



ТЕХНОСФЕРА

Лекция 4 Задача классификации

Николай Анохин

17 апреля 2015 г.

План занятия

Деревья решений

Деревья решений

Задача

Дано:

обучающая выборка из профилей
нескольких десятков тысяч
человек

- ▶ пол (binary)
- ▶ возраст (numeric)
- ▶ образование (nominal)
- ▶ и еще 137 признаков
- ▶ наличие интереса к косметике

Задача:

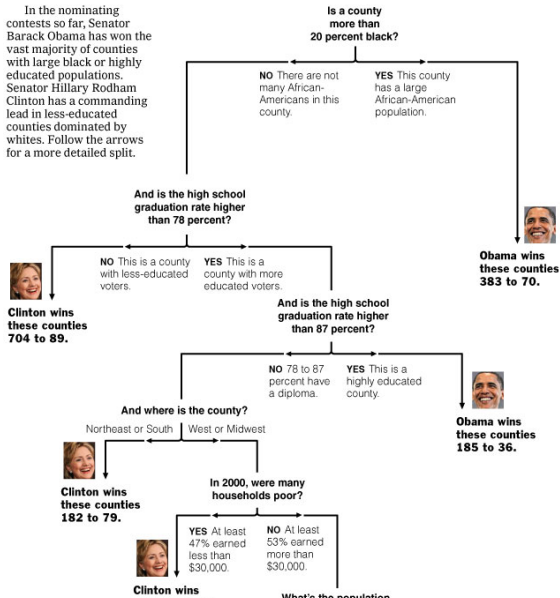
Для рекламной кампании
определить, характеристики
людей, интересующихся
косметикой



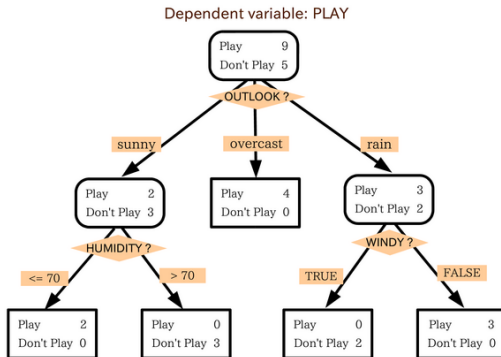
Обама или Клинтон?

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.

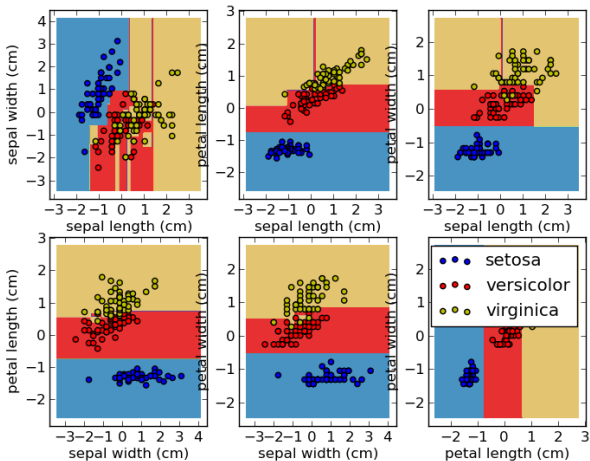


Хороший день для партии в гольф



Регионы принятия решений

Decision surface of a decision tree using paired features



Рекурсивный алгоритм

```
1 function decision_tree(X_N):  
2     if X_N satisfies leaf criterion:  
3         L = create_leaf(X_N)  
4         assign_class(L)  
5     else:  
6         L = create_node(X_N)  
7         X_1,...,X_S = split(L)  
8         for i in 1..S:  
9             C = decision_tree(X_i)  
10            add_child(L, C)  
11    return L
```


CART

Classification And Regression Trees

1. Как происходит разделение?
2. На сколько детей разделять каждый узел?
3. Какой критерий листа выбрать?
4. Как укоротить слишком большое дерево?
5. Как выбрать класс каждого листа?
6. Что делать, если часть значений отсутствует?

Чистота узла

Задача

Выбрать метод, позволяющий разделить узел на два или несколько детей наилучшим образом

Ключевое понятие – *impurity* узла.

1. Misclassification

$$i(N) = 1 - \max_k p(x \in C_k)$$

2. Gini

$$i(N) = 1 - \sum_k p^2(x \in C_k) = \sum_{i \neq j} p(x \in C_i) p(x \in C_j)$$

3. Информационная энтропия

$$i(N) = - \sum_k p(x \in C_k) \log_2 p(x \in C_k)$$

Теория информации

Количество информации \sim “степень удивления”

$$h(x) = -\log_2 p(x)$$

Информационная энтропия $H[x] = E[h(x)]$

$$H[x] = -\sum p(x) \log_2 p(x) \quad \text{или} \quad H[x] = -\int p(x) \log_2 p(x) dx$$

Упражнение

Дана случайная величина x , принимающая 4 значения с равными вероятностями $\frac{1}{4}$, и случайная величина y , принимающая 4 значения с вероятностями $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$. Вычислить $H[x]$ и $H[y]$.

Выбор наилучшего разделения

Критерий

Выбрать признак и точку отсечения такими, чтобы было максимально уменьшение *impurity*

$$\Delta i(N, N_L, N_R) = i(N) - \frac{N_L}{N} i(N_L) - \frac{N_R}{N} i(N_R)$$

Замечания

- ▶ Выбор границы при числовых признаках: середина?
- ▶ Решения принимаются локально: нет гарантии глобально оптимального решения
- ▶ На практике выбор *impurity* не сильно влияет на результат

Если разделение не бинарное

Естественный выбор при разделении на B детей

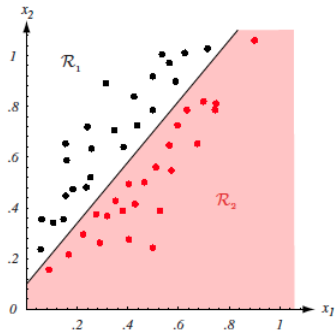
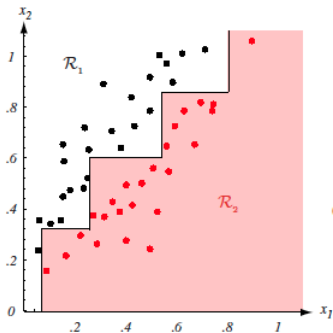
$$\Delta i(N, N_1, \dots, N_B) = i(N) - \sum_{k=1}^B \frac{N_k}{N} i(N_k) \rightarrow \max$$

Предпочтение отдается большим B . Модификация:

$$\Delta i_B(N, N_1, \dots, N_B) = \frac{\Delta i(N, N_1, \dots, N_B)}{-\sum_{k=1}^B \frac{N_k}{N} \log_2 \frac{N_k}{N}} \rightarrow \max$$

(gain ratio impurity)

Использование нескольких признаков



Практика

Задача

Вычислить наилучшее бинарное разделение корневого узла по одному признаку, пользуясь gini impurity.

№	Пол	Образование	Работа	Косметика
1	М	Высшее	Да	Нет
2	М	Среднее	Нет	Нет
3	М	Нет	Да	Нет
4	М	Высшее	Нет	Да
1	Ж	Нет	Нет	Да
2	Ж	Высшее	Да	Да
3	Ж	Среднее	Да	Нет
4	Ж	Среднее	Нет	Да

Когда остановить разделение

Split stopping criteria

- ▶ никогда
- ▶ использовать валидационную выборку
- ▶ установить минимальный размер узла
- ▶ установить порог $\Delta i(N) > \beta$
- ▶ статистический подход

$$\chi^2 = \sum_{k=1}^2 \frac{(n_{kL} - \frac{N_L}{N} n_k)^2}{\frac{N_L}{N} n_k}$$

Укорачиваем дерево

Pruning (a.k.a. отрезание ветвей)

1. Растим “полное” дерево T_0
2. На каждом шаге заменяем самый “слабый” внутренний узел на лист

$$R_\alpha(T_k) = \text{err}(T_k) + \alpha \text{size}(T_k)$$

3. Для заданного α из получившейся последовательности

$$T_0 \succ T_1 \succ \dots \succ T_r$$

выбираем дерево T_k , минимизирующее $R_\alpha(T_k)$

Значение α выбирается на основании тестовой выборки или CV

Какой класс присвоить листьям

1. Простейший случай:
класс с максимальным количеством объектов
2. Дискриминативный случай:
вероятность $p(C_k|x)$

Вычислительная сложность

Выборка состоит из n объектов, описанных m признаками

Предположения

1. Узлы делятся примерно поровну
2. Дерево имеет $\log n$ уровней
3. Признаки бинарные

Обучение. Для узла с k обучающими объектами:

Вычисление impurity по одному признаку $O(k)$

Выбор разделяющего признака $O(mk)$

Итог: $O(mn) + 2O(m\frac{n}{2}) + 4O(m\frac{n}{4}) + \dots = O(mn \log n)$

Применение. $O(\log n)$

Отсутствующие значения

- ▶ Удалить объекты из выборки
- ▶ Использовать отсутствие как отдельную категорию
- ▶ Вычислять impurity, пропуская отсутствующие значения
- ▶ Surrogate splits: разделяем вторым признаком так, чтобы было максимально похоже на первичное разделение

Surrogate split

$$c_1 : \quad x_1 = \begin{pmatrix} 0 \\ 7 \\ 8 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 1 \\ 8 \\ 9 \end{pmatrix}, \quad x_3 = \begin{pmatrix} 2 \\ 9 \\ 0 \end{pmatrix}, \quad x_4 = \begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix}, \quad x_5 = \begin{pmatrix} 5 \\ 2 \\ 2 \end{pmatrix}$$

$$c_2 : \quad y_1 = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}, \quad y_2 = \begin{pmatrix} 6 \\ 0 \\ 4 \end{pmatrix}, \quad y_3 = \begin{pmatrix} 7 \\ 4 \\ 5 \end{pmatrix}, \quad y_4 = \begin{pmatrix} 8 \\ 5 \\ 6 \end{pmatrix}, \quad y_5 = \begin{pmatrix} 9 \\ 6 \\ 7 \end{pmatrix}$$

primary split



$x_1, x_2, x_3, x_4, x_5, y_1, y_2, y_3, y_4, y_5$

first surrogate split



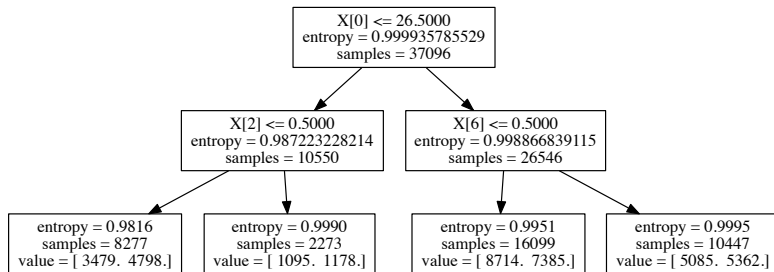
$x_3, x_4, x_5, y_1, y_2, y_3, y_4, y_5$
 x_1, x_2

*predictive association
 with primary split = 8*

Упражнение

Вычислить второй surrogate split

Задача о косметике



X_0 – возраст, X_4 – неоконченное высшее образование, X_6 – пол

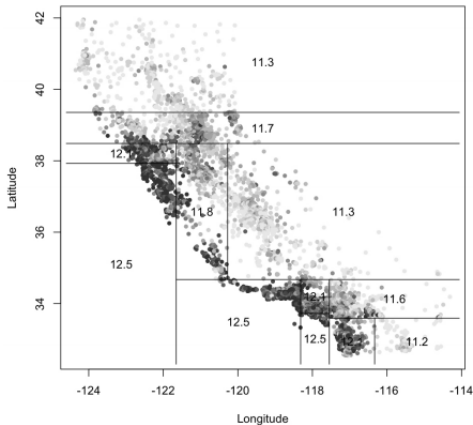
Задачи регрессии

Impurity узла N

$$i(N) = \sum_{y \in N} (y - \bar{y})^2$$

Присвоение класса листьям

- ▶ Среднее значение
- ▶ Линейная модель



Кроме CART

ID3 Iterative Dichotomiser 3

- ▶ Только номинальные признаки
- ▶ Количество детей в узле = количество значений разделяющего признака
- ▶ Дерево растет до максимальной высоты

C4.5 Улучшение ID3

- ▶ Числовые признаки – как в CART, номинальные – как в ID3
- ▶ При отсутствии значения используются **все** дети
- ▶ Укорачивает дерево, убирая ненужные предикаты в правилах

C5.0 Улучшение C4.5

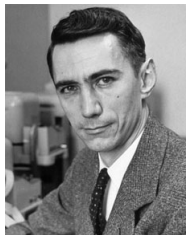
- ▶ Проприетарный

Решающие деревья. Итог

- + Легко интерпретируемы. Визуализация (ня!)
- + Любые входные данные
- + Мультикласс из коробки
- + Предсказание за $O(\log n)$
- + Поддаются статистическому анализу
- Склонны к переобучению
- Жадные и нестабильные
- Плохо работают при дисбалансе классов

Ключевые фигуры

- ▶ Claude Elwood Shannon
(Теория информации)
- ▶ Leo Breiman
(CART, RF)
- ▶ John Ross Quinlan
(ID3, C4.5, C5.0)



Вопросы

