Введение в Data Science
Занятие 3. Модели, основанные на правилах

Николай Анохин Михаил Фирулик

16 марта 2014 г.

План занятия

Деревья решений

Задача

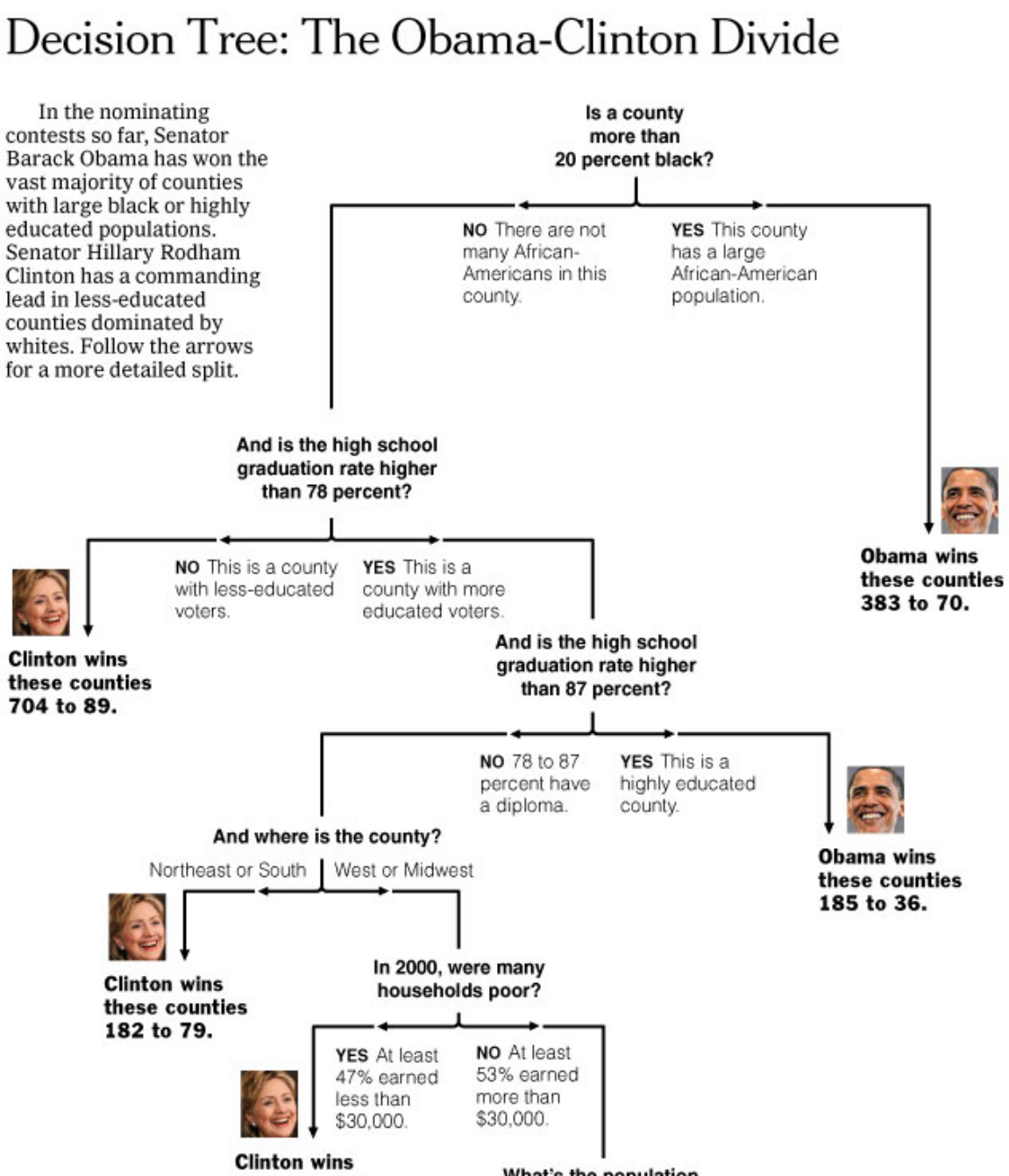
Дано:
обучающая выборка из профилей нескольких десятков тысяч человек

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

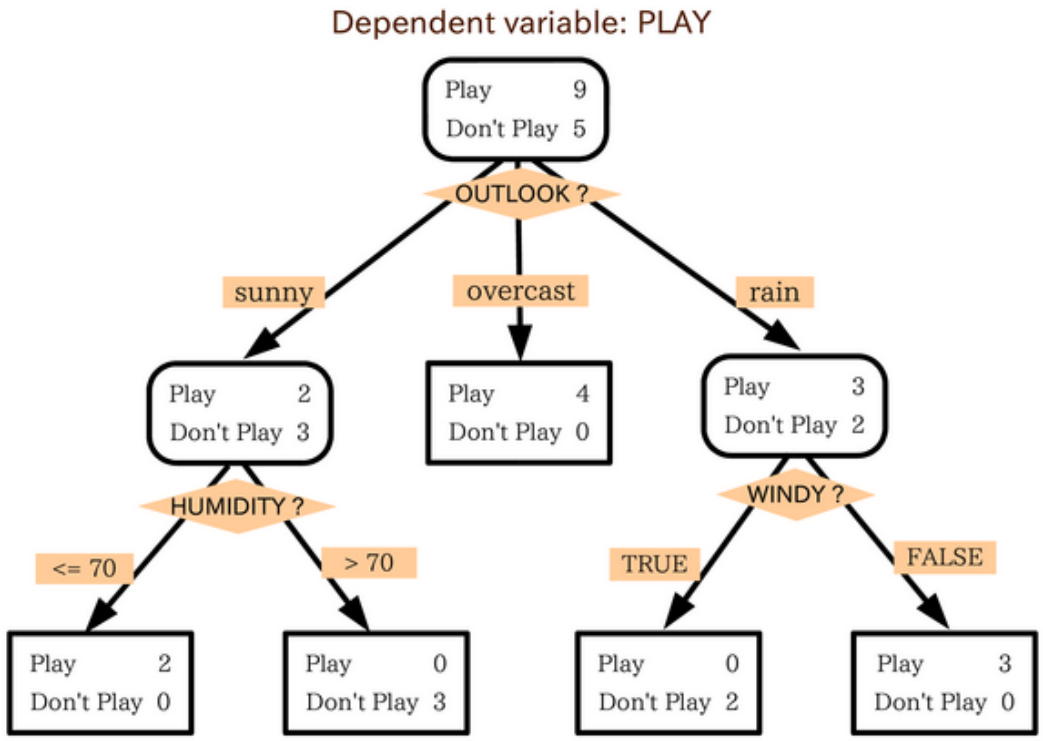
Задача:
Для рекламной кампании определить, характеристики людей, интересующихся косметикой

ТЕХНОСФЕРА @mail.ru

Обама или Клинтон?

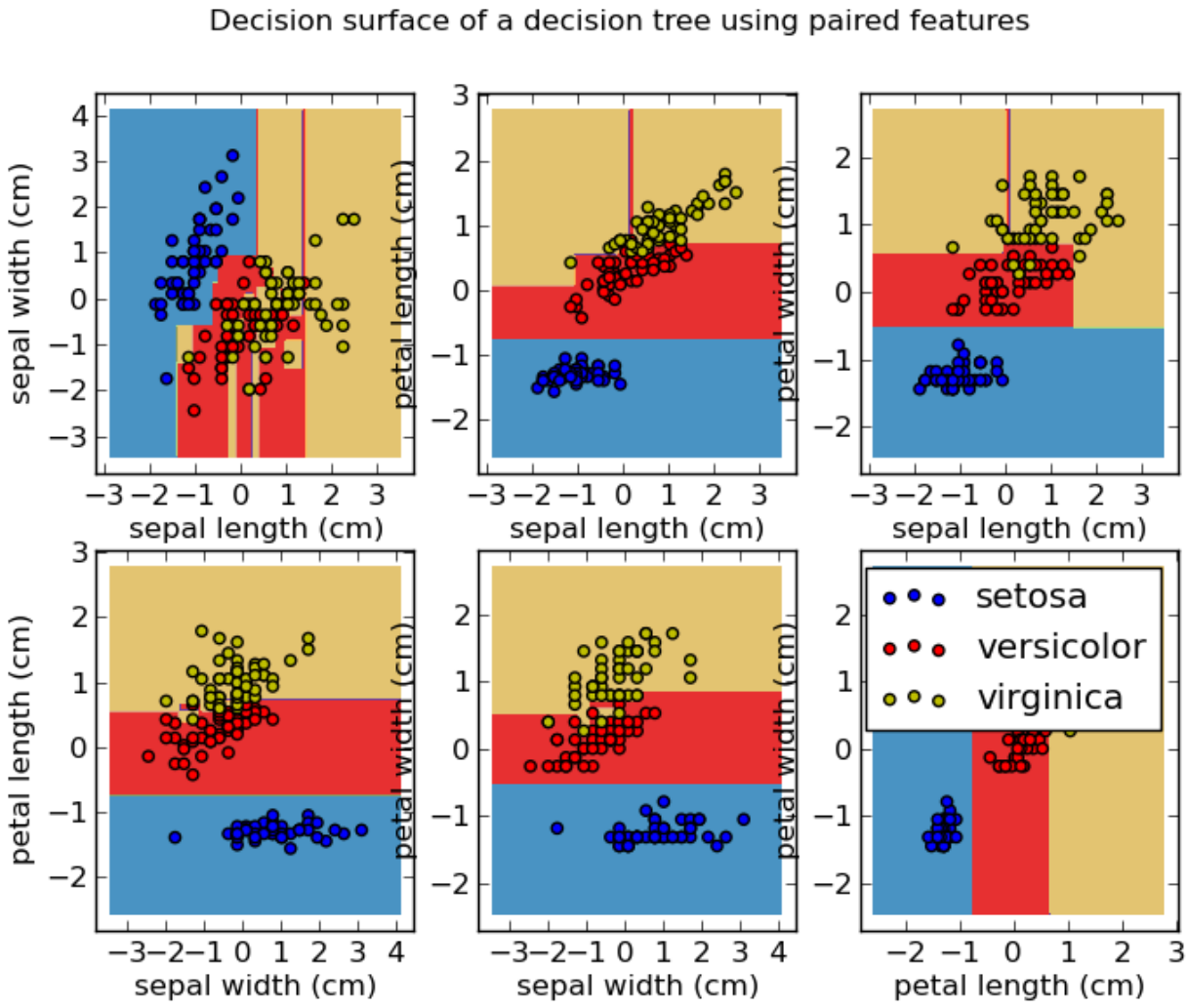


Хороший день для партии в гольф



ТЕХНОСФЕРА @mail.ru

Регионы принятия решений



ТЕХНОСФЕРА @mail.ru

Рекурсивный алгоритм

```
decision_tree(X_N):
  если X_N удовлетворяет критерию листа:
    создаем текущий узел N как лист
    выбираем подходящий класс C_N
  иначе:
    создаем текущий узел N как внутренний
    разбиваем X_N на подвыборки
    для каждой подвыборки X_j:
      n = decision_tree(X_j)
      добавляем n к N как ребенка
  возвращаем N
```

ТЕХНОСФЕРА @mail.ru

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

ТЕХНОСФЕРА @mail.ru

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 $H[y]$.

ТЕХНОСФЕРА @mail.ru

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

ТЕХНОСФЕРА @mail.ru

Если разделение не бинарное

Естественный выбор при разделении на B детей

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

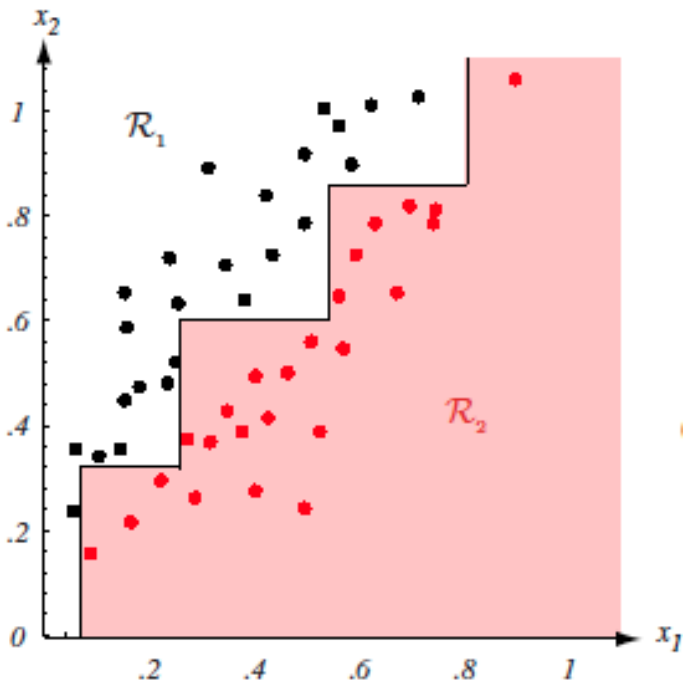
Предпочтение отдается большим B. Модификация:

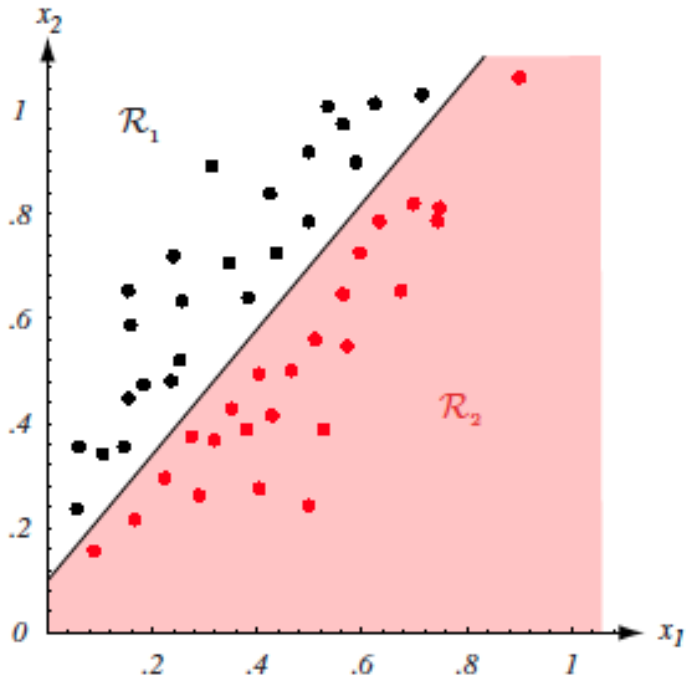
$$\Delta i_B(N, N_1, \ldots, N_B) = \frac{\Delta i(N, N_1, \ldots, N_B)}{-\sum_{k=1}^{B} \frac{N_k}{N} \log_2 \frac{N_k}{N}} \to \max$$

(gain ratio impurity)

ТЕХНОСФЕРА @mail.ru

Использование нескольких признаков





ТЕХНОСФЕРА @mail.ru

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

ТЕХНОСФЕРА @mail.ru

Когда остановить разделение

Split stopping criteria

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

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

ТЕХНОСФЕРА @mail.ru

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

$$R_\alpha(T_k) = err(T_k) + \alpha size(T_k)$$

3. Для заданного α из получившейся последовательности

$$T_0 \succ T_1 \succ \ldots \succ T_r$$

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

Значение α выбирается на основании тестовой выборки или CV

ТЕХНОСФЕРА @mail.ru

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

Применение. $O(\log n)$

ТЕХНОСФЕРА @mail.ru

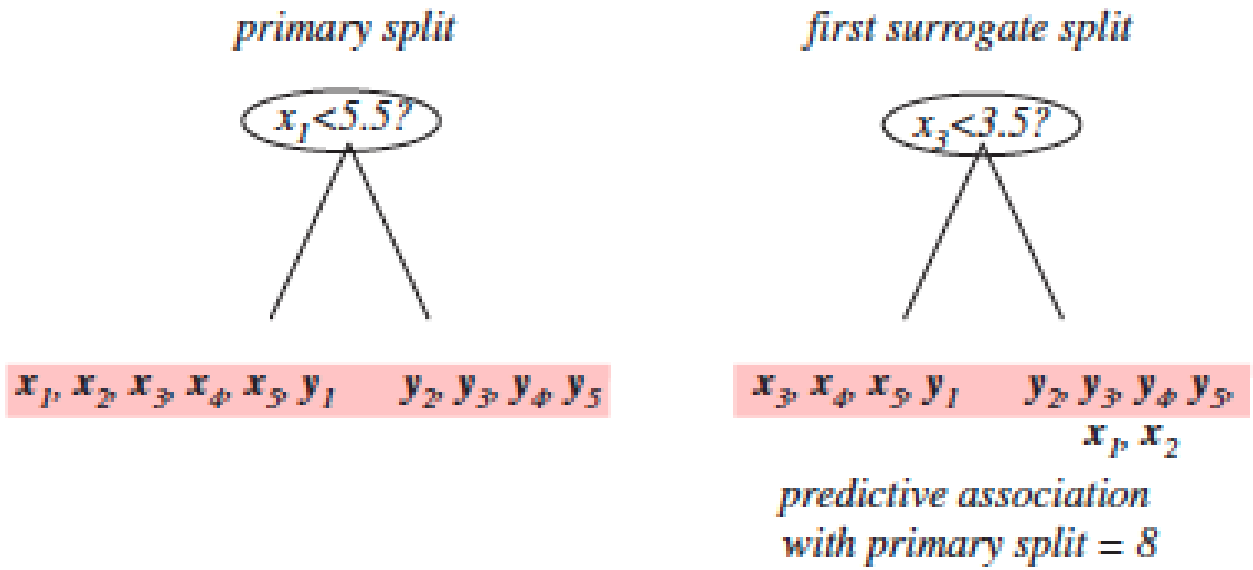
Отсутствующие значения

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

Surrogate split

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

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

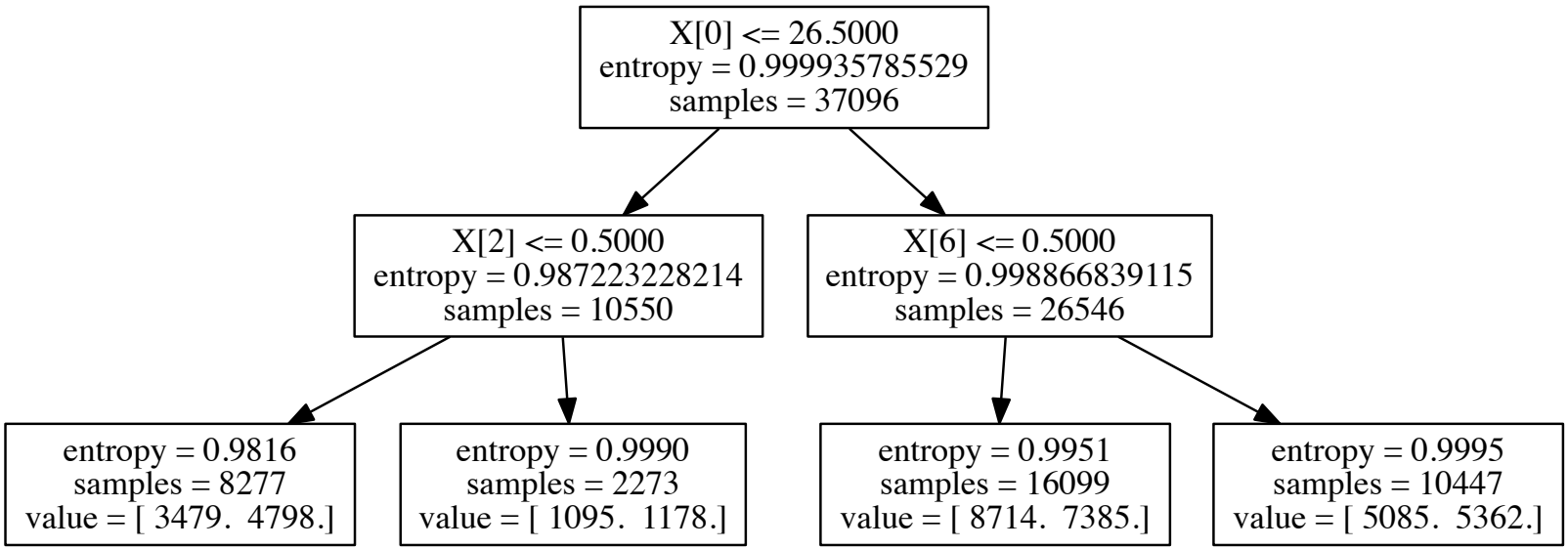


Упражнение

Вычислить второй surrogate split

ТЕХНОСФЕРА @mail.ru

Задача о косметике



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

ТЕХНОСФЕРА @mail.ru

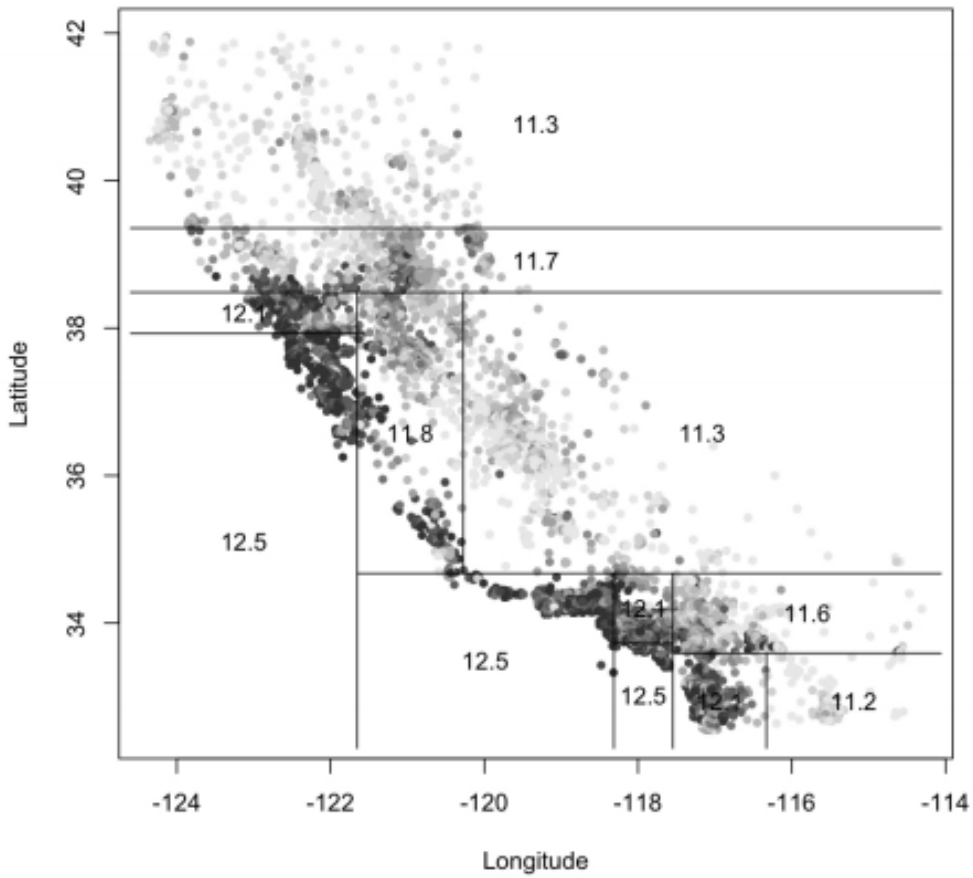
Задачи регрессии

Impurity узла N

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

Присвоение класса листьям

- Среднее значение
- Линейная модель



ТЕХНОСФЕРА @mail.ru

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

ТЕХНОСФЕРА @mail.ru

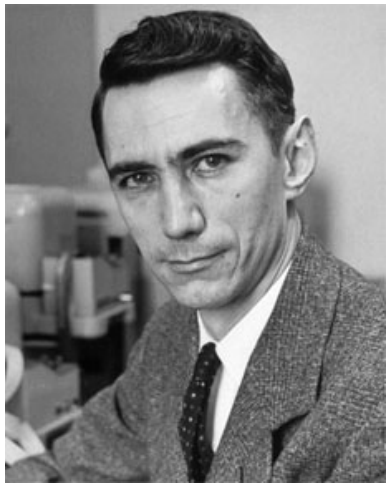
Решающие деревья. Итог

+ 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)



ТЕХНОСФЕРА @mail.ru

Другие модели, основанные на правилах

- Market Basket, Association Rules, A-Priori
- Logical inference, FOL

ТЕХНОСФЕРА @mail.ru

Заключение

"Binary Trees give an interesting and often illuminating way of looking at the data in classification or regression problems. They should not be used to the exclusion of other methods. We do not claim that they are always better. They do add a flexible nonparametric tool to the data analyst's arsenal."

–Breiman, Friedman, Olshen, Stone

Задача

Предсказать категорию семейного дохода на основании профилей пользователей с использованием дерева решений (имплементация из sklearn).

Метрика качества:

$$\mu = \frac{accuracy}{max_k P(C_k)}$$

Награда:
В ДЗ можно использовать любую готовую имплементацию DT

ТЕХНОСФЕРА @mail.ru

Домашнее задание 2

Деревья решений

Реализовать

- алгоритм CART для задачи регрессии
- алгоритм CART для задачи классификации

Поддержка: разные impurity, split stopping, pruning (+)

Ключевые даты

- До 2014/03/22 00.00 выбрать задачу и ответственного в группе
- До 2014/03/29 00.00 предоставить решение задания

Спасибо!

Обратная связь