011

Развитие курса

1. Выбор и синтез признаков, метод главных компонент (PCA)
2. Стимулируемое обучение (reinforcement learning)
3. Активное и аналитическое обучение (active learning, analytical learning)
4. Вычислительная теория обучения (COLT)
5. Параллельные алгоритмы машинного обучения, анализ больших объемов данных

Содержание курса

- Описание некоторых хорошоизученных задач машинного обучения
- Примеры решения реальных задач
- Реализации большинства методов
- Немного анализа эффективности, применимости и вычислительной сложности алгоритмов

Задачи этого курса

- Познакомить слушателя с машинным обучением и задачами, им решаемыми
- Показать применимость конкретных методов к реальным задачам
- Убедить, что теорию вероятности и линейную алгебру можно применить где-то кроме университета :-)

Реализации расмотренных в курсе методов

- Для большинства рассмотренных алгоритмов приложены примеры их реализации на языке Python
- Выбор языка обусловлен его простотой, распространённостью и не «эзотеричностью»
- Знание Python поможет, но не необходимо
- Сами алгоритмы будут описаны в виде последовательности действий на естественном языке

Немного об истории искусственного интеллекта

- Искусственный интеллект как научная дисциплина (или проект) развивается уже достаточно долгое время
- Наибольшая активность по попыткам создания искусственного интеллекта наблюдалась в 60-х и 70-х годах прошлого века
- «В течении десяти лет компьютер сможет стать чемпионом мира по шахматам», Ньюэлл, 1958
- «В течении 3-8 лет мы будем иметь машины с такими же умственными способностями, как у человека», Мински, 1970
- На проекты искусственного интеллекта выделялись огромные деньги
- Но всё оказалось не так радужно, проекты искусственного интеллекта стали «буксовать»

Проблемы искусственного интеллекта

- В 1973 Лайтхилл опубликовал отчёт, предрекший снижение финансирования и вообще интереса к ИИ
- Проблема комбинаторного взрыва и низкая производительность компьютеров
- Парадокс Моравеца: для машины многие простые задачи становятся сложными и наоборот
- Проблемы представления знаний «здравого смысла» (commonsense knowledge)
- Как результат – т.н. «зима искусственного интеллекта» – снижение финансирования проектов по созданию искусственного интеллекта
- Имеющиеся достижения не могли просто пропасть – потеря интереса к ИИ привела к развитию смежных дисциплин, но уже не в рамках такого грандиозного проекта

Смежные и составные дисциплины AI

- Обработка естественных языков (natural language processing)
- Компьютерное зрения и распознавание образов (computer vision, pattern recognition)
- Инженерия знаний (knowledge engineering)
- Вывод фактов (inference)
- Машинное обучение (machine learning)
- ...

Машинное обучение

- Подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных к обобщению и обучению
- Тесно связанная с теорией вероятностей и статистикой, линейной алгеброй, методами оптимизации и некоторыми другими дисциплинами
- Возникло на основе ранних работ по созданию искусственного интеллекта как способ решения труднопрограммируемых задач
- Как и статистика имеет практическую основу и большая часть исследований проводится на реальных данных
- Теоретические исследования проводятся в рамках вычислительной теории обучения (computational learning theory, COLT)

Предпосылки к развитию машинного обучения

- Проблема создания AI не только технологично трудна, но трудноразрешима даже философски
- Вместо создания полноценного искусственного интеллекта можно реализовать одну из самых важных частей «интеллектуальности» – способность обучаться
- Некоторые области знаний уже имели наработки в способности обучаться – статистическая теория обучения распознавания образов
- Улучшение вычислительных способностей компьютеров и постановка трудноразрешимых напрямую задач

Ниши машинного обучения

- Интеллектуальный анализ данных:
  - Выделение из данных знаний
  - Выяснение структуры данных
  - ...
- Приложения, которые тяжело или невозможно запрограммировать
  - Распознавание
  - Автоматическое управление
  - ...
- Адаптивные приложения
  - Коллаборативная фильтрация интересов
  - Фильтрация спама
  - ...

Что такое обучаемый алгоритм?

Определение Артура Самуэля, 1959 (вольный перевод)

Алгоритм, который может обучаться без явного запрограммирования.

Определение Тома Митчелла, 1998

Алгоритм обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T улучшается (по мере P) с приобретением опыта E.

Открытые вопросы машинного обучения

- Какой алгоритм решает некоторую задачу наилучшим образом? Как создать такой алгоритм?
- Сколько и какой информации необходимо для обучения?
- Какие данные выбирать для обучения, как этот выбор влияет на качество обучения?
- Как свести какую-либо задачу обучения к аппроксимации или оптимизации какой-то функции?
- Практические вопросы выбора моделей (как данных, так и моделей алгоритмов)

Типы машинного обучения

- Обучение с учителем (supervised learning)
  явная обратная связь от «учителя»
- Частичное обучение (semi-supervised learning)
  частичная обратная связь от «учителя»
- Активное обучение (active learning)
  обратная связь по запросу
- Стимулируемое обучение (reinforcement learning)
  отложенная обратная связь
- Обучение без учителя (unsupervised learning)
  нет обратной связи

Обучение с учителем и частичное обучение

Supervised (semi-supervised) learning

- На подмножестве рассматриваемых объектов известны соответствующие объектам ответы (их знает «учитель»)
- Учителем как правило является обучающая выборка пар (объект – ответ)
- Задачей такого обучения является нахождение закономерности, позволяющее узнать ответ на любом требуемом объекте
- В случае с частичным обучением большая часть ответов не известна

Активное обучение
Active learning

- Очень схоже с обучением с учителем, но «ответы» изначально неизвестны
- Основная идея состоит в том, что алгоритм может обучаться на малых выборках, если он сам будет выбирать какие данные ему нужны
- Таким образом, алгоритм составляет запросы (query), ответы на которые помогают ему обучиться

Обучение с подкреплением
Reinforcement learning

- Не существует никаких «правильных» ответов
- Можно наблюдать ответную реакцию среды
- Ситуация достаточно близкая к реальной
- Как правило используется формализация марковским процессом принятия решений
- Алгоритм пытается найти стратегию, оптимальную для данного процесса принятия решения

Обучение без учителя
Unsupervised learning

- Работа с данными без сторонней информации: ищется не зависимости (объект-ответ), а связи между объектами
- Критерии качества обучения сильно разнятся в зависимости от задачи
- Большая часть методов анализа данных производится именно в рамках этого типа задач

Переобучение и недообучение

Overfitting, underfitting

- Переобучение (overfitting) – явление, при котором алгоритм слишком приспособлен для данных, на которых обучался
- Недообучение (underfitting) – обратное переобучению явление, при котором алгоритм не использует полностью предоставленные ему данные
- В некоторых задачах такая проблема почти не возникает (без учителя), в некоторых она стоит достаточно остро (с учителем)

Выбор модели

- Модель в обучении – класс алгоритмов решающих задачу
- Алгоритм должен решать задачи без переобучения и недообучения
- Для некоторого алгоритма переобучение это излишняя его сложность, а недообучение – простота
- Единого способа оценки оптимальной сложности нет: SRM (структурная минимизация риска, Вапник), MDL (минимизация длины описания, Риссанен), ...
- Кроме того, при решении задачи необходимо выбирать представление данных

Представление объектов в машинном обучении

- Большая часть объектов, изучаемых в машинном обучении представляется в виде векторов

$$x = (x_1, x_2, \ldots, x_n)$$

или матриц

$$x = \begin{pmatrix} x_{11} & \ldots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \ldots & x_{mn} \end{pmatrix}$$

компоненты которых называются признаками (feature), а пространство объектов иногда называется признаковым пространством (feature space) или пространством объектов
- В некоторых случаях, признаки объекта не вещественны
- Иногда конкретные значения признаков объектов неизвестны, а известны только взаимосвязи (например расстояния между объектами)

Классификация
Classification

- Обучение с учителем (supervised learning)
- Синоним: распознавание образов (pattern recognition)
- Имеется подмножество множества объектов $\mathcal{X}$, разбитое некоторым образом на классы из $\mathcal{C}$ – обучающая выборка (классами являются любые сущности, характеризующие группы объектов)
- Необходимо найти классы для всех возможных объектов
- Примеры классификации:
  - Данные о клиенте банка, классы: выгоден, не выгоден
  - Симптомы заболевания, классы: виды заболеваний
  - Изображения лиц, классы: личности
  - Данные зондирования почв, классы: наличие полезных ископаемых или отсутствие
  - Оптическое распознавание символов (OCR)
  - Документы, классы: спам, не спам (или тематика документа)
- Классификацию мы рассмотрим в лекциях 2 и 3

Регрессия
Regression estimation

- Обучение с учителем (supervised learning)
- Задача статистического обучения, или вообще статистики
- В подмножестве множестве объектов $\mathcal{X}$ каждому объекту сопоставлено число из $\mathbb{R}$ (получается классификация с множеством классов $\mathcal{C} = \mathbb{R}$)
- Необходимо найти «ответы» на всех возможных объектах
- Очень многие методы классификации несложно модифицируются под регрессию
- Примеры восстановления регрессии:
  - Данные о жилье, оценка стоимости жилья
  - Медицинские параметры, оценка длительности реабилитации
  - Данные о клиенте банка, оценка максимальной допустимой суммы кредита

Прогнозирование
Forecasting

- Обучение с учителем (supervised learning)
- Исходные данные задачи – некоторые значения функции во времени, то есть временной ряд (time series)
- Задачей является нахождение значений функции за пределами имеющихся данных
- Является самой популярной задачей анализа данных
- Может решаться методами регрессии и классификации
- Примеры прогнозирования
  - Значения цены или относительного курса во времени, значения в будущем на некотором «горизонте прогнозирования»
  - Данные о сейсмоактивности, время следующего землятресения

Кластеризация объектов
Clustering

- Обучение без учителя (unsupervised learning)
- В отличие от классификации, классы не известны, их нужно построить (но называться они будут уже кластерами)
- Кластеры понимаются как непересекающиеся структуры, обладающие компактностью, связностью и пространственным разделением
- Примеры классификации
  - Данные о личности, кластеры: характерные группы
  - Тексты, кластеры: группы текстов одинаковой тематики
  - Некоторые объекты, кластеры: типичные объекты и нетипичные объекты
- Кластеризацию объектов мы будем рассматривать в лекции 4

Поиск ассоциативных правил
Association rule learning

- Обучение без учителя (unsupervised learning)
- Среди множества составных объектов ищутся взаимосвязи
- Примеры поиска ассоциативных правил:
  - Покупки в супермаркетах, поиск зависимостей
  - Записи в логах сервера, поиск частых переходов
  - Некоторые данные о поведении программы или пользователя, нахождение признаков вторжения
- Поиск ассоциативных правил мы будем рассматривать в лекции 5

Сокращение размерности данных
Dimensionality reduction

- Обучение без учителя (unsupervised learning)
- Практическая задача, решение которой позволяет повысить эффективность обработки данных
- Большие размерности данных неудобны для визуализации и сложны для анализа
- Формально для пространства объектов $\mathbb{R}^m$ задача сводится к нахождению $r : \mathbb{R}^m \to \mathbb{R}^n, m > n$
- Два подхода:
  - Выбор информативных признаков (feature selection): отображение в пространство меньшей размерности отбрасыванием неинформативных признаков
  - Синтез признаков (feature extraction): отображение в пространство меньшей размерности с помощью некоторой векторной функции

Фильтрация выбросов
Outliers detection

- Обучение без учителя (unsupervised learning)
- Выбросы (outliers) – любые объекты выборки, существенно отличающиеся от остальных объектов, в ней содержащихся
- Алгоритм должен уметь определять, что является выбросом
- Примеры фильтрации:
  - Описания процессов, обнаружение мошенничества, нетипичного поведения
  - Зашумлённые выборки данных, очистка моделей данных от шумов

Ранжирование
Learning to rank

- Обучение без учителя (unsupervised learning) или частичное обучение (semi-supervised learning)
- В обычном виде ранжирование часто решается статистически и не обучаемо, но существуют и обучаемые методы
- Примеры ранжирования:
  - Ранжирование поисковых запросов в соответствии с релевантностью и личными интересами пользователя
  - Ранжирование в системах коллаборативной фильтрации

Информация, данные, знания

- Понятия информации, данных и знаний достаточно интуитивны:
  - Информация – любые сведения
  - Данные – совокупность некоторых фактов, идей
  - Знания – совокупность фактов, закономерностей и правил их вывода
- В распоряжении современного общества имеется немыслимое количество информации и не меньшее количество данных
- Но действительно важны только знания и чтобы справляться с потоками информации необходимы способы выделения знаний из данных

Интеллектуальный анализ данных
Data mining)

- Процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности (определение Григория Пиатецкого-Шапиро)
- Data mining тесно связан с машинным обучением:
  - Классификация задач практически аналогична
  - Различия кроются в самом процессе анализа данных – машинное обучение как таковое не изучает подготовительные этапы
- Потребности в data mining растут с каждым днём (ещё в 2007 году объём информации, произведённый человеком достиг 295 эксабайт)

Области, в которых решаются задачи data mining

- Астрономия, биоинформатика
- Сегментация рынков, предсказание ухода клиентов
- Веб-поиск
- Антитерроризм, детектирование противозаконных действий
- И многие другие области, связанные с большими объёмами сырых данных

Summary

- История появления машинного обучения
- Машинное обучение, типы обучения
- Модели данных и алгоритмов, переобучение и недообучение
- Задачи машинного обучения
- Data mining

Веб-ресурсы

1. UC Irvine Machine Learning Repository
   Репозиторий задач машинного обучения университета Калифорнии Ирвайн
2. MachineLearning.ru
   Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных
3. AutonLab
   Лаборатория Auton университета Carnegie Mellon
4. arXiv.org
   Открытая библиотека научных статей. Разделы, связанные с машинным обучением: cs.AI, stat.ML, cs.IR, cs.CV, cs.LG

Основные источники

1. Владимир Вапник «Восстановление зависимостей по эмпирическим данным»
2. Николай Загоруйко «Прикладные методы анализа данных и знаний»
3. Tom Mitchell "Machine learning"
4. Trevor Hastie et al. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"
5. Christopher Bishop "Pattern recognition and machine learning"
6. Ethem Alpaydin "Introduction to machine learning"

Курсы машинного обучения

1. К.В. Воронцов «Машинное обучение», ФУПМ МФТИ, ВМиК МГУ, школа анализа данных Яндекса
2. Д.П. Ветров, Д.А. Кропотов «Байесовские методы машинного обучения», ВМиК МГУ
3. Н.Ю. Золотых «Машинное обучение», ВМК НижГУ
4. С.И. Николенко «Машинное обучение», Computer Science Club ПОМИ

Курсы машинного обучения

1. Andrew Ng, CS 229, Machine learning, Stanford University
2. Tom Mitchell et al., 10-701/15-781, Machine learning, Carnegie Mellon University
3. Tommi Jakkola, 6.867, Machine learning, Massachusetts Institute of Technology

Другие источники

- B. Allen "How to think like computer scientist: Learning with Python"
- T. Segaran "Programming collective intelligence"
- D. Hand et al. "Principles of Data Mining"
- I. Witten, E. Frank "Data mining: practical machine learning tools and techniques"
- D. MacKay "Information Theory, Inference, and Learning Algorithms"
- И. Чубукова "Data Mining"

Источники для этой лекции

- С.И. Николенко «Введение во всё-всё-всё»
- E. Xing "Introduction to ML and decision trees"
- К.В. Воронцов «Вычислительные методы обучения по прецедентам. Введение»
- Д.П. Ветров, Д.А. Кропотов «Различные задачи машинного обучения»
- Н.Ю. Золотых «Машинное обучение»
- G. Piatetsky-Shapiro «Machine Learning and Data Mining – Course notes»