ТЕХНОСФЕРА

Лекция 4
Задача классификации

Николай Анохин

15 октября 2014 г.

План занятия

Задачи классификации и регрессии

Подходы к моделированию

Теория принятия решений

Оценка результатов классификации

Деревья решений

Задачи классификации и регрессии

Классификация: интуиция

Задача

Разработать алгоритм, позволяющий определить класс произвольного объекта из некоторго множества

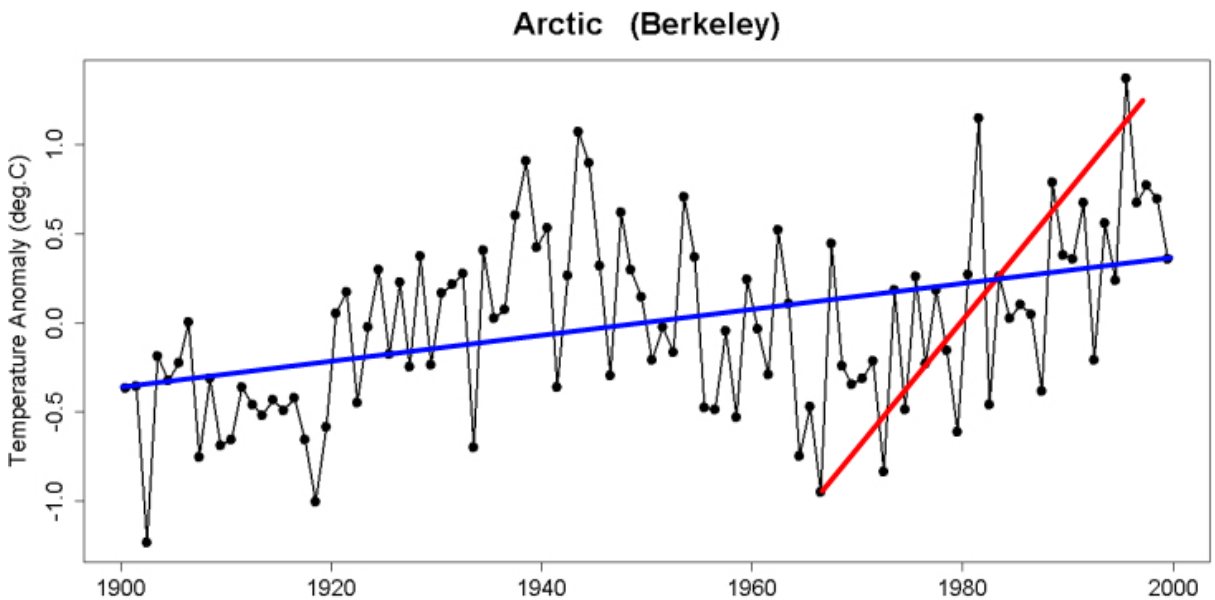
- Дана *обучающая выборка*, в которой для каждого объекта известен класс

Регрессия: интуиция

Задача

Разработать алгоритм, позволяющий предсказать числовую характеристику произвольного объекта из некоторого множества

- Дана *обучающая выборка*, в которой для каждого объекта известно значение числовой характеристики



Постановка задачи

Пусть дан набор объектов $\mathcal{D} = \{(\mathbf{x}_i, y_i)\},\ \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{Y},\ i \in 1, \ldots, N$, полученный из неизвестной закономерности $y = f(\mathbf{x})$. Необходимо выбрать из семейства параметрических функций

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathcal{Y}\}$$

такую $h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$, которая наиболее точно апроксимирует $f(\mathbf{x})$.

Задачи

- Классификация: $|\mathcal{Y}| < C$
- Регрессия: $\mathcal{Y} = [a, b] \subset \mathbb{R}$

Как решать

M Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathcal{Y}\},$$

которая могла бы решить нашу задачу (model selection)

L Выбираем наилучшие параметры модели θ^*, используя **алгоритм обучения**

$$A(X, Y) : (\mathcal{X}, \mathcal{Y})^N \to \Theta$$

(learning/inference)

D Используя полученную модель $h^*(x) = h(x, \theta^*)$, классифицируем неизвестные объекты (decision making)

Подходы к моделированию

Виды моделей

Генеративные модели. Смоделировать $p(x|C_k)$ и $p(C_k)$, применить теорему Байеса

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

и использовать $p(C_k|x)$ для принятия решения
(NB, Bayes Networks, MRF)

Дискриминативные модели. Смоделировать $p(C_k|x)$ и использовать ее для принятия решения
(Logistic Regression, Decision Trees)

Функции решения. Смоделировать напрямую $h^*(x) : \mathcal{X} \to \mathcal{Y}$
(Linear Models, Neural Networks)

Вероятностные модели VS Функции решения

- Отказ от классификации (reject option)
- Дисбаланс в выборке
- Ансамбли моделей
- Сильные предположения о природе данных
- Излишняя (вычислительная) сложность

Байесовский подход к моделированию

Идея. Вместо фиксированного, но неизвестного θ^* ищем апостериорное распределение $p(\theta|\mathcal{D})$
Дано. $p(y_i)$, $p(\theta)$, $p(\mathbf{x}|\theta)$

$$p(y_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|y_i,\mathcal{D})p(y_i|\mathcal{D})}{\sum_j p(\mathbf{x}|y_j,\mathcal{D})p(y_j|\mathcal{D})} = \frac{p(\mathbf{x}|y_i,\mathcal{D})p(y_i)}{\sum_j p(\mathbf{x}|y_j,\mathcal{D})p(y_j)}$$

$$p(\mathbf{x}|y_i,\mathcal{D}) = \int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})d\theta$$

Апостериорное распределение

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta} = \frac{\prod_n p(\mathbf{x}_n|\theta)p(\theta)}{\int \prod_n p(\mathbf{x}_n|\theta)p(\theta)d\theta}$$

Обучение модели

$$\textit{LEARNING} = \textit{representation} + \textit{evaluation} + \textit{optimization}$$

Pedro Domingos

Evaluation – критерий, который оптимизируем

- эмпирический риск $\to$ min
- KL-дивергенция $\to$ min
- функция правдоподобия $\to$ max
- information gain $\to$ max

Optimization – как оптимизируем

- unconstrained (GD, Newton+)
- constrained (linear programming, quadratic programming)

Эмпирический риск

Функция потерь $\mathcal{L}(\mathbf{x}, y, \theta)$ - ошибка, которую для данного $\mathbf{x}$ дает модель $h(\mathbf{x}, \theta)$ по сравнению с реальным значением y

Эмпирический риск – средняя ошибка на обучающей выборке

$$Q(X, Y, \theta) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathbf{x}_n, y_n, \theta)$$

Задача – найти значение θ^*, минимизирующее эмпирический риск

$$\theta^* = \theta^*(X, Y) = \text{argmin}_\theta Q(X, Y, \theta)$$

Некоторые функции потерь

- Индикатор ошибки

$$\mathcal{L}(\mathbf{x}, y, \theta) = 0 \text{ if } h(\mathbf{x}, \theta) = y \text{ else } 1$$

- Функция Минковского

$$\mathcal{L}(\mathbf{x}, y, \theta) = |y - h(\mathbf{x}, \theta)|^q$$

Частные случаи: квадратичная $q = 2$, абсолютная ошибка $q = 1$

- Hinge

$$\mathcal{L}(\mathbf{x}, y, \theta) = \max(0, 1 - y \times h(\mathbf{x}, \theta))$$

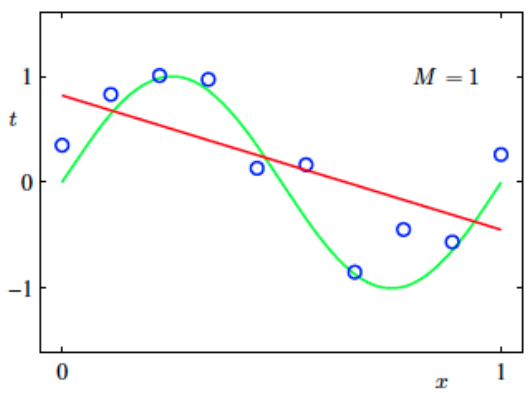
- Информационная

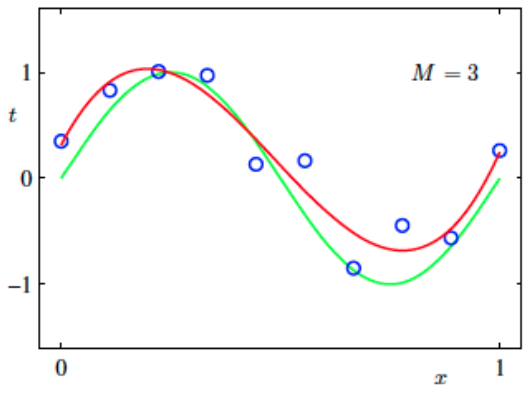
$$\mathcal{L}(\mathbf{x}, y, \theta) = -\log_2 p(y|\mathbf{x}, \theta)$$

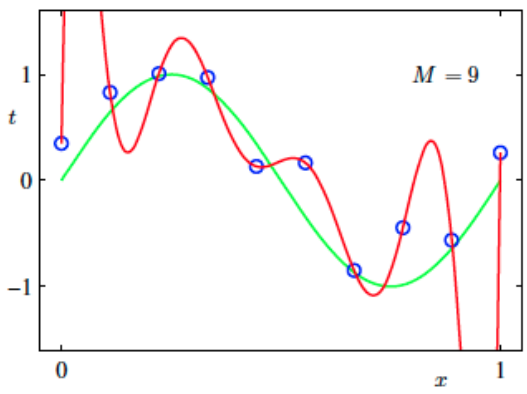
Проблема 1. Переобучение

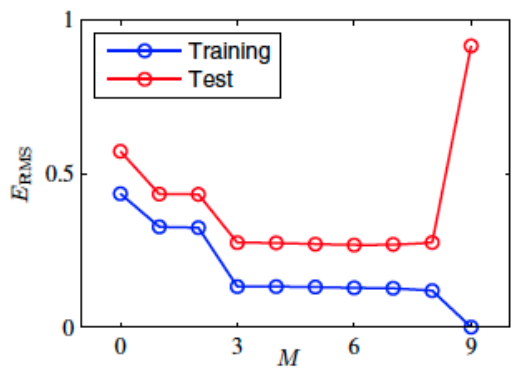
Задача

Аппроксимировать обучающую выборку полиномом M степени





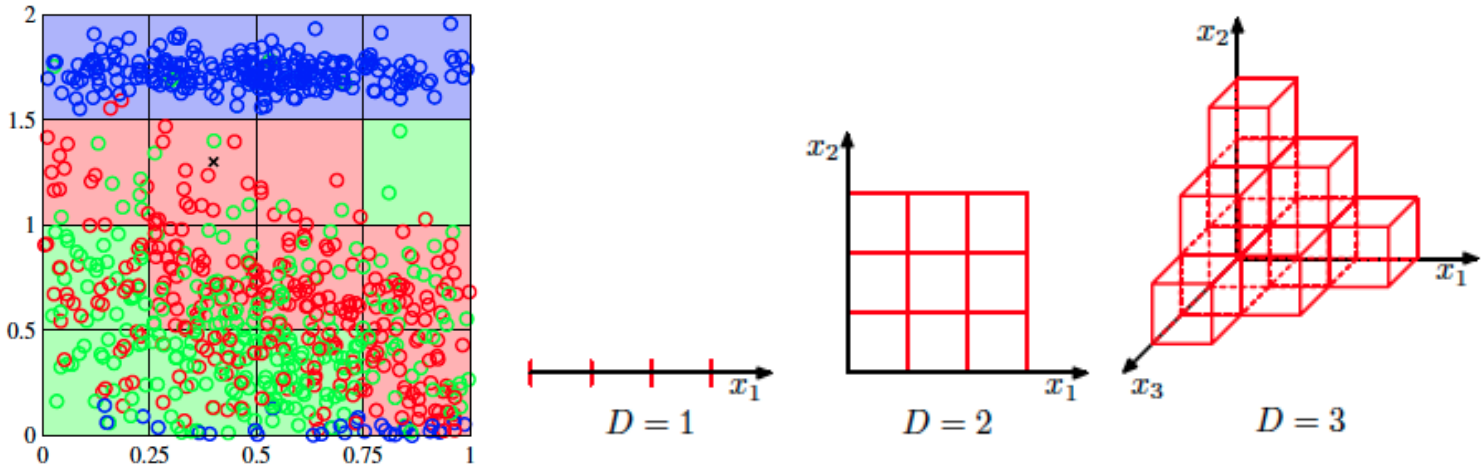




Проблема 2. Проклятие размерности

Задача

Классифицировать объекты.



Теория принятия решений

Классификация

Пусть
$\mathcal{R}_k$ – область, такая что все $\mathbf{x} \in \mathcal{R}_k$ относим к y_k
Дано
R_{kj} – риск, связанный с отнесением объекта класса y_k к классу y_j
Найти
$\forall k : \mathcal{R}_k$, такие, что математическое ожидание риска $E[R]$ минимально.

$$E[R] = \sum_k \sum_j \int_{\mathcal{R}_j} R_{kj} p(y_k|\mathbf{x}) p(\mathbf{x}) dx$$

Медицинская диагностика

Матрица риска $[R_{kj}]$

	sick	normal
sick	0	10
normal	1	0

Условные вероятности $p(y_k|x)$

$$p(\texttt{normal}|\texttt{moving}) = 0.9,\ p(\texttt{normal}|\texttt{not moving}) = 0.3$$

Вероятности $p(x)$

$$p(\texttt{moving}) = 0.7$$

Требуется определить $\mathcal{R}_{\texttt{sick}}$, $\mathcal{R}_{\texttt{normal}}$

Регрессия

Те же виды моделей: **генеративные**, **дискриминативные**, **функция решения**

Задана функция риска

$$R(y, h(\mathbf{x}))$$

Математическое ожидание $E[R]$

$$E[R] = \int\int R(y, h(\mathbf{x}))p(\mathbf{x}, y)d\mathbf{x}dy$$

Для квадратичной функции риска $R(y, h(\mathbf{x})) = [y - h(\mathbf{x})]^2$

$$h(x) = E_y[h|\mathbf{x}] = \int yp(y|\mathbf{x})dy$$

Оценка результатов классификации

Как оценить различные модели?

Идея

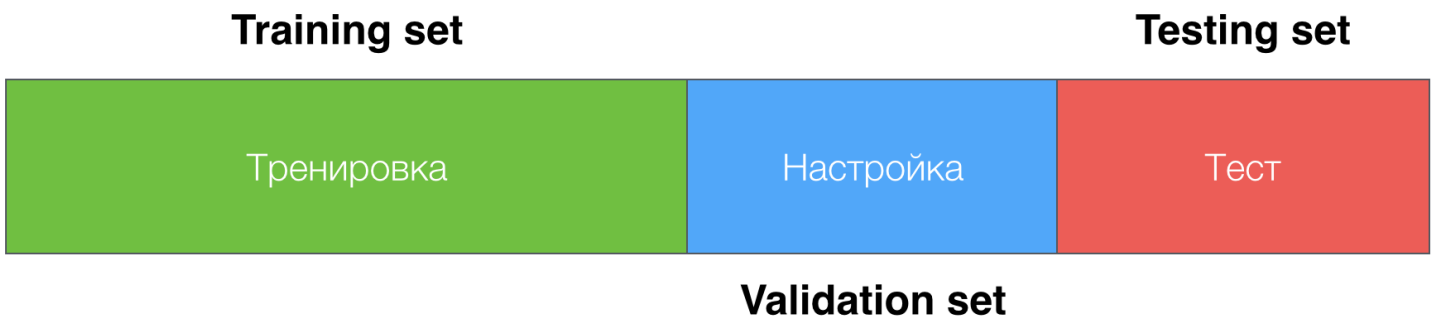
использовать долю неверно классифицированных объектов (error rate)

Важное замечание

error rate на обучающей выборке **НЕ** является хорошим показателем качества модели

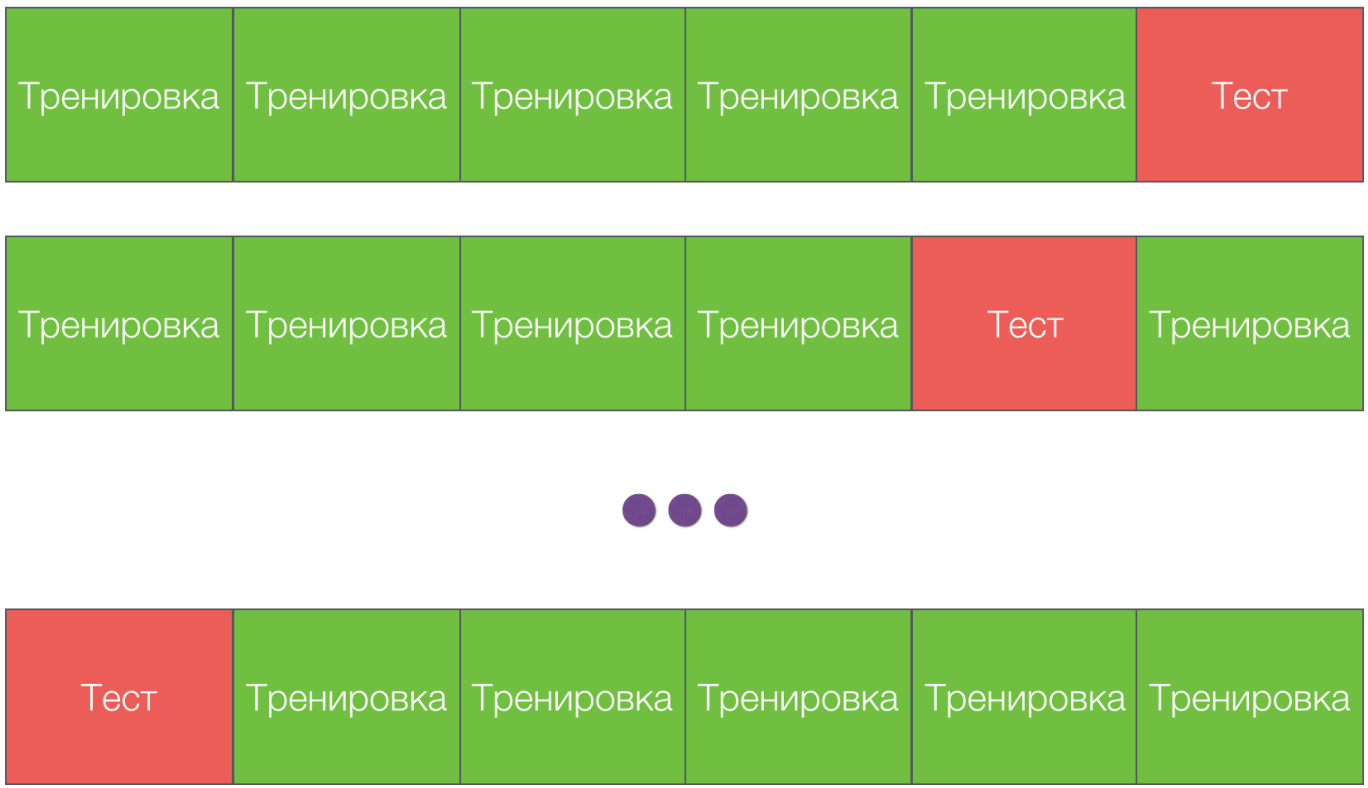
Решение 1: разделение выборки

Делим обучающую выборку на **тренировочную**, **валидационную** и **тестовую**



Решение 2: скользящий контроль

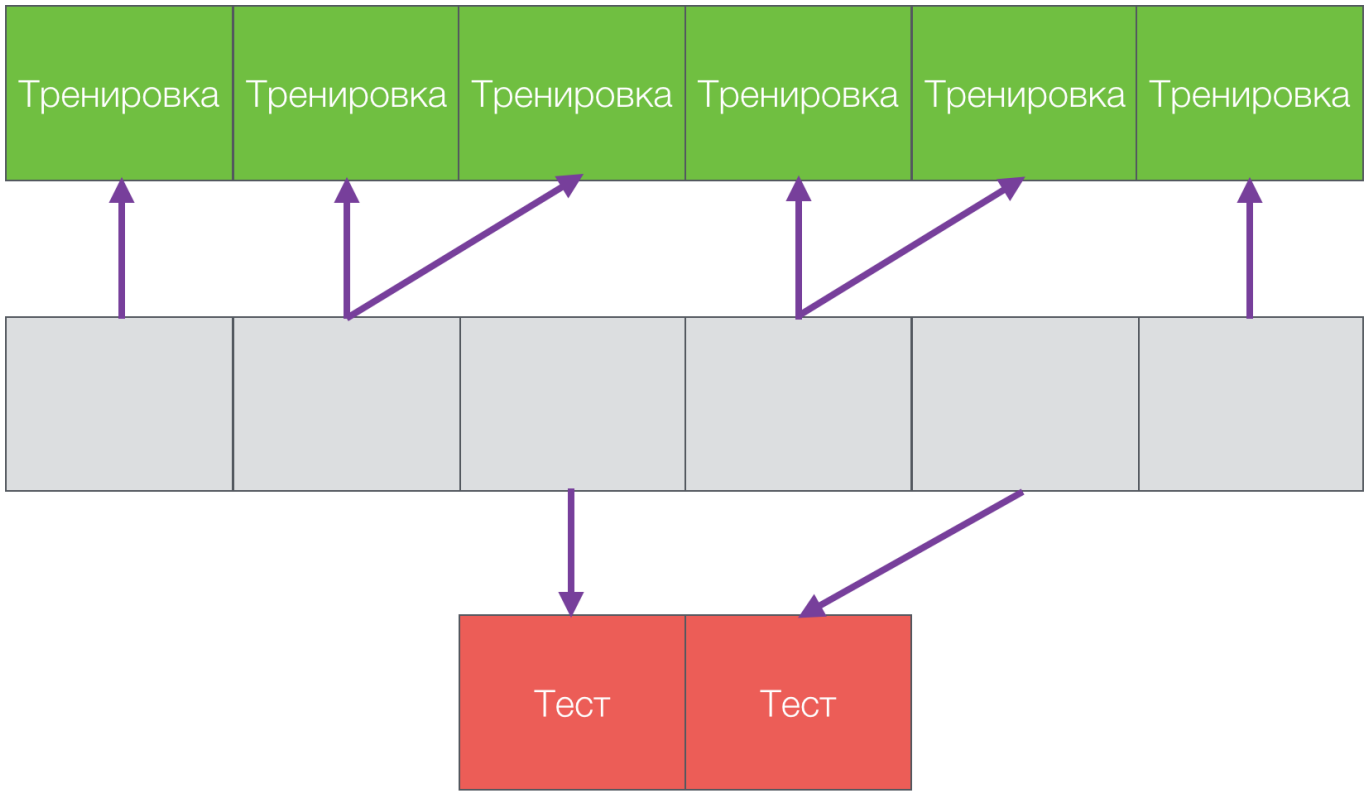
(n-times) (stratified) cross-validation



частный случай: leave-one-out

Решение 3: bootstrap

выбираем в тренировочную выбоку n объектов с возвращением



упражнение: найти математическое ожидание размера тестовой выборки.

Доверительный интервал для success rate

При тестировании на $N = 100$ объектах было получено 25 ошибок. Таким образом измеренная вероятность успеха (success rate) составила $f = 0.75$. Найти доверительный интервал для действительной вероятности успеха с уровнем доверия $\alpha = 0.8$.

Решение

Пусть p – действительная вероятность успеха в испытаниях бернулли, тогда

$$f \sim \mathcal{N}\left(p, p(1-p)/N\right).$$

Воспользовавшись табличным значением $P(-z \leq \mathcal{N}(0,1) \leq z) = \alpha$, имеем

$$P\left(-z \leq \frac{f-p}{\sqrt{p(1-p)/N}} \leq z\right) = \alpha,$$

откуда

$$p \in \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right) = [0.69, 0.80]$$

Метрики качества. Вероятностные модели.

Пусть y_i - действительный класс для объекта $\mathbf{x}_i$

- Information loss
$$-\frac{1}{N}\sum_i \log_2 p(y_i|\mathbf{x}_i)$$
- Quadratic loss
$$\frac{1}{N}\sum_j (p(y_j|\mathbf{x}_i) - a_j(\mathbf{x}_i))^2,$$

  где
$$a_j(\mathbf{x}_i) = \begin{cases} 1, \text{ если } C_j = y_i \\ 0, \text{ иначе} \end{cases}$$

Метрики качества. Функции решения.

		Предсказанный	
		true	**false**
Действительный	**true**	TP	FN
	false	FP	TN

$$success\ rate = accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

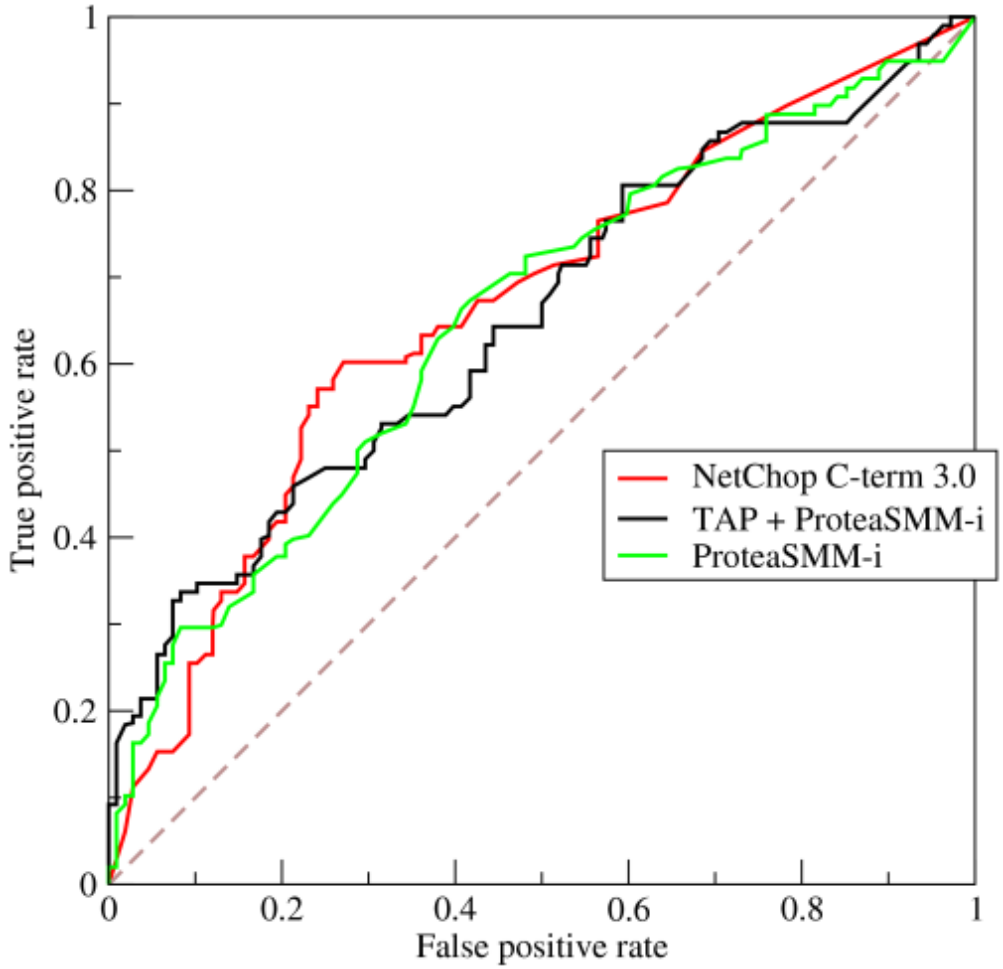
$$recall = TPR = \frac{TP}{TP + FN}; \quad precision = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{FP + TN}$$

$$affinity = lift = \frac{accuracy}{p}$$

Receiver Operating Characteristic

$$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN}$$



Упражнение

Простые классификаторы

В генеральной совокупности существуют объекты 3 классов, вероятность появления которых $p_1 < p_2 < p_3$. Первый классификатор относит все объекты к классу с большей вероятностью (то есть к третьему). Второй классификатор случайно относит объект к одному из классов в соответствии с базовым распределением. Рассчитать precision и recall, которые эти классификаторы дают для каждого из 3 классов.

Метрики качества. Регрессия

$$MSE = \frac{1}{N}\sum(h(\mathbf{x}_i) - y_i)^2, \quad RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{N}\sum|h(\mathbf{x}_i) - y_i|, \quad RMAE = \sqrt{MAE}$$

$$RSE = \frac{\sum(h(\mathbf{x}_i) - y_i)^2}{\sum(y_i - \bar{y})^2}$$

$$correlation = \frac{S_{hy}}{\sqrt{S_h S_y}}; \quad S_{yh} = \frac{\sum(h(i) - \overline{h(i)})(y_i - \bar{y})}{N-1}$$

$$S_h = \frac{\sum(h(i) - \overline{h(i)})^2}{N-1}; \quad S_y = \frac{\sum(y_i - \bar{y})^2}{N-1}$$

NFLT, MDL, AIC и все такое

No free lunch theorem

Не существует единственной лучшей модели, решающей все задачи

Minimum description length

Лучшая гипотеза о данных – та, которая ведет к самому краткому их описанию

Akaike information criterion (AIC)

$$model = \arg\max \ln p(\mathcal{D}|\theta_{ML}) - \|\theta\|$$

Деревья решений

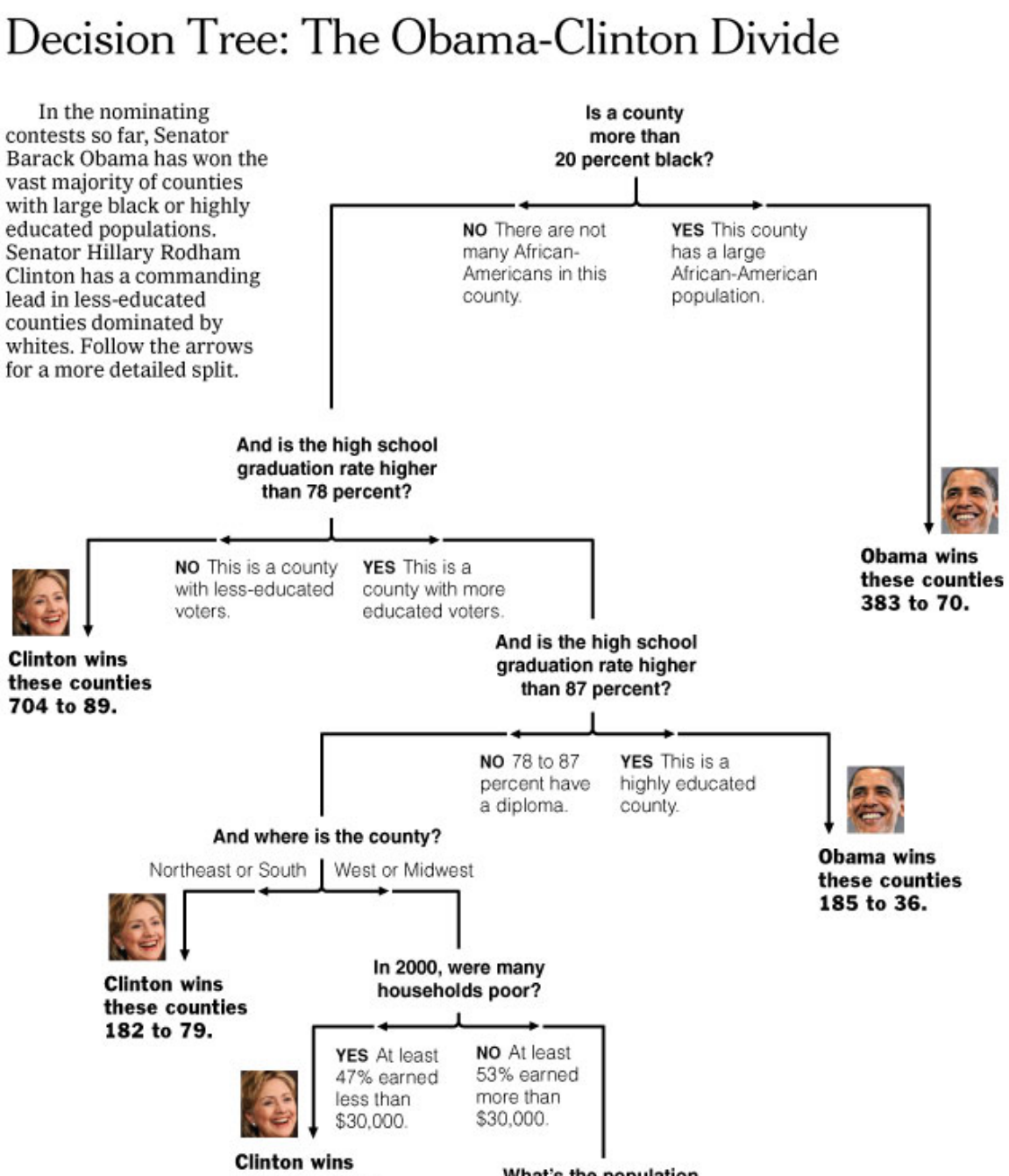
Задача

Дано:
обучающая выборка из профилей нескольких десятков тысяч человек

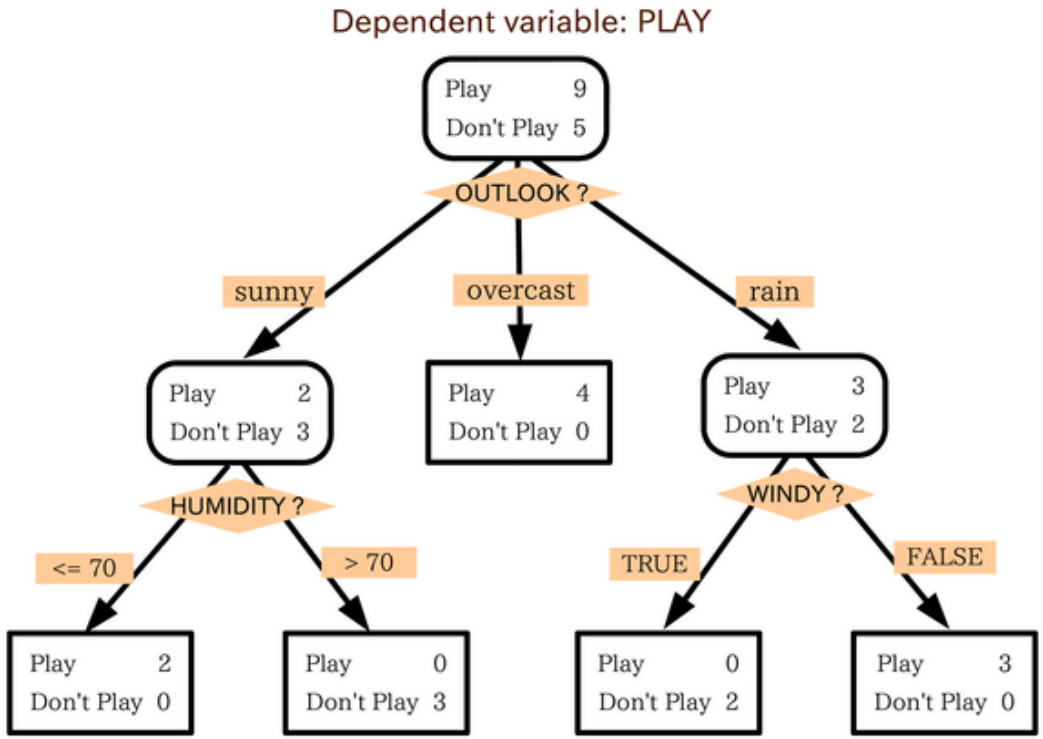
- пол (binary)
- возраст (numeric)
- образование (nominal)
- и еще 137 признаков
- наличие интереса к косметике

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

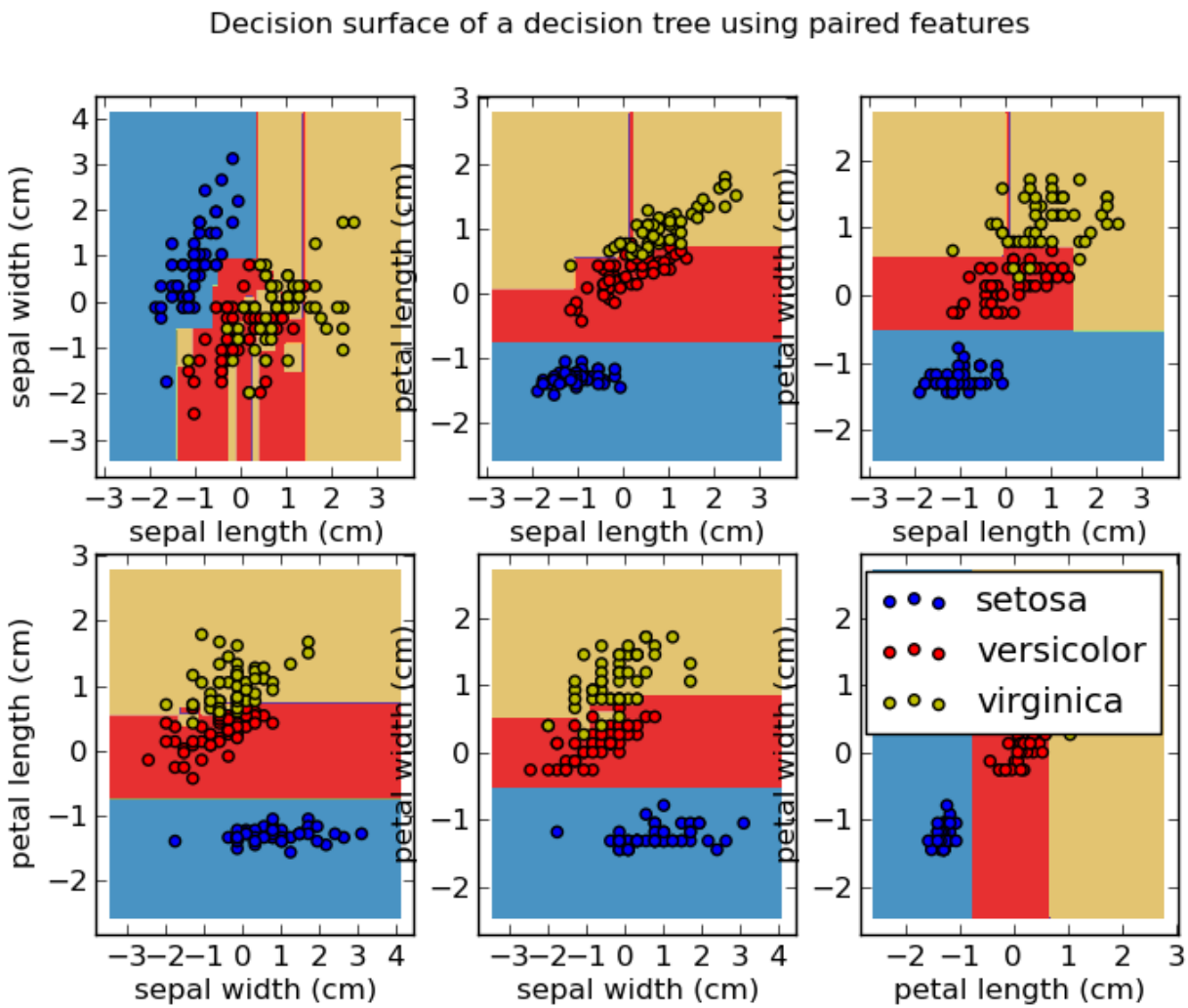
Обама или Клинтон?



Хороший день для партии в гольф



Регионы принятия решений



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

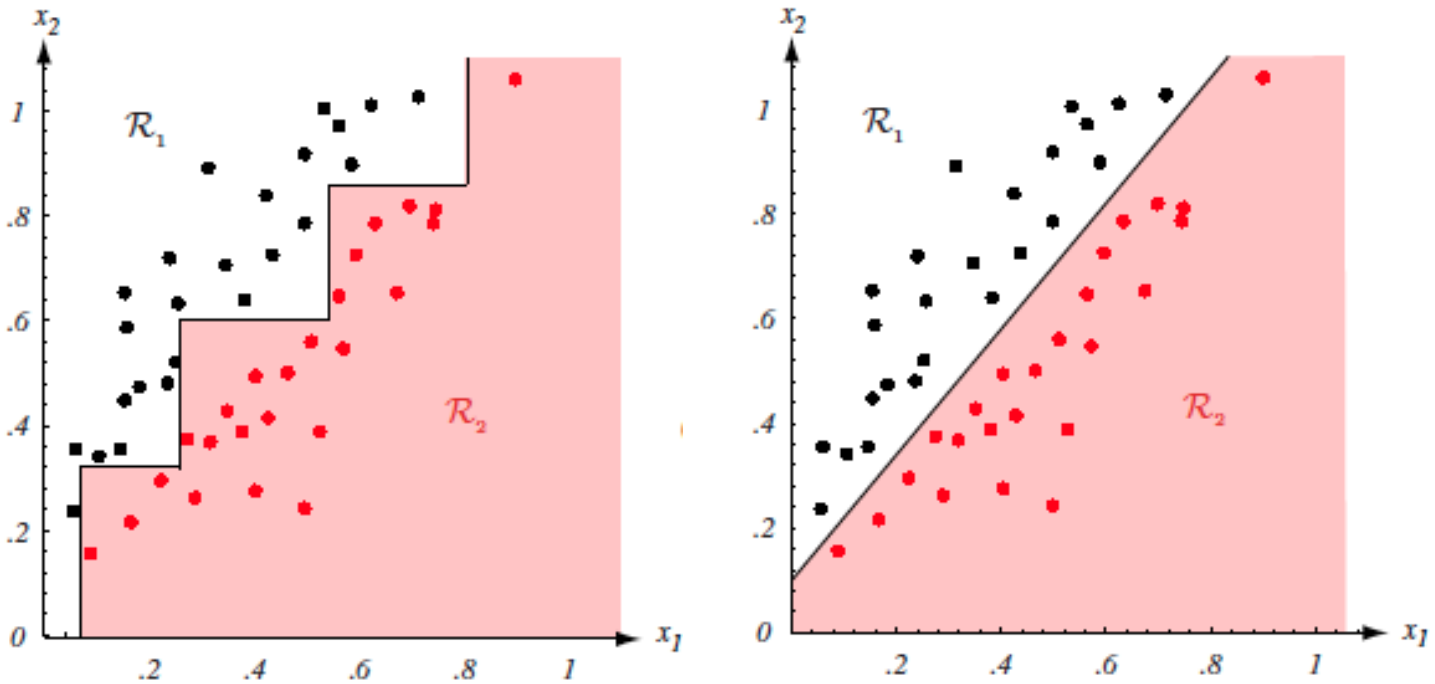
$$\Delta i(N, N_1, \ldots, N_B) = i(N) - \sum_{k=1}^{B} \frac{N_k}{N} i(N_k) \to \max$$

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)

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

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

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

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. $O(\log n)$

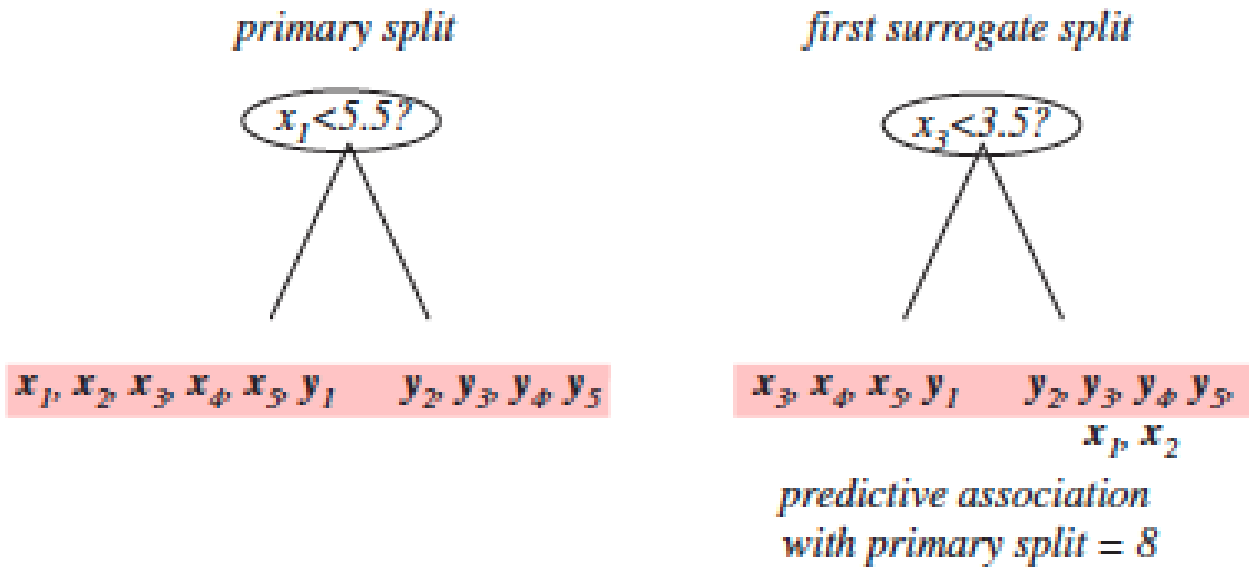
Отсутствующие значения

- 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

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

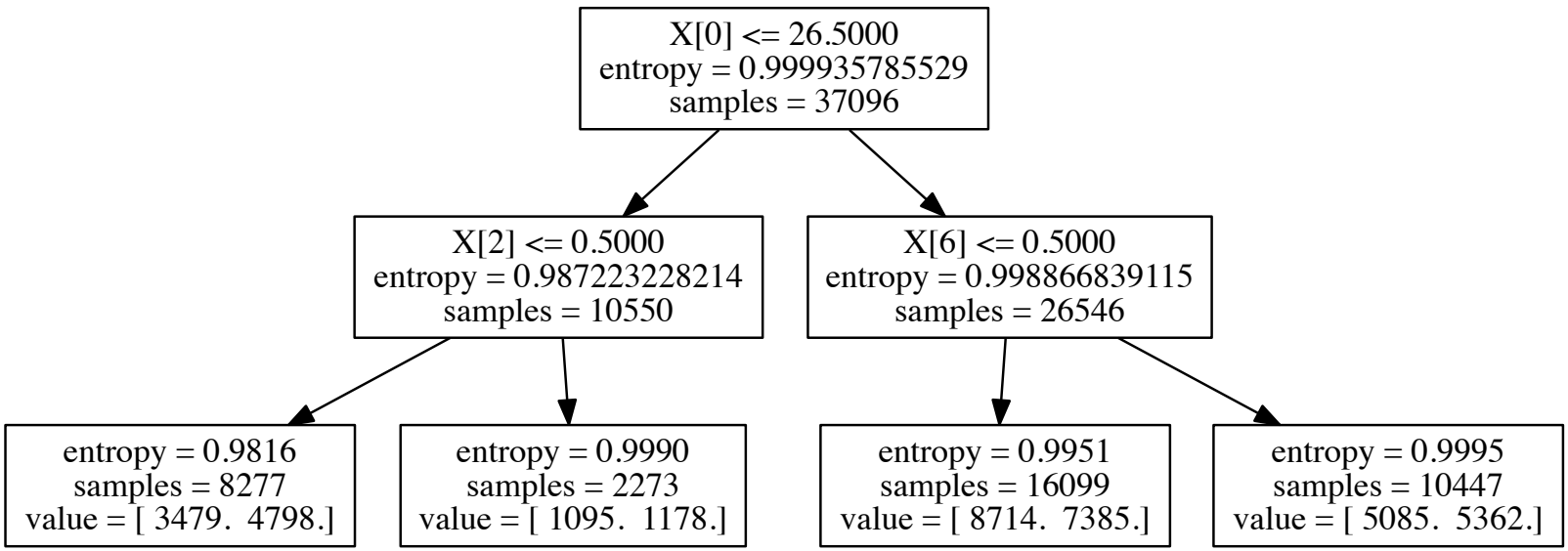
$$c_2: \quad y_1 = \begin{pmatrix}3\\3\\3\end{pmatrix}, \; y_2 = \begin{pmatrix}6\\0\\4\end{pmatrix}, \; y_3 = \begin{pmatrix}7\\4\\5\end{pmatrix}, \; y_4 = \begin{pmatrix}8\\5\\6\end{pmatrix}, \; y_5 = \begin{pmatrix}9\\6\\7\end{pmatrix}$$



Упражнение

Вычислить второй surrogate split

Задача о косметике



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

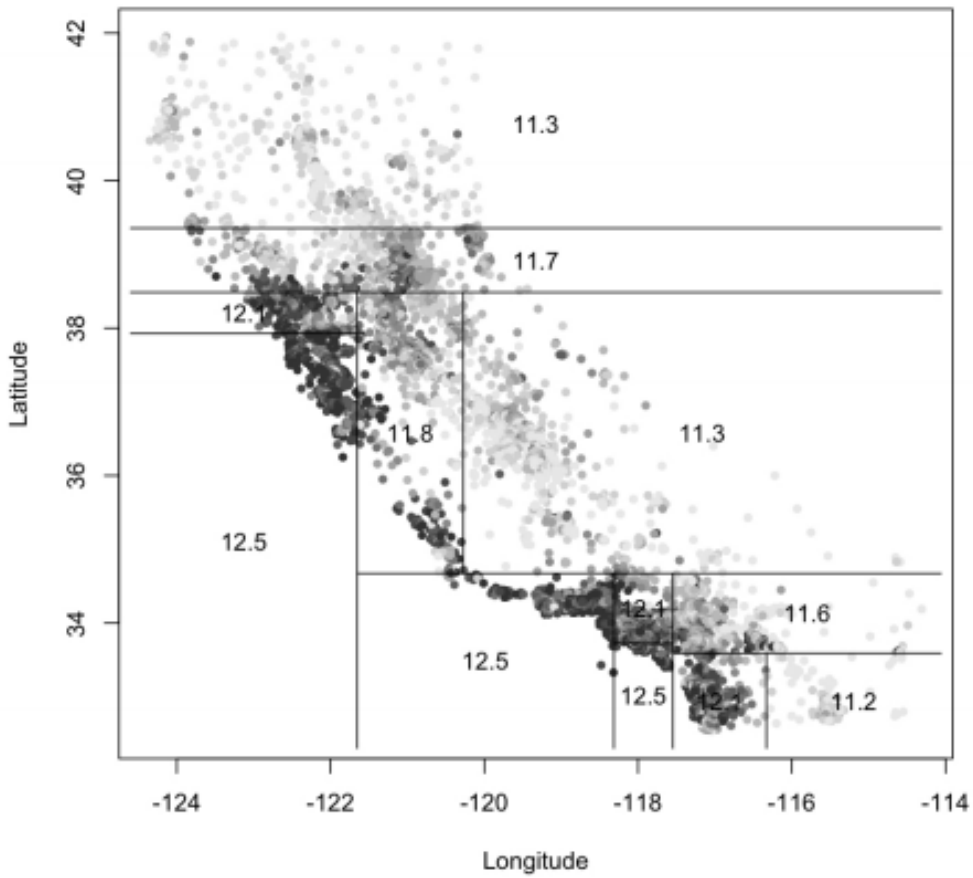
Задачи регрессии

Impurity узла N

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

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

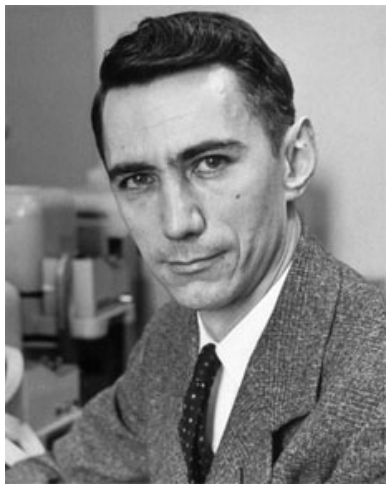
Решающие деревья. Итог

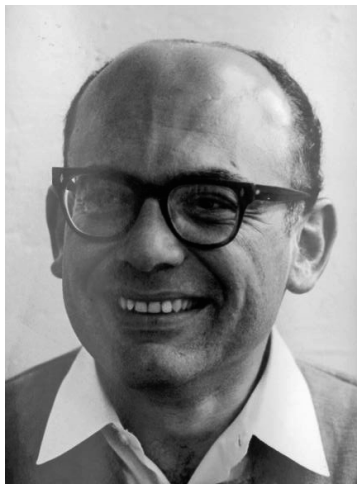
+ Легко интерпретируемы. Визуализация (ня!)
+ Любые входные данные
+ Мультикласс из коробки
+ Предсказание за $O(\log n)$
+ Поддаются статистическому анализу

– 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