



КЛАССИФИКАЦИЯ

В человеческой сфере невозможна никакая линейная классификация.
Э. Менье

Технологии анализа данных

Содержание

2

- Понятия классификации и прогноза
- Процесс классификации
- Классификация с помощью деревьев решений
- Классификация с помощью k ближайших соседей

Технологии анализа данных © М.Л. Цымблер

Классификация и прогноз

3

- *Классификация* – определение классов, к которым принадлежат объекты заданного множества, по характеристикам этих объектов. При этом множество классов, к которым может быть отнесен объект, заранее известно.
- *Прогноз* – моделирование непрерывной функции, предсказание неизвестных или отсутствующих значений.
- Применение
 - Выдача кредита
 - Целевой маркетинг
 - Медицинская диагностика
 - Выявление мошенничества

Технологии анализа данных © М.Л. Цымблер

Классификация как процесс

4

- **Построение модели** – описание множества predetermined классов
 - Каждый кортеж множества принадлежит некоторому predetermined классу, значение которого фиксировано в *классификационном атрибуте*
 - Множество кортежей для построения модели – *обучающая выборка (training set)*.
 - Модель может быть представлена в виде классификационных правил, деревьев решений или математическими формулами
- **Использование модели** для классификации будущих или неизвестных объектов
 - Оценка точности модели
 - Сравнение значения классификационного атрибута у кортежей *тестовой выборки* с результатом классификации этих кортежей, полученного с помощью модели.
 - Точность модели – доля (%) кортежей, корректно классифицированных с помощью модели.
 - Тестовая и обучающая выборки должны быть независимыми, иначе имеет место *подгонка (overfitting)*.
 - Если полученная точность приемлема, использовать модель для классификации кортежей, значение атрибута классификации которых неизвестно.

Технологии анализа данных © М.Л. Цымблер

Классификация

5

Обучающая выборка

ФИО	Доход	Возраст	Выдать кредит
Иванов	Низкий	<30	НЕТ
Петров	Средний	30..40	ДА
Сидоров	Высокий	<30	ДА
Егоров	Средний	>40	ДА
Бендер	Низкий	30..40	ДА
Балаганов	Средний	<30	НЕТ



ЕСЛИ Доход=Высокий
ИЛИ Возраст>30
ТО ВыдатьКредит=ДА

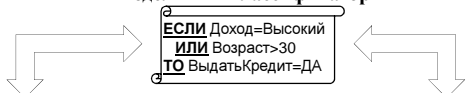
Модельный классификатор

Технологии анализа данных © М.Л. Цымблер

Классификация

6

Модельный классификатор



Тестовая выборка

ФИО	Доход	Возраст	Выдать кредит
Васечкин	Низкий	<30	НЕТ
Петров	Средний	<30	НЕТ
Воробьянинов	Высокий	>40	ДА
Бонд	Средний	30..40	ДА

Точность модели

Выдать кредит	ОК
НЕТ	✓
ДА	✗
ДА	✓
ДА	✓

75%


Технологии анализа данных © М.Л. Цымблер

Классификация

7

Модельный классификатор

ЕСЛИ Доход=Высокий
ИЛИ Возраст>30
ТО ВыдатьКредит=ДА



Новые corteжи

ФИО	Доход	Возраст	Выдать кредит	Выдать кредит
Берлиоз	Низкий	<30	?	НЕТ
Паркер	Низкий	60	?	ДА

Технологии анализа данных © М.Л. Цымблер

Методы классификации

8

- ☐ Деревья решений
- ☐ Метод k ближайших соседей
- ☐ Нейронные сети
- ☐ Байесовская классификация
- ☐ Генетические алгоритмы
- ☐ Нечеткие множества
- ☐ ...

Технологии анализа данных © М.Л. Цымблер

Оценка методов классификации

9

- ☐ Точность прогноза
 - ☐ Способность модели корректно предсказать класс
- ☐ Скорость и масштабируемость
 - ☐ Время построения модели
 - ☐ Время использования модели
 - ☐ Эффективность для сверхбольших баз данных
- ☐ Устойчивость
 - ☐ Обработка шумов и отсутствующих значений
- ☐ Интерпретируемость
 - ☐ Уровень понимания и знания, предоставляемых моделью

Технологии анализа данных © М.Л. Цымблер

Деревья решений

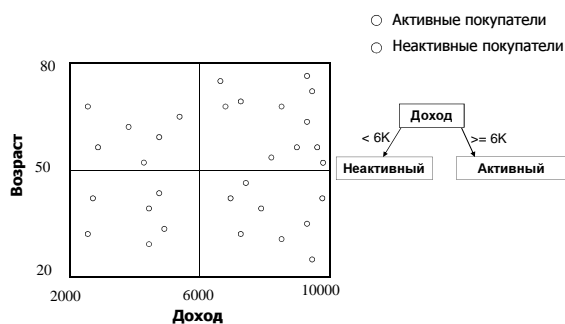
10

- Дерево решений – дерево, в котором
 - внутренние узлы представляют собой операции проверки значения указанного атрибута
 - ветви представляют собой переходы в соответствии с результатом проверки значения указанного атрибута
 - листья представляют собой метки класса или их диапазоны

Технологии анализа данных © М.Л. Цымблер

Деревья решений: пример

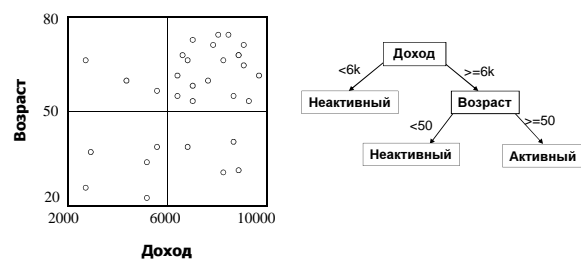
11



Технологии анализа данных © М.Л. Цымблер

Деревья решений: пример

12



Технологии анализа данных © М.Л. Цымблер

Деревья решений: пример

13

Прогноз	Температура	Влажность	Ветер	Игра?
Солнце	Жарко	Высокая	Нет	Нет
Солнце	Жарко	Высокая	Да	Нет
Облачность	Жарко	Высокая	Нет	Да
Дождь	Умеренно	Высокая	Нет	Да
Дождь	Холодно	Нормальная	Нет	Да
Дождь	Холодно	Нормальная	Да	Нет
Облачность	Холодно	Нормальная	Да	Да
Солнце	Умеренно	Высокая	Нет	Нет
Солнце	Холодно	Нормальная	Нет	Да
Дождь	Умеренно	Нормальная	Нет	Да
Солнце	Умеренно	Нормальная	Да	Да
Облачность	Умеренно	Высокая	Да	Да
Облачность	Жарко	Нормальная	Нет	Да
Дождь	Умеренно	Высокая	Да	Нет



Технологии анализа данных © М.Л. Цымблер

Построение дерева решений

14

- Построение
 - Поместить все кортежи обучающей выборки в корень дерева.
 - Разбивать множество кортежей рекурсивно на основе отбираемых атрибутов.
- Сокращение
 - удалить ветви дерева, которые могут отражать шумы в обучающей выборке и привести к ошибкам при классификации тестовых данных (повышение точности классификации)

Технологии анализа данных © М.Л. Цымблер

Спецификация условия отбора

15

- В зависимости от типа атрибута
 - Номинальный
 - Порядковый
 - Непрерывный
- В зависимости от степени разбиения
 - Бинарное
 - N-арное

Технологии анализа данных © М.Л. Цымблер

Разбиение по номинальным атрибутам

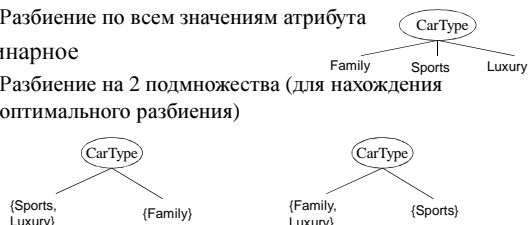
16

□ N-арное

- ▣ Разбиение по всем значениям атрибута

□ Бинарное

- ▣ Разбиение на 2 подмножества (для нахождения оптимального разбиения)



Технологии анализа данных © М.Л. Цымбалер

Разбиение по порядковым атрибутам

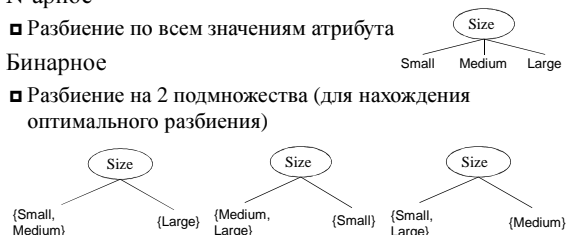
17

□ N-арное

- ▣ Разбиение по всем значениям атрибута

□ Бинарное

- ▣ Разбиение на 2 подмножества (для нахождения оптимального разбиения)



Технологии анализа данных © М.Л. Цымбалер

Разбиение по непрерывным атрибутам

18

□ Дискретизация для формирования порядкового атрибута-категории

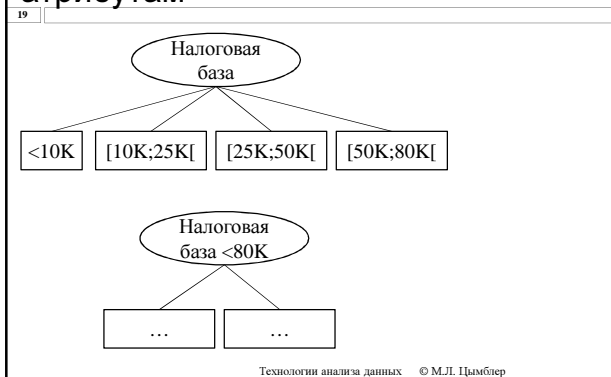
- ▣ Статическая дискретизация – однажды перед построением дерева
- ▣ Динамическая дискретизация – промежутки могут быть найдены, например, с помощью кластеризации

□ Бинарное разбиение ($A < v$) OR ($A \geq v$)

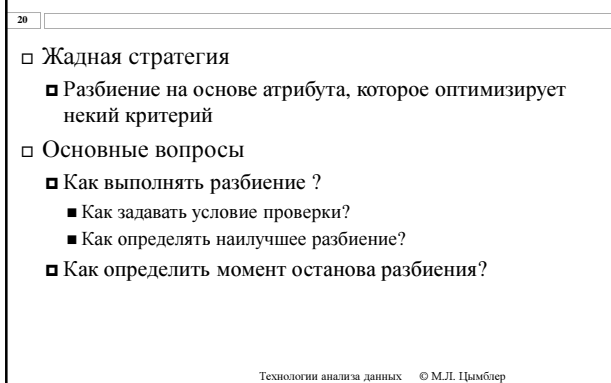
- ▣ рассмотреть все возможные разбиения и взять наилучшее
- ▣ может увеличить вычислительную сложность

Технологии анализа данных © М.Л. Цымбалер

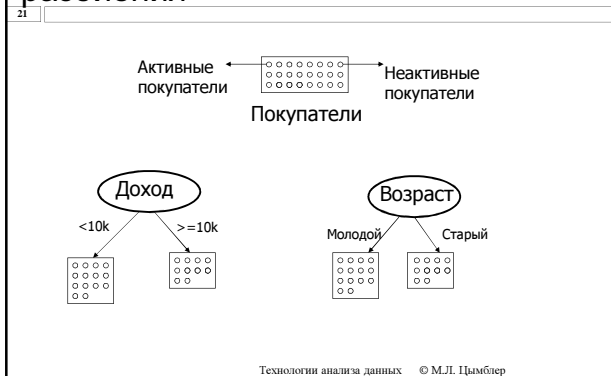
Разбиение по непрерывным атрибутам



Индуктивное построение дерева



Определение наилучшего разбиения



Определение наилучшего разбиения

22

- Жадный подход
 - Предпочтительнее узлы с однородным распределением
- Мера информационных примесей в узле



Высокая
степень
примесей

50%
50%



Низкая
степень
примесей

75% **75%**
25% **25%**



Отсутствие
примесей

100% **100%**
0% **0%**

Технологии анализа данных © М.Л. Цымблер

Меры инф. примесей

23

- Information gain (прирост информации)
- Gain ratio (соотношение прироста информации и информации, необходимой для разбиения)
- Gini index (индекс Джини)

Технологии анализа данных © М.Л. Цымблер

Алгоритм индуктивного построения дерева решений

24

- Основной (жадный) алгоритм
 - Дерево строится рекурсивно сверху вниз методом "разделяй и властвуй"
 - Вначале все элементы помещаются в корень дерева
 - Атрибуты – категории (в случае непрерывных предварительно выполняется дискретизация)
 - Разбиение выполняется рекурсивно на основе выбранных атрибутов
 - Атрибуты выбираются с помощью эвристики или статистической меры (например, прирост информации)
- Условия останова разбиения
 - Все элементы данного узла принадлежат одному классу
 - Не осталось атрибутов для последующего разбиения – выполняется "голосование" и определение класса листа по большинству
 - Не осталось элементов для последующего разбиения

Технологии анализа данных © М.Л. Цымблер

Алгоритм ID3

25

- Выбирается атрибут, дающий наибольший прирост информации
- p_i – вероятность принадлежности произвольного элемента из множества D классу C_i , $p_i = |C_{i,D}|/|D|$
- *Ожидаемое количество информации (энтропия)*, необходимое для классификации элемента из D :

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

- *Информация, необходимая для классификации множества D* (после использования атрибута A для разбиения множества D на v подмножеств):

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$

- *Прирост информации* при разбиении по атрибуту A :

$$Gain(A) = Info(D) - Info_A(D)$$

Технологии анализа данных © М.Л. Цымблер

Алгоритм ID3

26

№ п/п	Возраст	Доход	Студент?	Кред. рейтинг	Купит компьютер?
1	<=30	Высокий	Нет	Удовл.	Нет
2	<=30	Высокий	Нет	Отличный	Нет
3	31...40	Высокий	Нет	Удовл.	Да
4	>40	Средний	Нет	Удовл.	Да
5	>40	Низкий	Да	Удовл.	Да
6	>40	Низкий	Да	Отличный	Нет
7	31...40	Низкий	Да	Отличный	Да
8	<=30	Средний	Нет	Удовл.	Нет
9	<=30	Низкий	Да	Удовл.	Да
10	>40	Средний	Да	Удовл.	Да
11	<=30	Средний	Да	Отличный	Да
12	31...40	Средний	Нет	Отличный	Да
13	31...40	Высокий	Да	Удовл.	Да
14	>40	Средний	Нет	Отличный	Нет

- Классы
 - C1 Купит компьютер = Да
 - C2 Купит компьютер = Нет
- $Info(D) = -\frac{9}{14} \log_2(\frac{9}{14}) - \frac{5}{14} \log_2(\frac{5}{14}) = 0.940$
- $Info_{Возраст}(D) = \frac{5}{14} \times (-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5}) + \frac{4}{14} \times (-\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4}) + \frac{5}{14} \times (-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5}) = 0.694 \text{ bits.}$
- $Gain(Возраст) = 0.940 - 0.694 = 0.246$

Технологии анализа данных © М.Л. Цымблер

Алгоритм ID3

27

№ п/п	Возраст	Доход	Студент?	Кред. рейтинг	Купит компьютер?
1	<=30	Высокий	Нет	Удовл.	Нет
2	<=30	Высокий	Нет	Отличный	Нет
3	31...40	Высокий	Нет	Удовл.	Да
4	>40	Средний	Нет	Удовл.	Да
5	>40	Низкий	Да	Удовл.	Да
6	>40	Низкий	Да	Отличный	Нет
7	31...40	Низкий	Да	Отличный	Да
8	<=30	Средний	Нет	Удовл.	Нет
9	<=30	Низкий	Да	Удовл.	Да
10	>40	Средний	Да	Удовл.	Да
11	<=30	Средний	Да	Отличный	Да
12	31...40	Средний	Нет	Отличный	Да
13	31...40	Высокий	Да	Удовл.	Да
14	>40	Средний	Нет	Отличный	Нет

- $Gain(Возраст) = 0.246$
- $Gain(Доход) = 0.029$
- $Gain(Студент?) = 0.151$
- $Gain(КредРейтинг) = 0.048$

Технологии анализа данных © М.Л. Цымблер

Алгоритм ID3

28

№ п/п	Доход	Студент?	Кред. рейтинг	Купит комп-р?
1	Высокий	Нет	Удовл.	Нет
2	Высокий	Нет	Отличный	Нет
3	Средний	Нет	Удовл.	Нет
4	Низкий	Да	Удовл.	Да
5	Средний	Да	Отличный	Да

№ п/п	Доход	Студент?	Кред. рейтинг	Купит комп-р?
1	Высокий	Нет	Удовл.	Да
2	Низкий	Да	Отличный	Да
3	Средний	Нет	Отличный	Да
4	Высокий	Да	Удовл.	Да

Технологии анализа данных © М.Л. Цымблер

Алгоритм C4.5

29

- Прирост информации тяготеет к атрибутам с большим количеством значений
- C4.5 является последователем ID3 преодолевает данную проблему с помощью нормализации значения прироста
 - $GainRatio(A) = Gain(A) / SplitInfo_A(D)$
- $SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$
- Пример: $SplitInfo_{Студент}(D) = - \frac{5}{14} \times \log_2 \left(\frac{5}{14} \right) - \frac{4}{14} \times \log_2 \left(\frac{4}{14} \right) - \frac{5}{14} \times \log_2 \left(\frac{5}{14} \right) = 0.926$
- Для разбиения выбирается атрибут с макс. значением $GainRatio$.

Технологии анализа данных © М.Л. Цымблер

Алгоритм CART

30

- Индекс Джини множества D , элементы которого принадлежат n классам ($p_i = |C_{i,D}| / |D|$):

$$Gini(D) = 1 - \sum_{j=1}^n p_j^2$$
- Индекс Джини разбиения множества D по атрибуту A применяется для бинарного разбиения дискретных атрибутов и вычисляется как взвешенная сумма информационных примесей каждого результирующего разбиения

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$
- Редукция информационной примеси

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$
- Для разбиения выбирается атрибут с мин. $Gini_A$ (= макс. $\Delta Gini$)

Технологии анализа данных © М.Л. Цымблер

Алгоритм CART

31					
№ п/п	Возраст	Доход	Студент?	Кред. рейтинг	Купит комп-р?
1	<=30	Высокий	Нет	Удовл.	Нет
2	<=30	Высокий	Нет	Отличный	Нет
3	31...40	Высокий	Нет	Удовл.	Да
4	>40	Средний	Нет	Удовл.	Да
5	>40	Низкий	Да	Удовл.	Да
6	>40	Низкий	Да	Отличный	Нет
7	31...40	Низкий	Да	Отличный	Да
8	<=30	Средний	Нет	Удовл.	Нет
9	<=30	Низкий	Да	Удовл.	Да
10	>40	Средний	Да	Удовл.	Да
11	<=30	Средний	Да	Отличный	Да
12	31...40	Средний	Нет	Отличный	Да
13	31...40	Высокий	Да	Удовл.	Да
14	>40	Средний	Нет	Отличный	Нет

□ Классы

□ C1 Купит компьютер = Да

□ C2 Купит компьютер = Нет

□ Индекс Джини

$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$

□ Индекс Джини разбиения D по атрибуту $D_{доход} = \{\text{Низкий, Средний}\} \cup \{\text{Высокий}\}$

$Gini_{D_{доход} \in \{\text{Низкий, Средний}\}}(D) = \left(\frac{10}{14}\right) Gini(D_1) + \left(\frac{4}{14}\right) Gini(D_2)$

$= \frac{10}{14} \left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) + \frac{4}{14} \left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) = 0.450$

□ $Gini_{D_{доход} \in \{\text{Средний, Высокий}\}}(D) = \mathbf{0.300}$

□ $Gini_{D_{доход} \in \{\text{Низкий, Высокий}\}}(D) = 0.315$

Технологии анализа данных © М.Л. Цымблер

Сравнение мер для выбора атрибута разбиения

32

- Все три меры дают хорошие результаты.

- Относительные недостатки

- InfoGain

- тяготеет к атрибутам с большим количеством значений

- Gain ratio

- предпочитает несбалансированные разбиения, где одна часть существенно меньше остальных

- Gini index

- тяготеет к атрибутам с большим количеством значений

- сложнее при большом количестве классов

Технологии анализа данных

© М.Л. Цымблер

Популярность деревьев решений

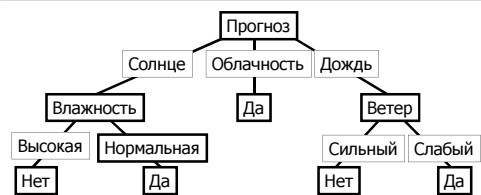
33	
----	--

- Более быстрая скорость обучения (в сравнении с др. методами классификации)
- Возможность преобразования в простые и понятные правила классификации
- Сравнимая точность классификации с др. методами

Технологии анализа данных © М.Л. Цымблер

Деревья решений и правила классификации

34



- R₁: IF (Прогноз=Солнце) AND (Влажность=Высокая) THEN Игра=Нет
- R₂: IF (Прогноз=Солнце) AND (Влажность=Нормальная) THEN Игра=Да
- R₃: IF (Прогноз=Облачность) THEN Игра=Да
- R₄: IF (Прогноз=Дождь) AND (Ветер=Сильный) THEN Игра=Нет
- R₅: IF (Прогноз=Дождь) AND (Ветер=Слабый) THEN Игра=Да

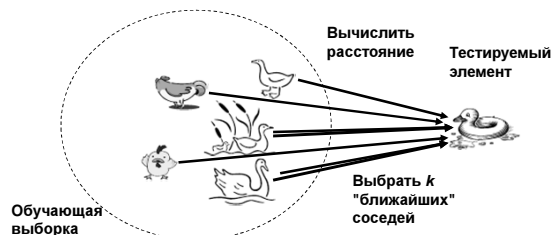
Технологии анализа данных © М.Л. Цымблер

Классификация по k ближайшим соседям

35

- Основная идея – утиный тест:

*If it looks like a duck, swims like a duck and quacks like a duck,
then it probably is a duck.*



Технологии анализа данных © М.Л. Цымблер

Обучение по примерам

36

Атр1	...	АтрN	Класс
			C2
			C1
			C3
			C1
			C4
			C3
...

- Элементы обучающей выборки сохраняются и используются для (поочередной) классификации элементов тестовой выборки.

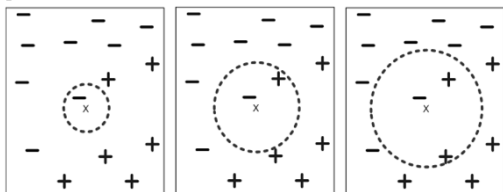
Атр1	...	АтрN	Класс
			?
			?
			?
...

Технологии анализа данных © М.Л. Цымблер

Определение ближайшего соседа

37

- k ближайших соседей (kNN , k Nearest Neighbors) элемента x – k элементов, имеющих минимальное расстояние до x .



(a) 1-nearest neighbor

(b) 2-nearest neighbor

(c) 3-nearest neighbor

Технологии анализа данных © М.Л. Цымблер

Классификация по k ближайшим соседям

38

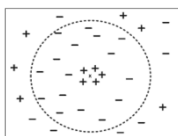
- Входные данные
 - Обучающая выборка D
 - Метрика d для вычисления расстояния
 - Количество ближайших соседей k
- Классификация
 - Вычислить расстояние от элемента тестовой выборки до элементов обучающей выборки
 - Евклидово расстояние
 - Манхэттенское расстояние
 - ...
 - Найти k ближайших соседей
 - Определить класс элемента тестовой выборки по меткам классов k ближайших соседей
 - Голосование по большинству
 - Взвешенное голосование, вес $w=1/d^2$

Технологии анализа данных © М.Л. Цымблер

kNN: вопросы реализации

39

- Иногда необходима нормализация некоторых атрибутов для исключения их доминантного влияния на расстояние
 - Рост: 1,5 м – 2 м
 - Вес: 50 кг – 100 кг
 - Доход: 10000 руб. – 1000000 руб.
- Выбор k
 - Слишком малое k – чувствительность к точкам шума
 - Слишком большое k – среди соседей может быть неоправданно много представителей др. классов



Технологии анализа данных © М.Л. Цымблер

Заключение

40

- Классификация – определение заранее известных классов, к которым принадлежат объекты заданного множества.
- Классификация предполагает
 - построение модели на основе обучающей выборки,
 - затем оценку точности модели на основе тестовой выборки
 - и последующее использование модели для определения класса не рассматривавшихся ранее кортежей.
- Дерево решений – дерево, в котором
 - внутренние узлы представляют собой операции проверки значения указанного атрибута
 - ветви представляют собой переходы в соответствии с результатом проверки значения указанного атрибута
 - листья представляют собой метки класса или их диапазоны.

Технологии анализа данных © М.Л. Цымблер

Заключение

41

- Меры информационной примеси, используемые при построении деревьев решений
 - Information gain (алгоритм ID3)
 - Gain ratio (алгоритм C4.5)
 - Gini index (алгоритм CART)
- Обучение по примерам – классификация методом k ближайших соседей.

Технологии анализа данных © М.Л. Цымблер
