Лекция 1. Различные задачи машинного обучения

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

Цели курса

- Ознакомление с классическими методами обработки данных, особенностями их применения на практике и их недостатками
- Представление современных проблем теории машинного обучения
- Введение в байесовские методы машинного обучения
- Изложение последних достижений в области практического использования байесовских методов
- Напоминание основных результатов из смежных дисциплин (теория кодирования, анализ, матричные вычисления, статистика, линейная алгебра, теория вероятностей, случайные процессы)

Структура курса

- 1 семестр, 12 лекций, 24 аудиторных часа + 12 часов для самостоятельной работы
- В каждой лекции секция ликбеза, содержащая краткое напоминание полезных фактов из смежных областей математики
- В конце курса экзамен. Три вопроса в билете, один из секции ликбеза + задача
- Каждая лекция сопровождается показом презентации
- Методические материалы (включая презентации), а также большая часть рекомендуемой литературы доступна на сайте http://mmphome.1gb.ru
- Лекторы: Дмитрий Ветров (VetrovD@yandex.ru) и Дмитрий Кропотов (DKropotov@yandex.ru)

План лекции

Некоторые задачи машинного обучения

- Задача классификации
- Задача восстановления регрессии
- Задача кластеризации (обучения без учителя)
- Задача идентификации
- Задача прогнозирования
- Задача извлечения знаний

Основные проблемы машинного обучения

- Малый объем обучающей выборки
- Некорректность входных данных
- Переобучение

Ликбез

- Основные понятия мат. статистики

Концепция машинного обучения

- Решение задач путем обработки прошлого опыта (case-based reasoning)
- Альтернатива построению математических моделей (model-based reasoning)
- Основное требование – наличие обучающей информации
- Как правило в качестве таковой выступает выборка прецедентов – ситуационных примеров из прошлого с известным исходом
- Требуется построить алгоритм, который позволял бы обобщить опыт прошлых наблюдений/ситуаций для обработки новых, не встречавшихся ранее случаев, исход которых неизвестен.

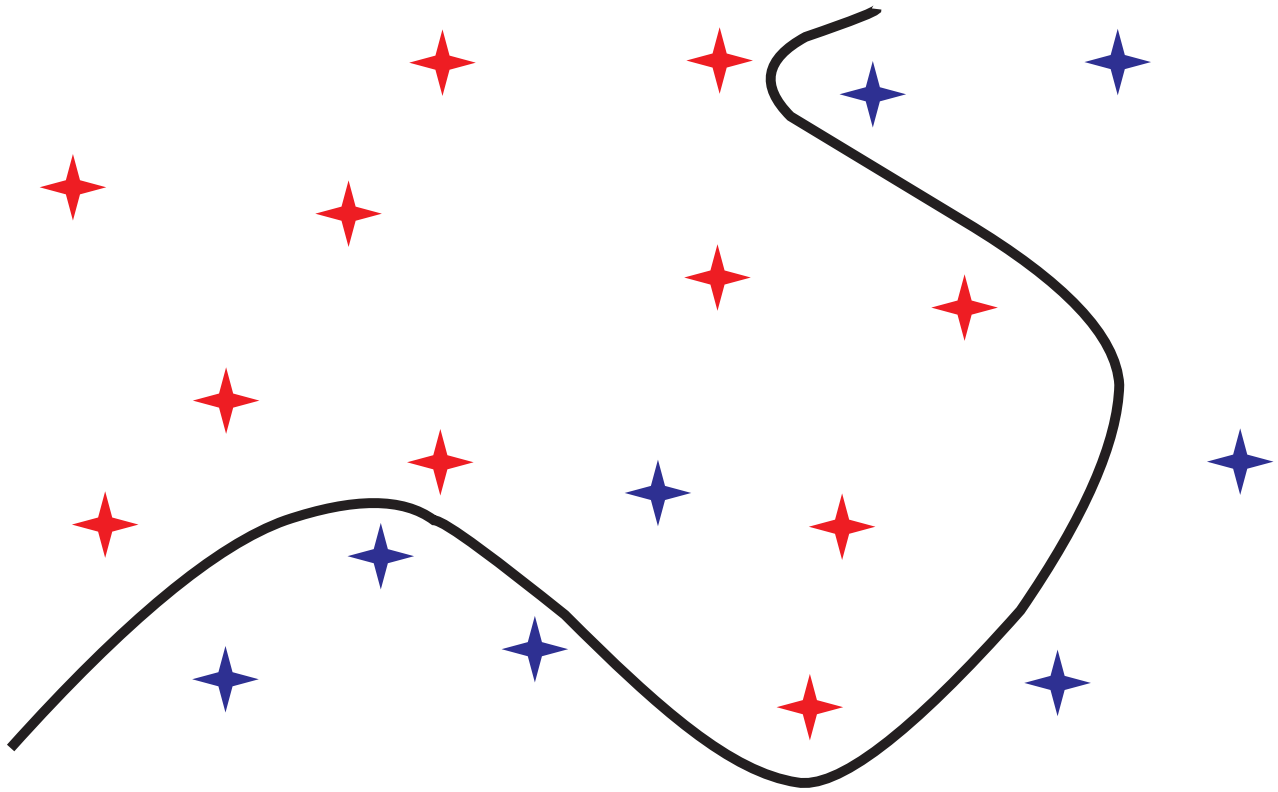
План лекции

Некоторые задачи машинного обучения
Задача классификации
Задача восстановления регрессии
Задача кластеризации (обучения без учителя)
Задача идентификации
Задача прогно-зирования
Задача извлечения знаний
Основные проблемы машинного обучения
Ликбез

Классификация

- Исторически возникла из задачи машинного зрения, поэтому часто употребляемый синоним – распознавание образов
- В классической задаче классификации обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- В качестве исхода объекта $\boldsymbol{x}$ фигурирует переменная t, принимающая конечное число значений, обычно из множества $\mathcal{T} = \{1, \ldots, l\}$
- Требуется постросить алгоритм (классификатор), который по вектору признаков $\boldsymbol{x}$ вернул бы метку класса $\hat{t}$ или вектор оценок принадлежности (апотериорных вероятностей) к каждому из классов $\{p(s|\boldsymbol{x})\}_{s=1}^l$

Классификация



Примеры задач классификации

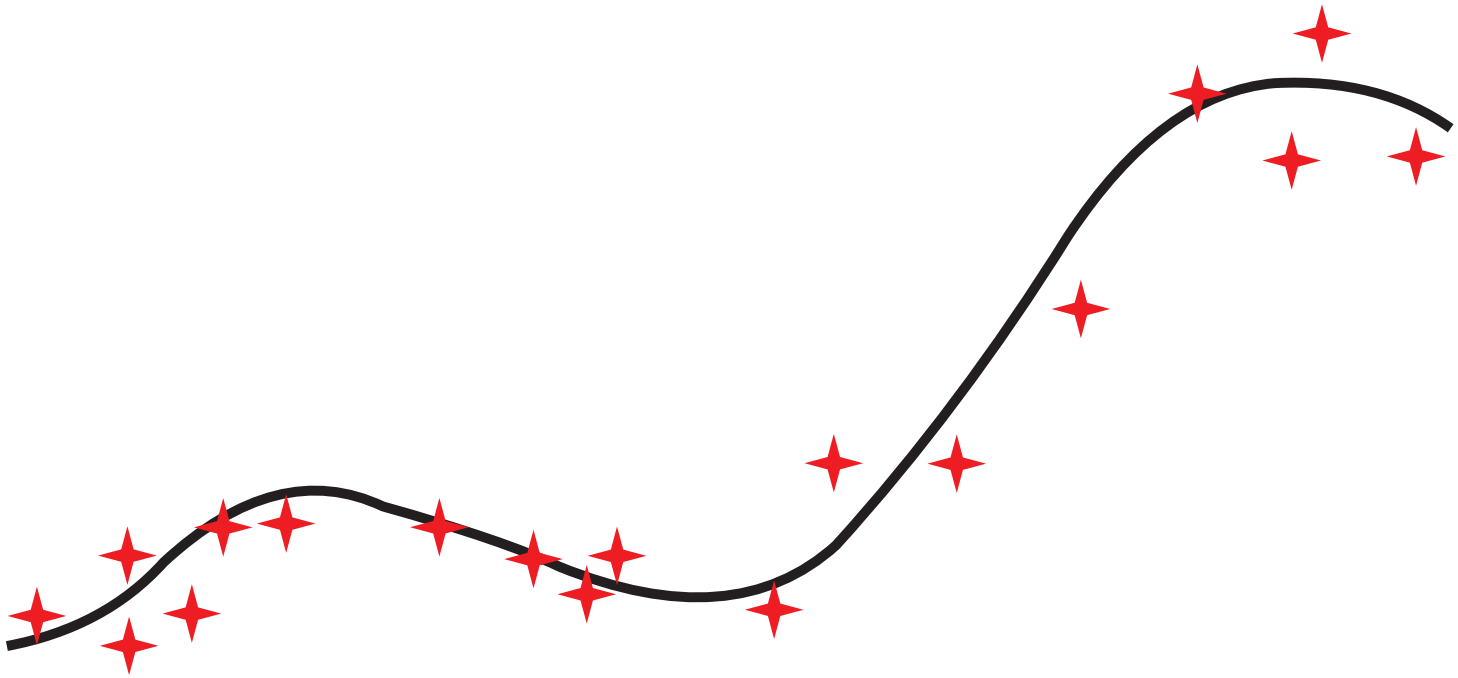
- Медицинская диагностика: по набору медицинских характеристик требуется поставить диагноз
- Геологоразведка: по данным зондирования почв определить наличие полезных ископаемых
- Оптическое распознавание текстов: по отсканированному изображению текста определить цепочку символов, его формирующих
- Кредитный скоринг: по анкете заемщика принять решение о выдаче/отказе кредита
- Синтез химических соединений: по параметрам химических элементов спрогнозировать свойства получаемого соединения

План лекции

Регрессия

- Исторически возникла при исследовании влияния одной группы непрерывных случайных величин на другую группу непрерывных случайных величин
- В классической задаче восстановления регрессии обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- В качестве исхода объекта $\boldsymbol{x}$ фигурирует непрерывная вещественнозначная переменная t
- Требуется постросить алгоритм (регрессор), который по вектору признаков $\boldsymbol{x}$ вернул бы точечную оценку значения регрессии $\hat{t}$, доверительный интервал (t_-, t_+) или апостериорное распределение на множестве значений регрессионной переменной $p(t|\boldsymbol{x})$

Регрессия



Примеры задач восстановления регрессии

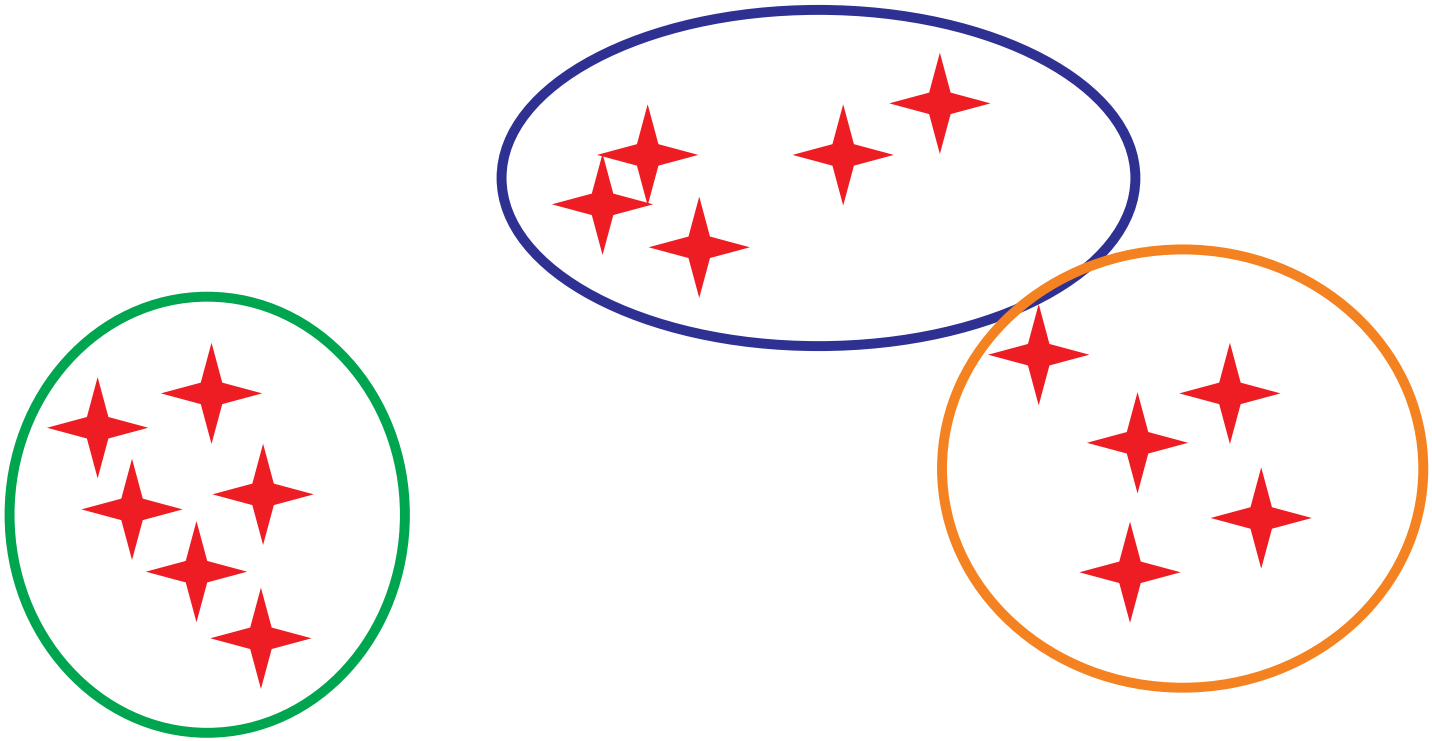
- Оценка стоимости недвижимости: по характеристике района, экологической обстановке, транспортной связности оценить стоимость жилья
- Прогноз свойств соединений: по параметрам химических элементов спрогнозировать температуру плавления, электропроводность, теплоемкость получаемого соединения
- Медицина: по постоперационным показателям оценить время заживления органа
- Кредитный скоринг: по анкете заемщика оценить величину кредитного лимита
- Инженерное дело: по техническим характеристикам автомобиля и режиму езды спрогнозировать расход топлива

План лекции

Кластеризация

- Исторически возникла из задачи группировки схожих объектов в единую структуру (кластер) с последующим выявлением общих черт
- В классической задаче кластеризации обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- Требуется постросить алгоритм (кластеризатор), который разбил бы выборку на непересекающиеся группы (кластеры) $X = \bigcup_{j=1}^k \boldsymbol{C}_k$, $\boldsymbol{C}_j \subset \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$, $\boldsymbol{C}_i \cap \boldsymbol{C}_j = \emptyset$
- В каждый класс должны попасть объекты в некотором смысле похожие друг на друга

Кластеризация



Примеры задач кластерного анализа

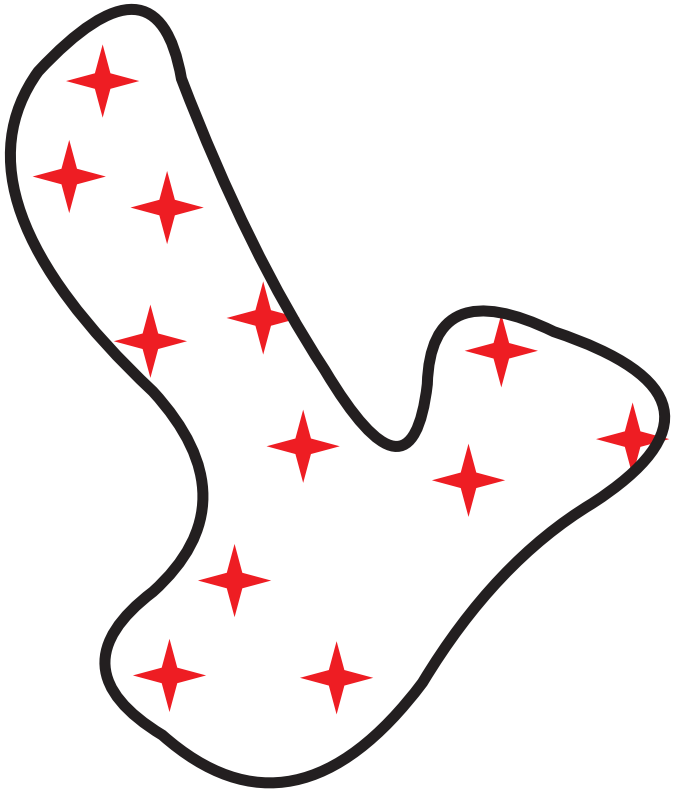
- Экономическая география: по физико-географическим и экономическим показателям разбить страны мира на группы схожих по экономическому положению государств
- Финансовая сфера: по сводкам банковских операций выявить группы «подозрительных», нетипичных банков, сгуппировать остальные по степени близости проводимой стратегии
- Маркетинг: по результатам маркетинговых исследований среди множества потребителей выделить характерные группы по степени интереса к продвигаемому продукту
- Социология: по результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества, а также характерные фокус-группы населения

План лекции

Идентификация

- Исторически возникла из классификации, необходимости отделить объекты, обладающие определенным свойством, от «всего остального»
- В классической задаче идентификации обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$, обладающих некоторым свойством $\chi_A(\boldsymbol{x}) = 1$
- Особенностью задачи является то, что все объекты принадлежат одному классу, причем не существует возможности сделать репрезентативную выборку из класса «все остальное»
- Требуется построить алгоритм (идентификатор), который по вектору признаков $\boldsymbol{x}$ определил бы наличие свойства A у объекта $\boldsymbol{x}$, либо вернул оценку степени его выраженности $p(\chi_A(\boldsymbol{x}) = 1|\boldsymbol{x})$

Идентификация



Примеры задач идентификации

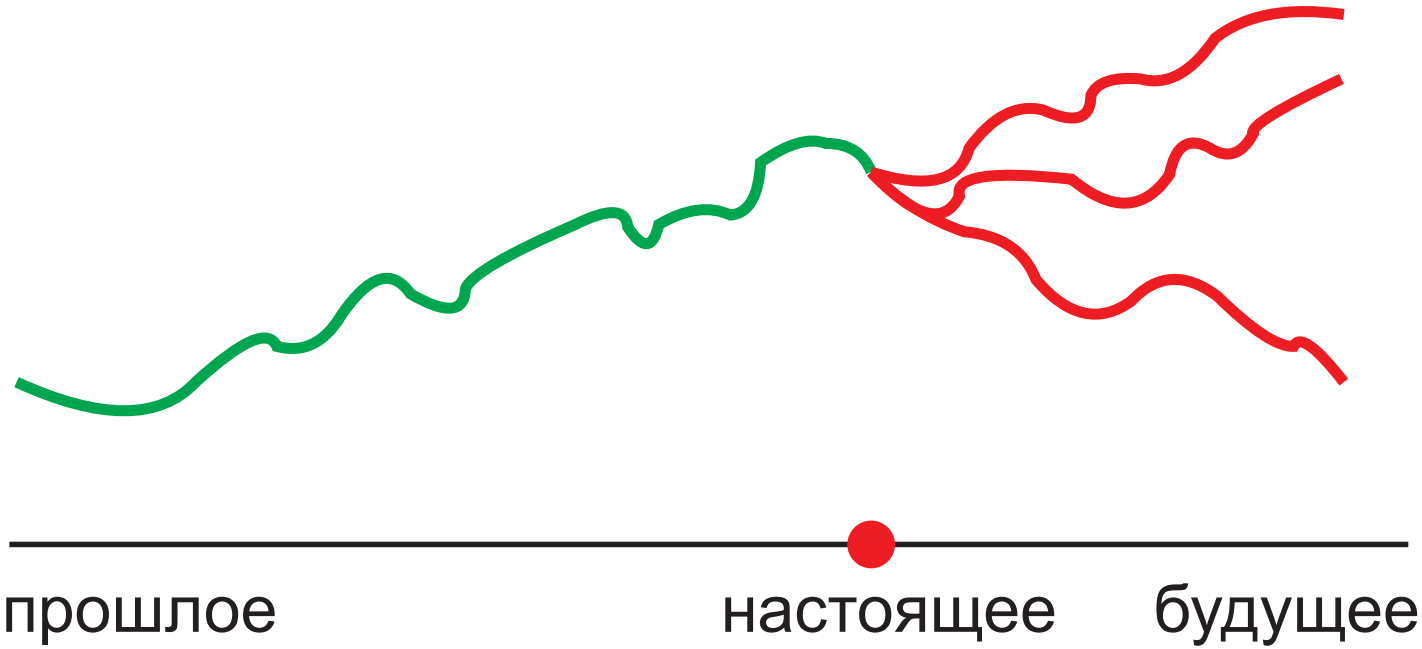
- Медицинская диагностика: по набору медицинских характеристик требуется установить наличие/отсутствие конкретного заболевания
- Системы безопасности: по камерам наблюдения в подъезде идентифицировать жильца дома
- Банковское дело: определить подлинность подписи на чеке
- Обработка изображений: выделить участки с изображениями лиц на фотографии
- Искусствоведение: по характеристикам произведения (картины, музыки, текста) определить, является ли его автором тот или иной автор

План лекции

Прогнозирование

- Исторически возникла при исследовании временных рядов и попытке предсказания их значений через какой-то промежуток времени
- В классической задаче прогнозирования обучающая выборка представляет собой набор измерений $X = \{\boldsymbol{x}[i]\}_{i=1}^{n}$, представляющих собой вектор вещественнозначных величин $\boldsymbol{x}[i] = (x_1[i], \ldots, x_d[i])$, сделанных в определенные моменты времени
- Требуется постросить алгоритм (предиктор), который вернул бы точечную оценку $\{\hat{\boldsymbol{x}}[i]\}_{i=n+1}^{n+q}$, доверительный интервал $\{(\boldsymbol{x}_-[i], \boldsymbol{x}_+[i])\}_{i=n+1}^{n+q}$ или апостериорное распределение $p(\boldsymbol{x}[n+1], \ldots, \boldsymbol{x}[n+q] | \boldsymbol{x}[1], \ldots, \boldsymbol{x}[n])$ прогноза на заданную глубину q
- В отличие от задачи восстановления регрессии, здесь осуществляется прогноз по времени, а не по признакам

Прогнозирование



Примеры задач прогнозирования

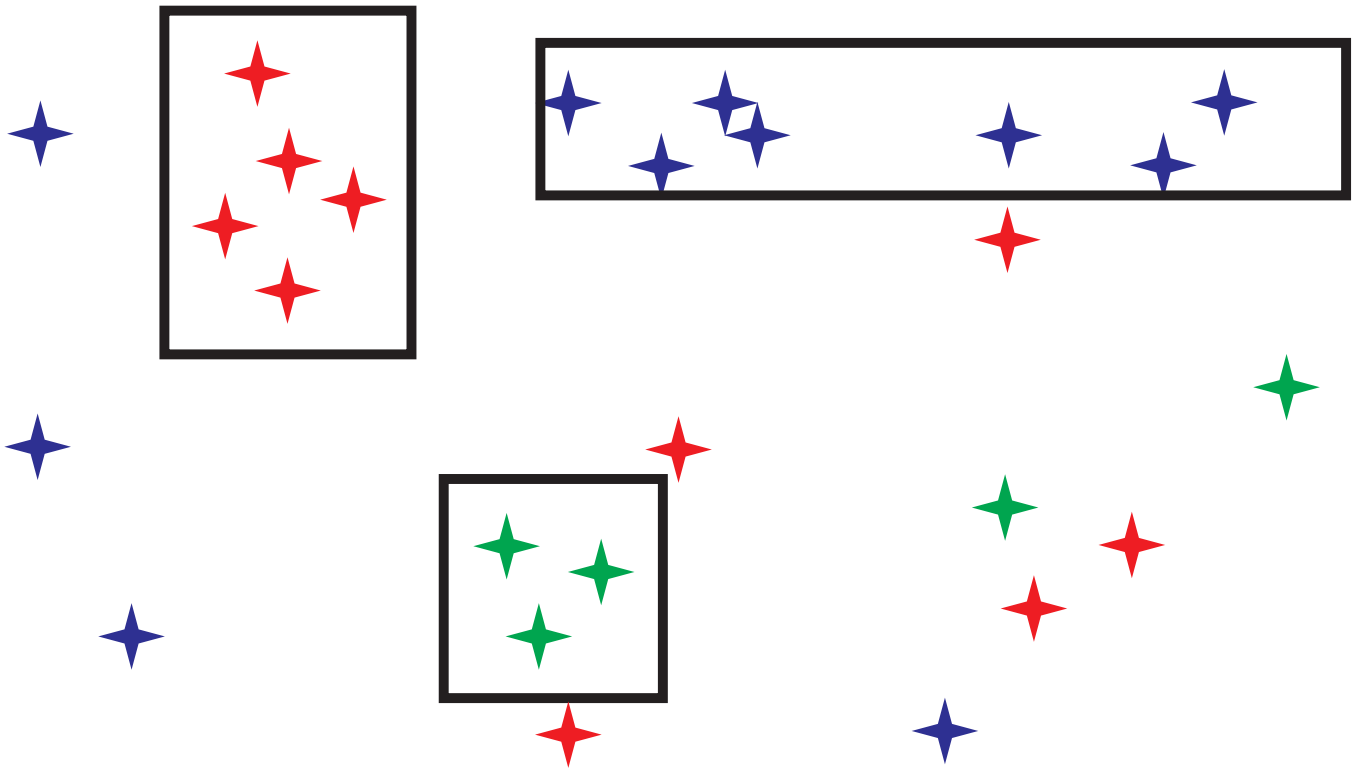
- Биржевое дело: прогнозирование биржевых индексов и котировок
- Системы управления: прогноз показателей работы реактора по данным телеметрии
- Экономика: прогноз цен на недвижимость
- Демография: прогноз изменения численности различных социальных групп в конкретном ареале
- Гидрометеорология: прогноз геомагнитной активности

План лекции

Извлечение знаний

- Исторически возникла при исследовании взаимозависимостей между косвенными показателями одного и того же явления
- В классической задаче извлечения знаний обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, \boldsymbol{x}_{i,d})$
- Требуется постросить алгоритм, генерирующий набор объективных закономерностей между признаками, имеющих место в генеральной совокупности
- Закономерности обычно имеют форму предикатов «ЕСЛИ ... ТО ...» и могут выражаться как в цифровых терминах $((0.45 \leq x_4 \leq 32.1)\&(-6.98 \leq x_7 \leq -6.59) \Rightarrow (3.21 \leq x_2 \leq 3.345))$, так и в текстовых («ЕСЛИ Давление – низкое И (Реакция – слабая ИЛИ Реакция – отсутствует) ТО Пульс – нитевидный»)

Извлечение знаний



Примеры задач извлечения знаний

- Медицина: поиск взаимосвязей (синдромов) между различными показателями при фиксированной болезни
- Социология: определение факторов, влияющих на победу на выборах
- Генная инженерия: выявление связанных участков генома
- Научные исследования: получение новых знаний об исследуемом процессе
- Биржевое дело: определение закономерностей между различными биржевыми показателями

План лекции

Объем выборки I

- Основным объектом работы любого метода машинного обучения служит обучающая выборка
- Большой объем выборки позволяет
  - Получить более надежные результаты
  - Использовать более сложные модели алгоритмов
  - Оценить точность обучения
  - НО: Время обучения быстро растет
- При малых выборках
  - Можно использовать только простые модели алгоритмов
  - Скорость обучения максимальна – можно использовать методы, требующие много времени на обучение
  - Высока вероятность переобучения при ошибке в выборе модели

Объем выборки II

- Одна и та же выборка может являться большой для простых моделей алгоритмов и малой для сложных моделей.
- Для методов с т.н. бесконечной емкостью Вапника-Червоненкиса любая выборка является малой.
- С ростом числа признаков увеличивается количество объектов, необходимое для корректного анализа данных
- Часто рассматривается т.н. эффективная размерность выборки $\frac{n}{d}$
- При объемах данных порядка десятков и сотен тысяч встает проблема уменьшения выборки с сохранением ее репрезентативности (active learning)

План лекции

Неполнота признакового описания

- Отдельные признаки могут отсутствовать у некоторых объектов. Это может быть связано с отсутствием данных об измерении данного признака для данного объекта, а может быть связано с принципиальным отсутствием данного свойства у данного объекта
- Такое часто встречается в медицинских и химических данных
- Необходимы специальные процедуры, позволяющие корректно обрабатывать пропуски в данных
- Одним из возможных способов такой обработки является замена пропусков на среднее по выборке значение данного признака
- По возможности, пропуски следует игнорировать и исключать из рассмотрения при анализе соответствующего объекта

Противоречивость данных

- Объекты с одним и тем же признаковым описанием могут иметь разные исходы (принадлежать к разным классам, иметь отличные значения регрессионной переменной и т.п.)
- Многие методы машинного обучения не могут работать с такими наборами данных
- Необходимо заранее исключать или корректировать противоречащие объекты
- Использование вероятностных методов обучения позволяет корректно обрабатывать противоречивые данные
- При таком подходе предполагается, что исход t для каждого признакового описания $\boldsymbol{x}$ есть случайная величина, имеющая некоторое условное распределение $p(t|\boldsymbol{x})$

Разнородность признаков

- Хотя формально предполагается, что признаки являются вещественнозначными, они могут быть дискретными и номинальными
- Номинальные признаки отличаются особенностями метрики между значениями
- Стандартная практика состоит в замене номинальных признаков на набор бинарных переменных по числу значений номинального признака
- Текстовые признаки, признаки-изображения, даты и пр. необходимо заменить на соответствующие номинальные либо числовые значения

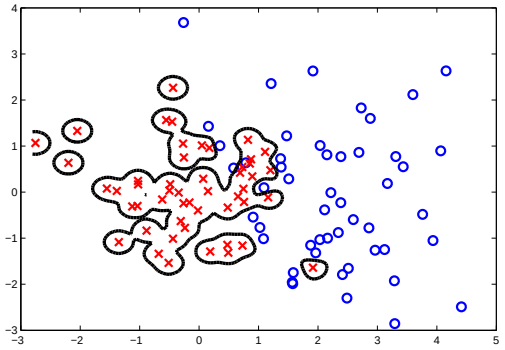
План лекции

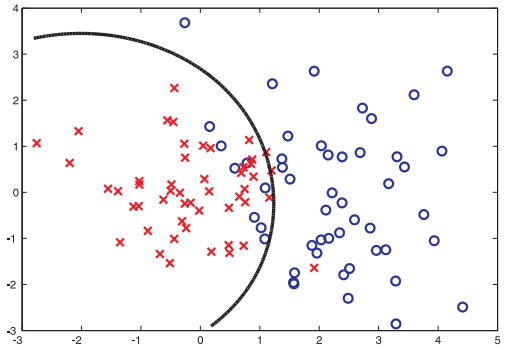
Идея машинного обучения

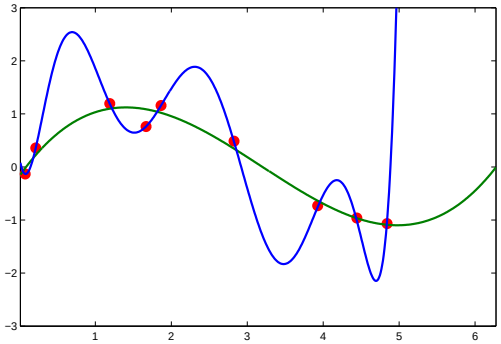
- Задача машинного обучения заключается в восстановлении зависимостей по конечным выборкам данных (прецедентов)
- Пусть $(X, \boldsymbol{t}) = (\boldsymbol{x}_i, t_i)_{i=1}^n$ – обучающая выборка, где $\boldsymbol{x}_i \in \mathbb{R}^d$ – признаковое описание объекта, а $t \in \mathcal{T}$ – значение скрытой компоненты (классовая принадлежность (не по Марксу!), значение прогноза, номер кластера и т.д.)
- При статистическом подходе к решению задачи МО предполагается, что обучающая выборка является выборкой из некоторой генеральной совокупности с плотностью $p(\boldsymbol{x}, t)$
- Требуется восстановить $p(t|\boldsymbol{x})$, т.е. знание о скрытой компоненте объекта по измеренным признакам

Проблема переобучения

Прямая минимизация невязки на обучающей выборке ведет к получению решающих правил, способных объяснить все что угодно и найти закономерности даже там, где их нет.







Способы оценки и увеличения обобщающей способности

- На сегодняшний день единственным универсальным способом оценивания обобщающей способности является кросс-валидация
- Все попытки предложить что-нибудь отличное от метода проб и ошибок пока не привели к общепризнанному решению. Наиболее известны из них следующие:
  - Структурная минимизация риска (В. Вапник, А. Червоненкис, 1974)
  - Минимизация длины описания (Дж. Риссанен, 1978)
  - Информационные критерии Акаике и Байеса-Шварца (Акаике, 1974, Шварц, 1978)
  - Максимизация обоснованности (МакКай, 1992)
- Последний принцип позволяет надеяться на конструктивное решение задачи выбора модели

Примеры задач выбора модели

- Определение числа кластеров в данных
- Выбор коэффициента регуляризации в задаче машинного обучения (например, коэффициента затухания весов (weight decay) в нейронных сетях)
- Установка степени полинома при интерполяции сплайнами
- Выбор наилучшей ядровой функции в методе опорных векторов (SVM)
- Определение количества ветвей в решающем дереве
- и многое другое...

План лекции

Некоторые задачи машинного обучения
- Задача классификации
- Задача восстановления регрессии
- Задача кластеризации (обучения без учителя)
- Задача идентификации
- Задача прогнозирования
- Задача извлечения знаний

Основные проблемы машинного обучения
- Малый объем обучающей выборки
- Некорректность входных данных
- Переобучение

Ликбез
- Основные понятия мат. статистики

Краткое напоминание основных вероятностных понятий

- $X : \Omega \to \mathbb{R}$ – случайная величина
- Вероятность попадания величины в интервал (a, b) равна

$$P(a \leq X \leq b) = \int_a^b p(x)dx,$$

где $p(x)$ – плотность распределения X,

$$p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

- Если поведение случайной величины определяется некоторым параметром, возникают условные плотности $p(x|\theta)$. Если рассматривать условную плотность как функцию от параметра

$$f(\theta) = p(x|\theta),$$

то принято говорить о т.н. функции правдоподобия

Основная задача мат. статистики

- Распределение случайной величины X известно с точностью до параметра θ
- Имеется выборка значений величины X, $\boldsymbol{x} = (x_1, \ldots, x_n)$
- Требуется оценить значение θ
- Метод максимального правдоподобия

$$\hat{\theta}_{ML} = \arg\max f(\theta) = \arg\max p(\boldsymbol{x}|\theta) = \arg\max \prod_{i=1}^{n} p(x_i|\theta)$$

- Можно показать, что ММП является ассимптотически оптимальным при $n \to \infty$
- Увы, мир несовершенен. Величина n конечна и обычно не слишком велика
- Необходима регуляризация метода

Пример некорректного использования метода максимального правдоподобия

Задача восстановления смеси нормальных распределений

- $X \sim w_1\mathcal{N}(\mu_1, \sigma_1^2) + \cdots + w_m\mathcal{N}(\mu_m, \sigma_m^2)$
- Необходимо определить $\boldsymbol{\theta} = (m, \mu_1, \sigma_1^2, \ldots, \mu_m, \sigma_m^2, w_1, \ldots, w_m)$
- Применяем ММП

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) =$$

$$\prod_{i=1}^{n}\sum_{j=1}^{m} \frac{w_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right) \to \max_{\boldsymbol{\theta}}$$

- Решение

$$\hat{m}_{ML} = n$$

$$\hat{\mu}_{j,ML} = x_j$$

$$\hat{\sigma}^2_{j,ML} = 0$$

$$\hat{w}_{ML,j} = \frac{1}{n}$$

Выводы

- Не все параметры можно настраивать в ходе обучения
- Существуют специальные параметры (будем называть их структурными), которые должны быть зафиксированы до начала обучения
- !! В данном случае величина m (количество компонент смеси) является структурным параметром!!
- Основной проблемой машинного обучения является проблема выбора структурных параметров, позволяющих избегать переобучения