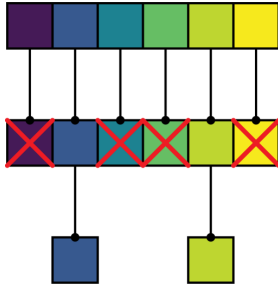


Отбор признаков

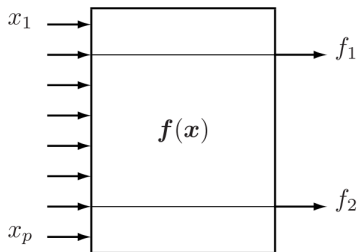
Виктор Китов

v.v.kitov@yandex.ru

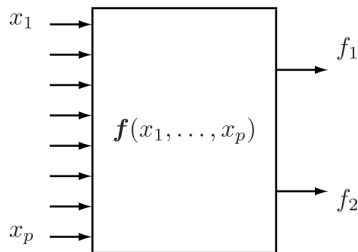


Задача отбора признаков

- Отбор признаков (feature selection) - выделение подмножества исходных признаков.
- Снижение размерности (dimensionality reduction) - преобразование исходных признаков в пространство меньшей размерности.



(a) feature selector



(b) feature extractor

Применения отбора признаков

- Применения отбора признаков:
 - ↑ точности прогнозов (убираем шумовые)
 - ↑ стабильности оценок параметров (например для лин. регрессии)
 - ↑ вычислительной эффективности
 - ↑ интерпретируемости моделей
 - ↓ стоимости сбора данных
- Некоторые методы умеют самостоятельно отбирать признаки:

Применения отбора признаков

- Применения отбора признаков:
 - \uparrow точности прогнозов (убираем шумовые)
 - \uparrow стабильности оценок параметров (например для лин. регрессии)
 - \uparrow вычислительной эффективности
 - \uparrow интерпретируемости моделей
 - \downarrow стоимости сбора данных
- Некоторые методы умеют самостоятельно отбирать признаки:
 - решающие деревья и их ансамбли (бэггинг, RF, ERT, бустинг)
 - линейная/нелинейная регрессия/классификация с L_1 регуляризацией
 - orthogonal matching pursuit регрессия

Типы признаков¹

f -признак, $G = \{f_1, f_2, \dots, f_D\}$ -полный набор, $\tilde{G} = G \setminus \{f\}$.

- **Сильно релевантный признак:**

$$p(y|f, \tilde{G}) \neq p(y|\tilde{G})$$

- **Слабо релевантный признак:**

$$p(y|f, \tilde{G}) = p(y|\tilde{G}), \text{ но } \exists S \subset \tilde{G} : p(y|f, S) \neq p(y|S)$$

- **Нерелевантный признак:**

$$\forall S \subset \tilde{G} : p(y|f, S) = p(y|S)$$

¹Приведите примеры признаков каждого типа.

Типы признаков¹

f -признак, $G = \{f_1, f_2, \dots, f_D\}$ -полный набор, $\tilde{G} = G \setminus \{f\}$.

- **Сильно релевантный признак:**

$$p(y|f, \tilde{G}) \neq p(y|\tilde{G})$$

- **Слабо релевантный признак:**

$$p(y|f, \tilde{G}) = p(y|\tilde{G}), \text{ но } \exists S \subset \tilde{G} : p(y|f, S) \neq p(y|S)$$

- **Нерелевантный признак:**

$$\forall S \subset \tilde{G} : p(y|f, S) = p(y|S)$$

Цель отбора признаков

Найти минимальный $G' \subset G$ такой, что $P(y|G') \approx P(y|G)$, т.е. оставить только сильно релевантные и минимальный набор слабо релевантных признаков.

¹Приведите примеры признаков каждого типа.

Категоризация методов отбора признаков

Полнота перебора вариантов:

- Полный перебор: сложность $O(2^D)^2$
- Субоптимальный перебор: нет гарантии на глобальный оптимум
 - детерминированные
 - случайные (детерминированные со случайностью / полностью случайные)

Взаимосвязь с методом прогнозирования:

- независимые (filter methods)
- использующие метод прогнозирования и \mathcal{L} (wrapper methods)
- интегрированные в метод прогнозирования (embedded methods)

² метод ветвей и границ не перебирает все варианты (при некоторых предположениях на $S(U)$), но сложность все равно $O(2^D)$

Содержание

1 Расчет важности признаков

- Внешние оценки значимости признаков
- Оценки значимости признаков по модели

2 Методы поиска набора признаков

Расчет важности признаков

- Оценим значимости каждого признака $I(f_1), I(f_2), \dots, I(f_D)$.
- Далее можем:
 - отбирать признаки по значимости
 - учитывать все признаки, но в разной степени, в зависимости от $I(\cdot)^3$.

³Как контролировать вклад признаков в прогноз для K-NN, линейных моделей, случайного леса?

Отбор признаков по значимости

- Упорядочим признаки по значимости $I(f)$:

$$I(f_1) \geq I(f_2) \geq \dots \geq I(f_D)$$

- выбрать топ m

$$\hat{F} = \{f_1, f_2, \dots, f_m\}$$

- выбрать по порогу: $f_i : I(f_i) \geq \text{threshold}$
- выбрать лучший набор из:

$$U = \{\{f_1\}, \{f_1, f_2\}, \dots, \{f_1, f_2, \dots, f_D\}\}$$

$$\hat{F} = \arg \max_{F \in U} S(F)$$

- Комментарии:
 - легко реализовать, вычислительно простые методы
 - будет включено много слабо релевантных зависимых признаков

1 Расчет важности признаков

- Внешние оценки значимости признаков
- Оценки значимости признаков по модели

Корреляция

- Регрессия или бинарная классификация:

$$I(f) = \frac{\sum_i (f_i - \bar{f})(y_i - \bar{y})}{[\sum_i (f_i - \bar{f})^2 \sum_i (y_i - \bar{y})^2]^{1/2}} = \frac{a}{b}$$

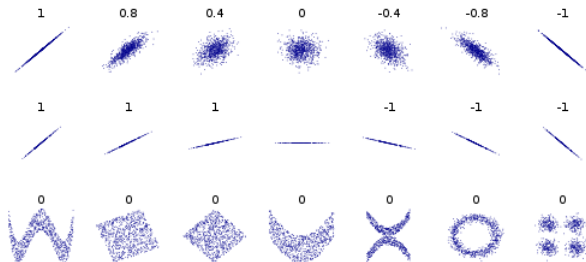
- Многоклассовая классификация:

$$I(f) = \frac{1}{C} \sum_{c=1}^C \frac{a_c}{b_c}, \quad I(f) = \max_c \left\{ \frac{a_c}{b_c} \right\}$$

- Корреляция:
 - легко вычисляется
 - выделяет только линейную зависимость

Корреляция выделяет только линейную зависимость

- Корреляция выделяет только линейную зависимость.



- Рассмотрим любую случайную величину X с симметричной (четной) плотностью распределения.
 - тогда $\mathbb{E}X = 0$, $\mathbb{E}X^3 = 0$
 - X и $Z = X^2$ зависимы, но $\text{corr}(X, Z) = 0$

Выделение монотонных зависимостей

- Рассмотрим $X = (X_1, X_2, \dots, X_N)$, $Y = (Y_1, Y_2, \dots, Y_N)$.
- Применим ранговое кодирование: $X \rightarrow R(X)$, $Y \rightarrow R(Y)$

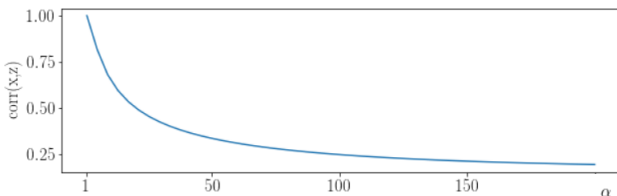
$IQ, X_i \blacktriangle$	Hours of TV per week, $Y_i \blacklozenge$	rank $x_i \blacklozenge$	rank $y_i \blacklozenge$
86	2	1	1
97	20	2	6
99	28	3	8
100	27	4	7
101	50	5	10
103	29	6	9
106	7	7	3
110	17	8	5
112	6	9	2
113	12	10	4

Ранговая корреляция Спирмена

- Ранговая корреляция Спирмена:

$$\text{corr}_{\text{Spearman}}(X, Y) = \text{corr}(R(X), R(Y))$$

- Рассмотрим $X = [0, 0.01, 0.02, \dots, 1]$, $Z = X^\alpha$.
- Существует монотонная зависимость между X и Z , но корреляция \downarrow при $\alpha \uparrow$:



- При этом

$$\text{corr}_{\text{Spearman}}(X, Y) = \text{corr}([1, 2, \dots], [1, 2, \dots]) = 1$$

Ранговая корреляция Кендалла

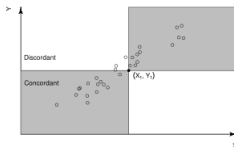
- Ранговая корреляция Кендалла:

- согласующиеся пары (concordant pairs)

$$C = \{[(X_i, Y_i), (X_j, Y_j)] : (X_j - X_i)(Y_j - Y_i) > 0\}$$

- несогласующиеся пары (discordant pairs)

$$D = \{[(X_i, Y_i), (X_j, Y_j)] : (X_j - X_i)(Y_j - Y_i) < 0\}$$



$$\text{corr}_{Kendall}(X, Y) = \frac{|C| - |D|}{\binom{N}{2}}$$

- Вместо самой корреляции можно судить о значимости признака по $p(\text{corr}(X, Y) = 0)$.

- это уровень значимости теста с $H_0 : \text{corr}(X, Y) = 0$

Определения

- **Энтропия** случайной величины Y :

$$H(Y) := - \sum_y p(y) \ln p(y)$$

- **Условная энтропия** Y при условии сл. вел. X :

$$H(Y|X) := - \sum_x p(x) \sum_y p(y|x) \ln p(y|x)$$

- **Расстояние Кульбака-Лейблера** между распределениями:

- дискретные исходы, $P(x)$, $Q(x)$ - вероятности исхода x :

$$KL(P||Q) := \sum_x P(x) \ln \frac{P(x)}{Q(x)}$$

- непрерывные исходы, $p(x)$, $q(x)$ - плотности вероятности:

$$KL(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$$

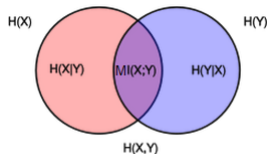
Взаимная информация

Взаимная информация измеряет насколько много общей информации между сл. вел. X и Y :

$$MI(X, Y) := \sum_{x,y} p(x, y) \ln \left[\frac{p(x, y)}{p(x)p(y)} \right] = KL(p(x, y) || p(x)p(y))$$

Свойства:

- $MI(X, Y) = MI(Y, X)$
- $MI(X, Y) = KL(p(x, y) || p(x)p(y)) \geq 0$
- X, Y - независимы $\Leftrightarrow MI(X, Y) = 0$
- $MI(X, Y) = H(Y) - H(Y|X)$
- $MI(X, Y) \leq \min \{H(X), H(Y)\}$
- X однозначно определяет $Y \Rightarrow MI(X, Y) = H(Y) \leq H(X)$



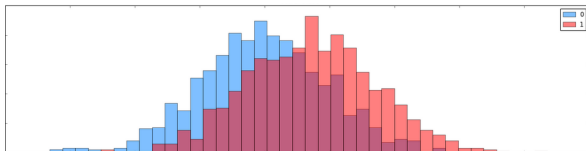
Нормированная взаимная информация

- Нормированная взаимная информация

$$NMI(X, Y) = \frac{MI(X, Y)}{H(Y)} \in [0, 1]$$

- $NMI(X, Y) = 0$ при независимости X и Y .
- $NMI(X, Y) = 1$, когда X однозначно определяет Y .
- Свойства MI и NMI :
 - выделяют зависимости любого вида
 - требуют оценки $p(X)$, $p(Y)$ и $p(X, Y)$.

Важность в задаче классификации



О взаимосвязи признака f и y можно судить по

$$\rho(p(f|y=i), p(f|y=j))$$

пример: $\int |p(x|y=1) - p(x|y=0)| dx$

Метрическая оценка $I(f)$: relief критерий для 1-NN

ВХОД:

Обучающая выборка $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$

Функция расстояния $\rho(x, x')$ # обычно Евклидова

для каждого объекта x_n, y_n :

найти ближайшего соседа $x_{s(n)}$ своего класса y_n

найти ближайшего соседа $x_{d(n)}$ чужого класса $\neq y_n$

для каждого признака $f_i \in \{f_1, f_2, \dots, f_D\}$:

рассчитать значимость $I(f_i) = \frac{1}{N} \sum_{n=1}^N \frac{|x_n^i - x_{d(n)}^i|}{|x_n^i - x_{s(n)}^i|}$

ВЫХОД:

значимости признаков $I(f_1), \dots, I(f_D)$

Метрическая оценка $I(f)$: relief критерий для K-NN

ВХОД:

Обучающая выборка $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$

Функция расстояния $\rho(x, x')$ # обычно Евклидова

Число соседей K

для каждого объекта x_n, y_n :

найти K ближайших соседей своего класса y_n :

$x_{s(n,1)}, x_{s(n,2)}, \dots x_{s(n,K)}$

найти K ближайших соседей чужого класса $\neq y_n$:

$x_{d(n,1)}, x_{d(n,2)}, \dots x_{d(n,K)}$

для каждого признака $f_i \in \{f_1, f_2, \dots f_D\}$:

рассчитать значимость $I(f_i) = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K \frac{|x_n^i - x_{d(n,k)}^i|}{|x_n^i - x_{s(n,k)}^i|}$

ВЫХОД:

значимости признаков $I(f_1), \dots I(f_D)$

1 Расчет важности признаков

- Внешние оценки значимости признаков
- Оценки значимости признаков по модели

Важность признаков по линейной модели

- В линейных моделях важность x^i можно считать по $|w_i|$.
 - при условии, что признаки приведены к единой шкале
 - `clf.coef_` в scikit-learn
- Учитывает линейную зависимость, как корреляция.

Важность признаков: mean decrease in impurity

- Важность признаков по изменению критерия информативности (mean decrease in impurity, MDI).
 - рассмотрим признак f
 - пусть $T(f)$ -множество всех вершин, использующих f в функции ветвления
 - эффективность разбиения в t :

$$\Delta\phi(t) = \phi(t) - \sum_{c \in \text{children}(t)} \frac{N(c)}{N(t)} \phi(c)$$

- значимость f :

$$\sum_{t \in T(f)} N(t) \Delta\phi(t)$$

- Поощряет признаки с большим количеством уникальных значений.

Важность признаков: mean decrease in impurity

В sklearn:

- важность рассчитывается методом *clf.feature_importances_*
 - доступен для композиций деревьев: RF, ERT, boosting.
- недостатки:
 - вычисляется на обучающей выборке
 - если модель переобучается на признаке, важность высока, но вклад в точность прогнозов мал.

Важность признаков: permutation feature importance

- Важность признаков по изменению критерия качества (permutation feature importance)
- Важность признака=разнице качества прогнозов на:
 - 1 исходной выборке
 - 2 исходной выборке, где значения j -го признака перемешаны

Важность признаков: permutation feature importance

- Важность признаков по изменению критерия качества (permutation feature importance)
- Важность признака = разнице качества прогнозов на:
 - ① исходной выборке
 - ② исходной выборке, где значения j -го признака перемешаны
- Значение рандомизированное, поэтому важно пересчитать несколько раз и усреднить.
- Показывает важность признака в разрезе
 - заданного критерия качества
 - заданной модели (не обязательно решающего дерева)
 - для плохой модели важный признак может оказаться неважным
 - поэтому нужно предварительно выбрать хорошую модель.

Важность признаков: permutation feature importance

- Можно считать по обучающей и валидационной выборке.
- Высокая важность на валидации=>признак усиливает обобщающую способность модели
- Высокая важность на обучении, но низкая на валидации=>на заданном признаке модель переобучается
- Если два признака скоррелированы, то при перемешивании одного признака модель имеет доступ к информации через другой
 - поэтому важность скоррелированных признаков занижена
 - важно исключать скоррелированные признаки.

Содержание

1 Расчет важности признаков

2 Методы поиска набора признаков

- Метод последовательной модификации набора признаков
- Лучевой поиск (beam search)
- Генетические алгоритмы

Поиск набора признаков

- Рассмотрим субоптимальные методы поиска подмножества признаков
 - вместо полного перебора со сложностью $O(2^D)$
- Пусть $S(U)$ -критерий качества набора признаков U .
 - например, точность модели на U
 - либо взвешенное сумма качества работы модели на U и штрафа за сложность $|U| = K$.
 - информационные критерии:

$$AIC = 2K - 2 \log P(Y|X)$$

$$BIC = K \ln N - 2 \log P(Y|X)$$

2 Методы поиска набора признаков

- Метод последовательной модификации набора признаков
- Лучевой поиск (beam search)
- Генетические алгоритмы

Метод последовательного включения признаков

- Метод последовательного включения признаков (sequential forward selection) реализует последовательное жадное добавление признаков один за другим, максимально увеличивающие $S(U)$.
- ВХОД:
 - максимальное #признаков K
 - критерий качества $S(U)$ для наборов признаков U
- ВЫХОД:
 - локально оптимальный набор U , $|U| \leq K$.

Метод последовательного включения признаков

Алгоритм жадного добавления признаков:

- инициализируем: $U = \{\}$
- пока $|U| \leq K - 1$:
 - $f^* = \arg \max_{f \in F \setminus S} S(U \cup \{f\})$
 - если $S(U \cup \{f^*\}) < S(U)$: выход
 - $U = U \cup \{f^*\}$
- вернуть U

Сложность $O(D |U|)$ без учета сложности расчета $S(U)$.

Модификации алгоритма

Модификации алгоритма:

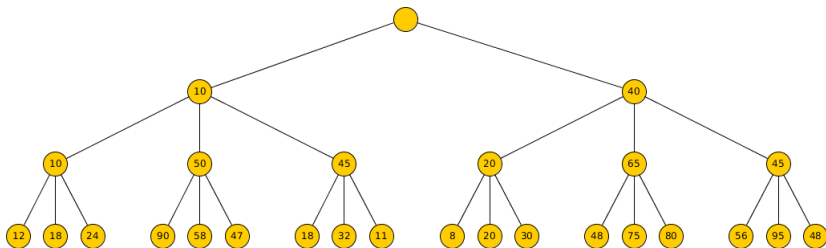
- последовательное исключение признаков (sequential backward selection)⁴
- последовательное включение лучшей группы из $\leq p$ признаков
- последовательное исключение худшей группы из $\leq p$ признаков
- композиция подходов добавления/удаления:
 - на каждом шаге пробовать удалить или добавить, что лучше (аналог GD)
 - на каждом шаге добавить, потом циклически удалять, пока приводит к $\uparrow S(U)$

⁴Что вычислительно эффективнее? Последовательное включения или исключения, если только 50% признаков релевантны?

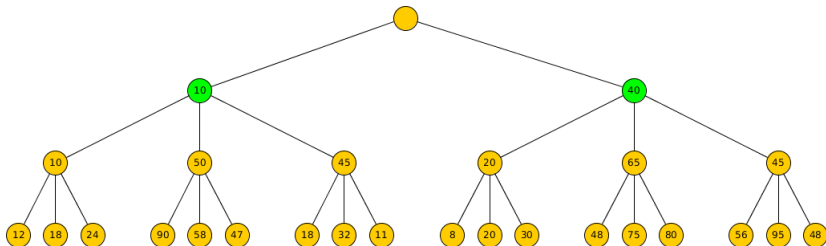
2 Методы поиска набора признаков

- Метод последовательной модификации набора признаков
- Лучевой поиск (beam search)
- Генетические алгоритмы

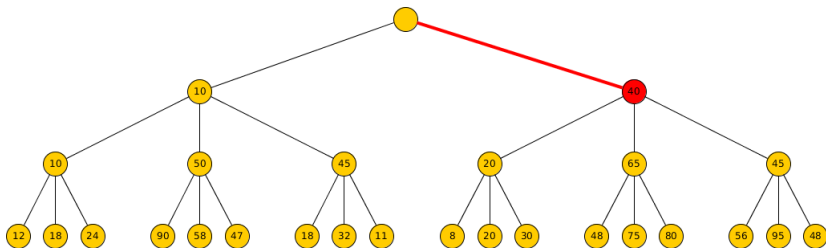
Жадный поиск (top-1)



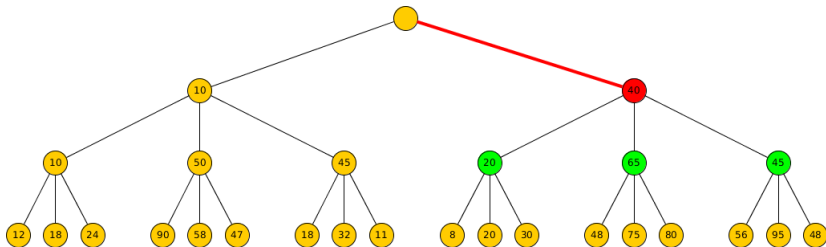
Жадный поиск (top-1)



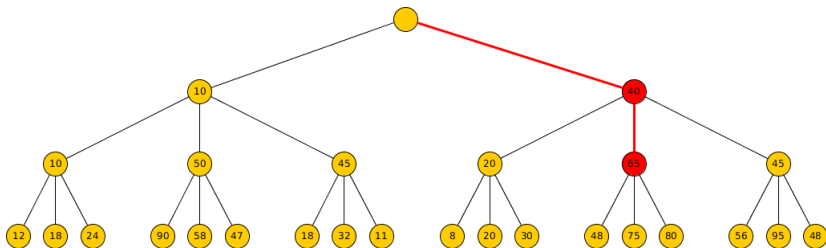
Жадный поиск (top-1)



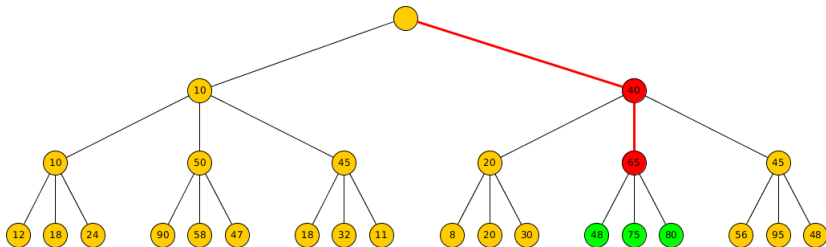
Жадный поиск (top-1)



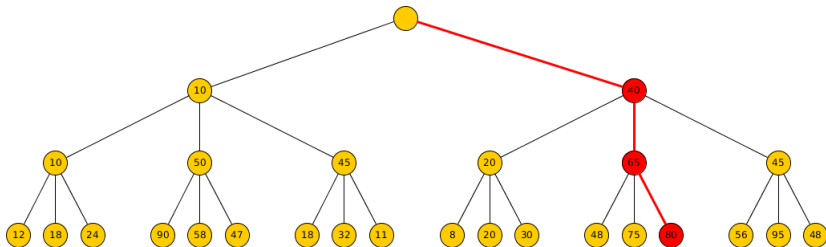
Жадный поиск (top-1)



Жадный поиск (top-1)



Жадный поиск (top-1)



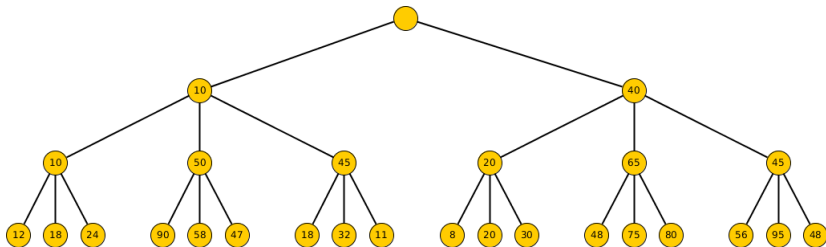
Лучевой поиск

- Лучевой поиск (beam search): при последовательном добавлении будем сохранять не один, а K лучших вариантов.
 - реализует жадный поиск в ширину (breadth first)
- Аналогично возможны обобщения последовательного исключения.

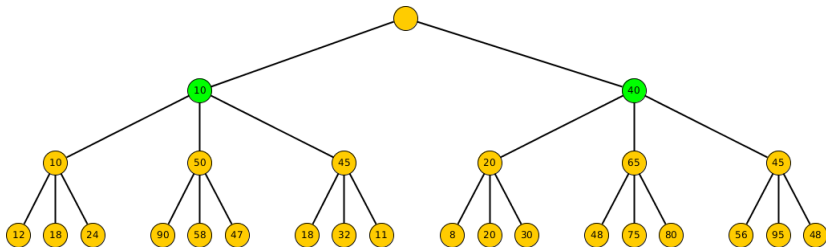
Принцип неоконченных решений Габора

Принимая решение, следует оставлять свободу выбора последующих решений.

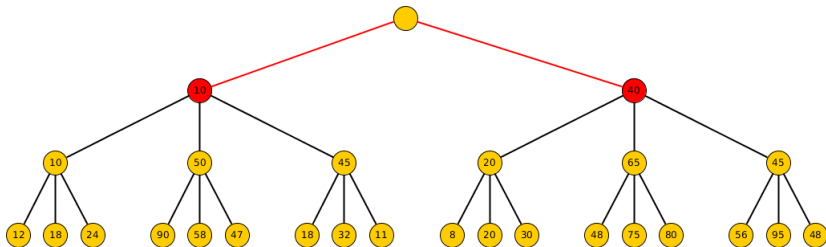
Лучевой поиск: $K = 2$



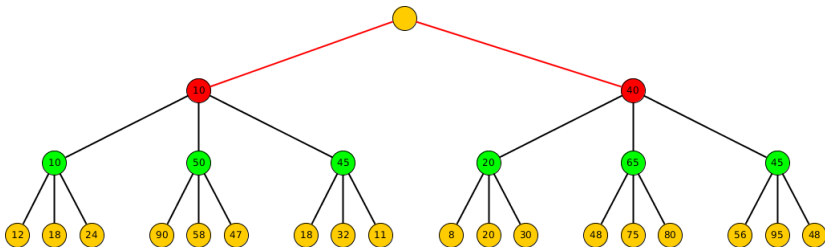
Лучевой поиск: $K = 2$



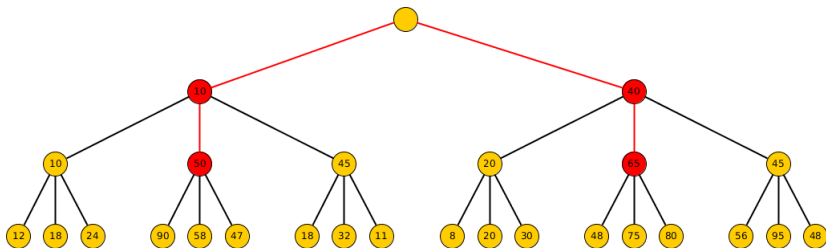
Лучевой поиск: $K = 2$



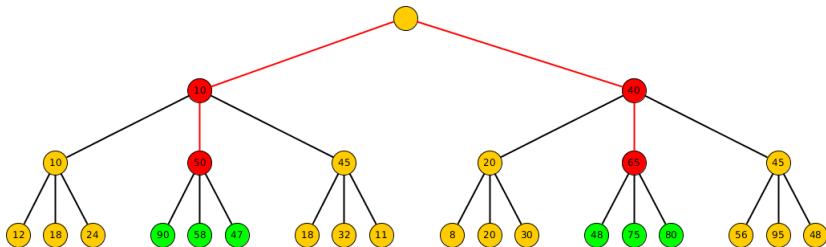
Лучевой поиск: $K = 2$



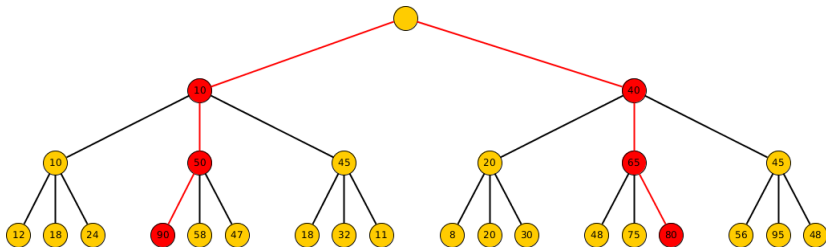
Лучевой поиск: $K = 2$

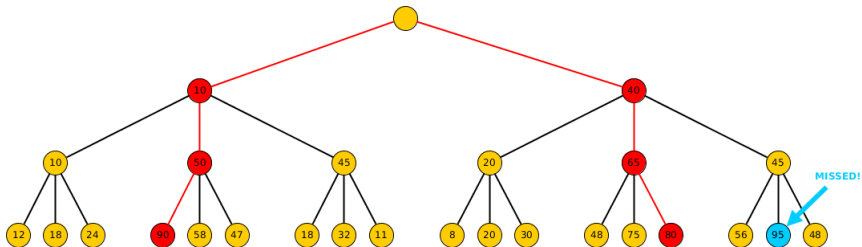


Лучевой поиск: $K = 2$



Лучевой поиск: $K = 2$



Лучевой поиск: $K = 2$ 

Комментарии

- Оптимизация: перебирать только признаки с максимальной информативностью.
- Для реализации нужна очередь с приоритетом (priority queue) с методами
 - `push(elements, scores)`: загрузить варианты с их оценками качества
 - `getKbest(K)`: выгрузить K лучших вариантов
- Сложность и полнота перебора:
 - Предположим, коэффициент ветвления B постоянный, а дерево поиска сбалансированное глубины D .
 - Тогда сложность поиска $O(KBD)$.
 - При достаточно большом K ($K \geq B^{D-1}$) превращается в полный перебор.

2 Методы поиска набора признаков

- Метод последовательной модификации набора признаков
- Лучевой поиск (beam search)
- Генетические алгоритмы

Генетические алгоритмы

- Каждый набор признаков $U = \{f_{i(1)}, f_{i(2)}, \dots, f_{i(K)}\}$ кодируется бинарным вектором $b = [b_1, b_2, \dots, b_D]$, где $b_i = \mathbb{I}[f_i \in U]$
- Жадное добавление/исключение работает быстро, но как аналог GD сходится к локальному оптимуму.
- Полный перебор - сложность $O(2^D)$.
 - Как увеличить широту перебора, не скатываясь к полному перебору?

Генетические алгоритмы

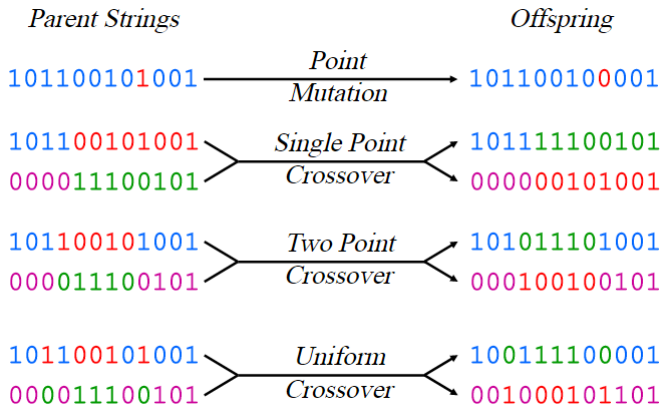
- Каждый набор признаков $U = \{f_{i(1)}, f_{i(2)}, \dots, f_{i(K)}\}$ кодируется бинарным вектором $b = [b_1, b_2, \dots, b_D]$, где $b_i = \mathbb{I}[f_i \in U]$
- Жадное добавление/исключение работает быстро, но как аналог GD сходится к локальному оптимуму.
- Полный перебор - сложность $O(2^D)$.
 - Как увеличить широту перебора, не скатываясь к полному перебору?

Гипотеза составного решения (building block hypothesis)

Хорошее решение состоит из комбинации других хороших решений.

- Генетические алгоритмы осуществляют поиск, комбинируя хорошие решения.

Операции скрещивания и мутации



Операции скрещивания и мутации⁵

- $mutation(b^1) = b$, where $b_i = \begin{cases} b_i^1 & \text{с вероятностью } 1 - \alpha \\ \neg b_i^1 & \text{с вероятностью } \alpha \end{cases}$
для некоторого $\alpha > 0$.
- $crossover(b^1, b^2) = b$, где $b_i = \begin{cases} b_i^1 & \text{с вероятностью } \frac{1}{2} \\ b_i^2 & \text{иначе} \end{cases}$
 - другие виды скрещивания: композиция с одним (single point) и двумя (2-point) разрывами.
- Биологическая аналогия: модификации генетических цепочек.

⁵Какая модификация этих операций приведет к аналогу градиентного подъема?

Генетический алгоритм

ВХОД:

размер популяции B и расширенной популяции B'
 параметры мутации и скрещивания
 макс. число итераций T , мин. изменение качества ΔS

АЛГОРИТМ:

сгенерировать B наборов признаков U_1, U_2, \dots, U_B случайно.
 инициализировать $t = 0$, $P^0 = \{S_1, S_2, \dots, S_B\}$, $S^0 = \max_{U \in P^0} S(U)$

пока $t \leq T$ и $S^t - S^{t-1} > \Delta S$:

$t = t + 1$

мутировать и скрещивать наборы из P^{t-1} :

$$U'_1, U'_2, \dots, U'_{B'} = \text{modify}(P^{t-1} | \theta)$$

упорядочить наборы по убыванию качества:

$$S(U'^t_{i(1)}) \geq S(U'^t_{i(2)}) \geq \dots S(U'^t_{i(B')})$$

загрузить в следующую популяцию B лучших наборов:

$$P^t = \{U'_{i(1)}, U'_{i(2)}, \dots, U'_{i(B)}\}$$

оценить качество по лучшему набору $S^t = \max_{U \in P^t} S(U)$

ВЫХОД: лучший набор признаков $\hat{U} = \arg \max_{U \in P^t} S(U)$

Улучшения генетического алгоритма

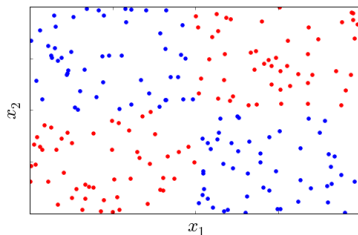
- **Ускорение:** генерировать вначале и мутацией f с $p \propto I(f)$.
- **Удлинить процесс оптимизации:**
 - прерывать процесс только если нет улучшения несколько итераций подряд.
 - при стагнации увеличить вероятность мутации
- **Бережнее модифицировать хорошие наборы и признаки:**
 - дополнять P^t лучшими наборами из P^{t-1} .
 - \downarrow вероятность мутации для хороших признаков (часто встречающиеся в наборах P^{t-1}).
 - \uparrow вероятность мутации для плохих признаков (редко встречающиеся в наборах P^{t-1}).
- **Увеличить широту поиска:**
 - скрещивание между > 2 наборами
 - вести несколько популяций из разных начальных условий, скрещивание лучших представителей между популяциями.

Важность признаков в контексте

Признаки могут влиять на y не по отдельности, а совместно:

$$p(y|x^1) = p(y), \quad p(y|x^2) = p(y)$$

$$p(y|x^1, x^2) \neq p(y)$$



Определение признаков, влияющих в контексте

Какие из методов могут определять признаки, влияющие в контексте?

- ❶ $corr(x^1, y), corr(x^1, y)$
- ❷ $MI(x^1, y), MI(x^1, y)$
- ❸ $MI([x^1, x^2], y)$
- ❹ критерий relief
- ❺ последовательное включение одного признака
- ❻ последовательное исключение одного признака
- ❼ важности признаков по дереву (дерево с ранней остановкой)
- ❽ важности признаков по дереву (дерево с обрезкой [prunning])

Заключение

- Отбор признаков позволяет быстрее настраивать модели.
 - модели точнее, если много шумовых признаков
- Предпочтение методам со встроенным отбором признаков.
- Методы отбора признаков, упорядоченные по сложности:
 - отбирать признаки по значимости
 - последовательное включение/исключение 1 признака
 - последовательное включение/исключение группы признаков
 - лучевой поиск с поддержкой K лучших групп признаков
 - генетический алгоритм генерации наборов
 - полный перебор
- Последовательное включение/исключение, лучевой поиск, генетический алгоритм применимы и для др. задач дискретной оптимизации (например подбор архитектуры нейросети).