Распознавание на основе скрытых марковских моделей (часть 2)

ОСТРОВСКИЙ Алексей Викторович

Институт кибернетики имени В. М. Глушкова НАН Украины

13 января 2015 г.

План доклада

1. Основные положения первой части доклада.
2. Общие сведения об алгебраическом подходе к распознаванию и алгоритмических композициях.
3. Линейные смеси марковских моделей.
4. Композиции с областями компетентности составляющих моделей.

Общая задача распознавания

Ген:

Нуклеотиды	ATGG…GCAA	GTAA…TCAG	ATTT…AAAG	$\cdots$	GTAA…CCAG	GGTG…ATAA
Структура	Экзон	Интрон	Экзон	$\cdots$	Интрон	Экзон

Белок:

Аминокислоты	NL	KLGFLV	KQPEE	PWFQTEWKFADKAGKDL	GF	EVIKIA	VPD	$\cdots$
Структура	$-$	β	$-$	α	$-$	β	$-$	$\cdots$

Состояния	ДНК	Белки
наблюдаемые (S)	нуклеотиды	аминокислоты
скрытые (H)	экзоны, интроны	α-спирали, β-листы, нерегулярные структуры

Вероятностная постановка задачи

Задача

Найти алгоритм $\mathcal{A}$, который для произвольной строки наблюдаемых состояний $\bar{s} \in S^*$ определяет строку скрытых состояний $\bar{h} \in H^*$, максимизирующую критерий качества $\mathcal{L}(\bar{s}, \bar{h})$.

Задача 1: поиск наиболее вероятной строки скрытых состояний.

$$\mathcal{A}(\bar{s}) = \arg\max_{\bar{h}} P(\bar{h} \mid \bar{s}, \Theta) = \arg\max_{\bar{h}} P(\bar{h}, \bar{s} \mid \Theta);$$

Θ — параметры модели.

Задача 2: поиск последовательности наиболее вероятных состояний.

$$h_i^* = \arg\max_{h} P(h_i = h, \bar{s} \mid \Theta) = \sum_{h_t, t \neq i} P(\bar{h}, \bar{s} \mid \Theta).$$

Зависимости между состояниями

		влияющие состояния					
Наблюдаемое состояние	D	L	G	F	E	V	I
Скрытое состояние	α	α	–	–	β	β	?
		влияющие состояния					

$$P(h_i, s_i \mid h_1 \dots h_{i-1}, s_1 \dots s_{i-1}) = P(h_i, s_i \mid h_{i-l} \dots h_{i-1}, s_{i-l} \dots s_{i-1}).$$

Предложение

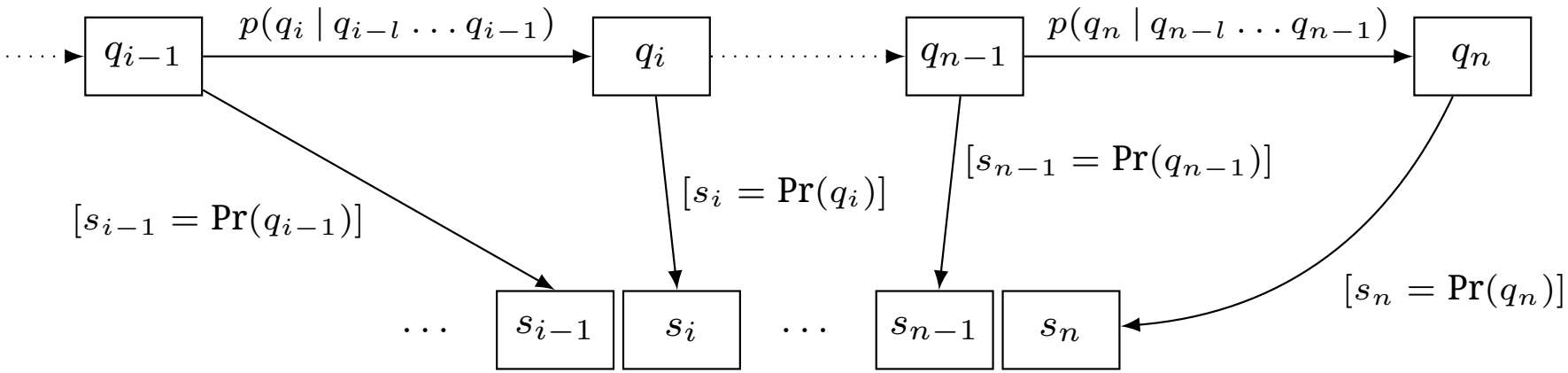
Для устранения зависимости между наблюдаемыми состояниями следует выполнить переход к скрытым состояниям

$$q_i = (s_i, h_i); \quad P(q_i \mid q_1 \dots q_{i-1}, \bar{s}) = P(q_i, s_i \mid q_{i-l} \dots q_{i-1}).$$

Множество скрытых состояний: $Q = S \times H$.

Зависимости между состояниями — модель $\mathcal{M}(l)$

Функция проекции: $\Pr : Q^* \to S^*, \quad \Pr((s_1, h_1)(s_2, h_2) \dots (s_n, h_n)) = s_1 s_2 \dots s_n.$

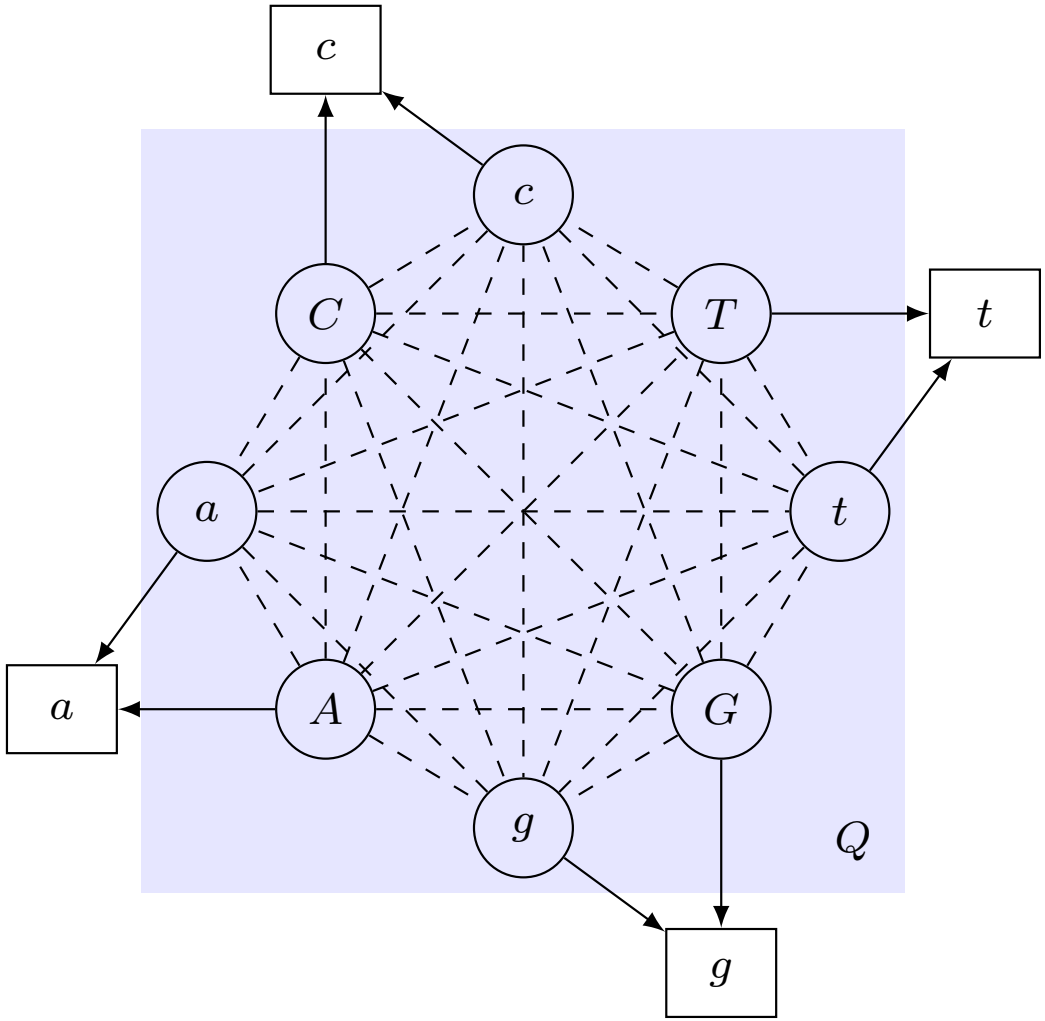


Совместная вероятность:

$$P(\bar{s}, \bar{q}) = P(\bar{q})[\Pr(\bar{q}) = \bar{s}];$$

$$P(\bar{q}) = \varphi(n)\pi(q_1 \dots q_l) \prod_{i=l+1}^{n} p(q_i \mid q_{i-l} \dots q_{i-1}), \quad n = |\bar{q}|.$$

Модель $\mathcal{M}(l)$ для генов



Скрытые состояния: A, C, G, T — нуклеотиды в экзонах, a, c, g, t — в интронах.

Решение задач распознавания

Задача 1: $P(\bar{q}, \bar{s}) \longrightarrow \max_{\bar{q}}$.

Метод решения: динамическое программирование (модификация алгоритма Витерби).

Вычислительная сложность: $\mathcal{O}(|\bar{s}| \cdot |H|^{l+1})$.

Задача 2: $P(q_i, \bar{s}) \longrightarrow \max_{q_i}, \quad i = 1, \ldots, |\bar{s}|$.

Метод решения: использование вспомогательных сумм для ускорения вычисления

$$P(q_i, \bar{s}) = \sum_{q_1} \cdots \sum_{q_{i-1}} \sum_{q_{i+1}} \cdots \sum_{q_n} P(\bar{q}, \bar{s}), \quad n = |\bar{s}|.$$

Вычислительная сложность: $\mathcal{O}(|\bar{s}| \cdot |H|^{2l})$.

Алгебраический подход к распознаванию [Журавлев]

Алгоритм распознавания: $\mathcal{A} : X \to Y$.

X — пространство объектов; Y — множество ответов.

$$\mathcal{A}(x) = C(b(x)), \quad x \in X;$$

$$X \xrightarrow{b} \mathcal{R} \xrightarrow{C} Y;$$

b — алгоритмический оператор; C — решающее правило; $\mathcal{R}$ — пространство оценок.

Пример: логистическая регрессия.

$$X = \mathbb{R}^n, Y = \{0, 1\}, \mathcal{R} = [0, 1];$$

$$b(x) = \sigma(w^T x), C(b) = [b > 0{,}5].$$

Задачи распознавания последовательностей

Объекты и ответы: $X = S^*, Y = Q^*$.

Задача 1: $P(\bar{q} \mid \bar{s}) \longrightarrow \max_{\bar{q}}$.

$$b(\bar{s}) = \{P(\bar{q} \mid \bar{s}) \mid \bar{q} \in Q^n\} \in \mathbb{R}^{|Q|^n};$$

$$C(b) = \arg\max_{\bar{q}} b(\bar{q}).$$

Задача 2: $P(q_i \mid \bar{s}) \longrightarrow \max_{q_i}, \quad i = 1, \ldots, |\bar{s}|$.

$$b(\bar{s}) = \{P(q_i \mid \bar{s}) \mid q_i \in Q,\ i = 1, \ldots, n\} \in \mathbb{R}^{|Q| \times n};$$

$$C(b) = \bigodot_{i=1}^{n} \arg\max_{q} b(q, i);$$

($\odot$ — операция конкатенации).

Алгоритмические композиции

Определение

Алгоритмической композицией, составленной из операторов $b_j(x) : X \to \mathcal{R}$, $j = 1, \ldots, k$, и корректирующей операции $F : \mathcal{R}^k \to \mathcal{R}$, называется алгоритм

$$\mathcal{A}(x) = C(F(b_1(x), b_2(x), \ldots, b_k(x))).$$

Смеси алгоритмов:

$$b(x) \equiv F(b_1(x), b_2(x), \ldots, b_k(x)) = \sum_{j=1}^{k} g_j(x) b_j(x);$$

g_j — функции компетентности (англ. *gate functions*).

$$\forall x \in X \quad g_j(x) \geqslant 0, \quad \sum_{j=1}^{k} g_j(x) = 1.$$

Функции компетентности

Гипотеза

Строка $\bar{s}$ порождается k вероятностными распределениями, которые характеризуются параметрами Θ_1, Θ_2,..., Θ_k:

$$\underbrace{P(\bar{q} \mid \bar{s})}_{b(\bar{s})} = \sum_{j=1}^{k} \underbrace{P(\bar{q} \mid \bar{s}, \Theta_j)}_{b_j(\bar{s})} \underbrace{P(\Theta_j \mid \bar{s})}_{g_j(\bar{s})};$$

$g_j(\bar{s}) = P(\Theta_j \mid \bar{s})$ — вероятность реализации j-го распределения на строке $\bar{s}$.

Строка наиболее вероятных скрытых состояний:

$$\underbrace{P(q_i \mid \bar{s})}_{b(\bar{s})} = \sum_{j=1}^{k} \underbrace{P(q_i \mid \bar{s}, \Theta_j)}_{b_j(\bar{s})} \underbrace{P(\Theta_j \mid \bar{s})}_{g_j(\bar{s})}.$$

Выбор функций компетентности

Варианты алгоритмических композиций:

- Линейная смесь алгоритмов: $g_j(\bar{s}) = w_j = \text{const}(\bar{s})$.
- Области компетентности: $g_j(\bar{s}) = [\bar{s} \in G_j]$.

  $\{G_j\}_{j=1}^k$ — покрытие множества S^*:

$$\forall j, t \in 1, \ldots, k \;\; G_i \cap G_j = \varnothing; \quad \bigcup_{j=1}^{k} G_j = S^*.$$

Линейная смесь алгоритмов

Параметры композиции: $\Theta = (w_1, \ldots, w_k, \Theta_1, \ldots, \Theta_k)$.

Обучение параметров: EM-алгоритм.

Дано: обучающая выборка $T = \{\bar{q}^i\} \subset Q^*$; начальное приближение параметров Θ^{init}.

1. **Ожидание:** определить апостериорные оценки вероятностей

$$\varphi_{ij} := P(\Theta_j \mid \bar{q}^i) = \frac{w_j P(\bar{q}^i \mid \Theta_j)}{\sum_t w_t P(\bar{q}^i \mid \Theta_t)}.$$

2. **Максимизация:** решить задачи максимизации взвешенного правдоподобия:

$$w_j := \frac{1}{|T|} \sum_i \varphi_{ij};$$

$$\Theta_j := \arg\max_{\Theta} \sum_i \varphi_{ij} \log P(\bar{q}^i \mid \Theta).$$

3. Повторять шаги 1 и 2 до сходимости.

Поиск решения

Проблема: для задачи поиска наиболее вероятной последовательности состояний вероятности $P(\bar{q} \mid \bar{s})$ не вычисляются явно.

Дано: строка $\bar{s}$, веса w_j, параметры моделей Θ_j.

$$\log P(\bar{q} \mid \bar{s}) = \log \left(\sum_{j=1}^{k} w_j P(\bar{q} \mid \bar{s}, \Theta_j) \right) \longrightarrow \max_{\bar{q}}.$$

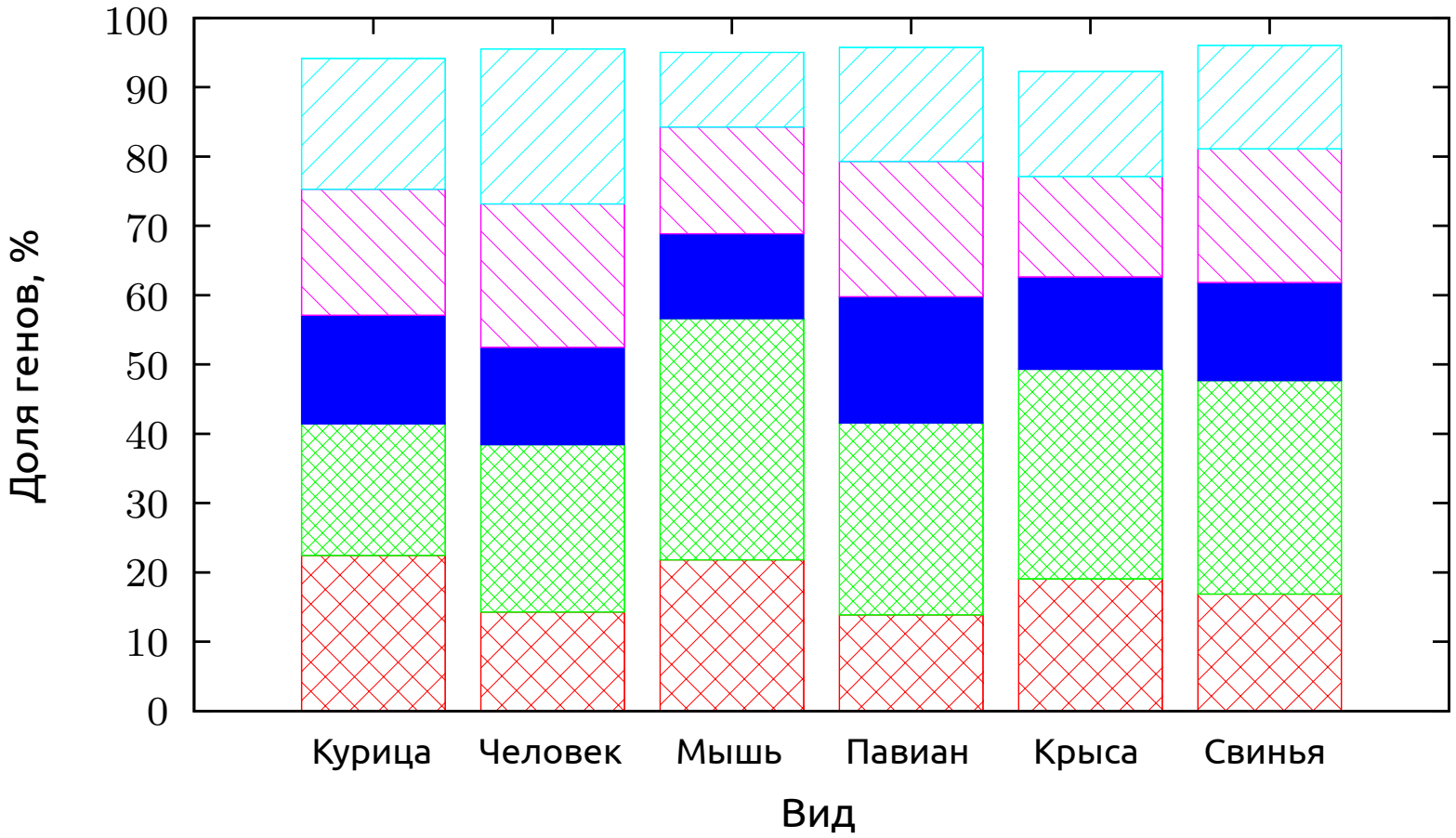
Неравенство Йенсена:

$$\log P(\bar{q} \mid \bar{s}) \geqslant \sum_{j=1}^{k} \varphi_j \log(w_j P(\bar{q} \mid \bar{s}, \Theta_j)) - \sum_{j=1}^{k} \varphi_j \log \varphi_j; \quad \varphi \in \Delta^k.$$

Оптимизация по $\bar{q}$: модель $\mathcal{M}(l)$ с параметрами, выражающимися через Θ_j.

Оптимизация по φ: $\varphi_j^* = P(\Theta_j \mid \bar{q})$.

Разбиение на зоны ответственности



Зона ответственности j-го алгоритма: $P(\Theta_j \mid \bar{q}) \geqslant 0{,}99$.

Качество композиций

Вид	k	Мера качества, %					
		отдельные состояния				границы	
		SSp	SSn	CC	ACP	ESp	ESn
Homo sapiens	1	35,58	89,56	49,48	76,45	27,64	31,11
(человек)	5	57,48	86,15	66,42	83,71	45,19	49,70
Sus scrofa	1	33,24	85,66	47,60	75,75	24,87	26,64
(свинья)	5	55,41	78,20	62,31	81,53	40,77	41,86
Mus musculus	1	59,97	85,32	67,20	83,97	42,22	40,39
(мышь)	5	76,28	81,58	75,96	87,99	53,47	47,87
Rattus norvegicus	1	61,73	83,59	67,47	84,01	40,75	36,49
(крыса)	4	76,30	78,34	74,00	87,00	48,85	40,63
Papio anubis	1	39,91	86,65	52,28	77,45	30,41	31,42
(павиан)	5	68,23	81,13	71,09	85,65	50,78	49,74

Итоги вычислительного эксперимента

- Качество распознавания при использовании смесей распределений повышается на 10–15 %.
- Смеси распределений устойчивы относительно способа выбора начального приближения.
- Оптимальное количество распределений в смеси — $k \in \{3, 4, 5\}$.
- Близким видам соответствуют схожие смеси (близость измеряется по логарифму правдоподобия для соответствующей выборки).

Композиции с областями компетентности

Алгоритмическая композиция:

$$\mathcal{A}(\bar{s}) = \begin{cases} \mathcal{A}_1(\bar{s}), & \text{если } \bar{s} \in G_1, \\ \vdots & \\ \mathcal{A}_k(\bar{s}), & \text{если } \bar{s} \in G_k; \end{cases}$$

где $\mathcal{A}_j$, $j = 1, \ldots, k$ — составляющие алгоритмы распознавания.

Параметры композиции: $\Theta = \{G_1, \ldots, G_k, \Theta_1, \ldots, \Theta_k\}$.

Утверждение

Параметры составляющей модели Θ_j, $j = 1, \ldots, k$, находятся с помощью обучения на выборке

$$T_j = \{\bar{q} \in T \mid \Pr(\bar{q}) \in G_j\}.$$

Определение оптимального разбиения

Предложение

Разбиение на области компетентности $\{G_j\}_{j=1}^k$ производится с помощью последовательности предикатов, зависящих от наблюдаемых состояний.

$$I : S^* \to \mathbb{R}^\alpha \to \{0,1\}, \quad \mathbb{R}^\alpha \text{ — признаки строки набл. состояний.}$$

Предикаты на основе концентраций состояний:

$$I(\bar{s}; U, \omega) = \Big[\sum_{u \in U} n(\bar{s}, u) \geqslant \omega\Big], \quad U \subset S, \omega \in (0,1).$$

1. вычислить суммарную концентрацию состояний из U в строке $\bar{s}$;
2. сравнить концентрацию с порогом ω.

Пример областей компетентности

Предикат: $I(\bar{s}) = [n(\bar{s}, \{C, G\}) \geqslant 0{,}492]$.

Области компетентности:

$$G_1 = [n(C) + n(G) < 0{,}492], \quad G_2 = S^* \setminus G_1 = [n(C) + n(G) \geqslant 0{,}492].$$

Ген	Содержание нуклеотидов, %					Часть разбиения
	$n(A)$	$n(C)$	$n(G)$	$n(T)$	$n(\{C, G\})$	
s_1	26,18	17,52	21,01	35,30	38,53	G_1
s_2	21,24	27,77	25,37	25,63	53,14	G_2
s_3	23,09	23,75	19,06	34,10	42,81	G_1
s_4	19,59	29,24	30,69	20,48	59,93	G_2
s_5	20,98	25,77	20,23	33,01	46,00	G_1
s_6	32,16	16,34	17,68	33,81	34,02	G_1

Алгоритм построения разбиения

$$\log P(T) \longrightarrow \max_{G_1,\ldots,G_k} \sim \max_{T_1,\ldots,T_k}.$$

1. Начальное разбиение состоит из всей выборки: $T_1 := T$, $k := 1$.
2. Определить *оптимальный* предикат:

$$(T^*, I^*) = \arg\max_{T_j, I} \Delta(T_j, I), \quad j = 1, \ldots, k.$$

3. Разбить часть выборки T^* на две согласно предикату I^*:

$$k := k + 1;$$
$$T_k := \{\bar{q} \in T^* \mid I(\Pr(\bar{q}))\};$$
$$T^* := \{\bar{q} \in T^* \mid \neg I(\Pr(\bar{q}))\}.$$

4. Если не выполнен критерий останова, вернуться к шагу 2.

Функционал качества предиката

Вычисление качества предиката I на выборке T:

1. $T^+ = \{\bar{q} \in T \mid I(\Pr(\bar{q}))\}, T^- = T \setminus T^+$;

   (части, на которые предикат делит выборку).

2. $\Theta[T] = \arg\max_{\Theta} P(T \mid \Theta)$, $\Theta[T^+] = \ldots, \Theta[T^-] = \ldots$

   (модели, обученные на выборке и ее частях).

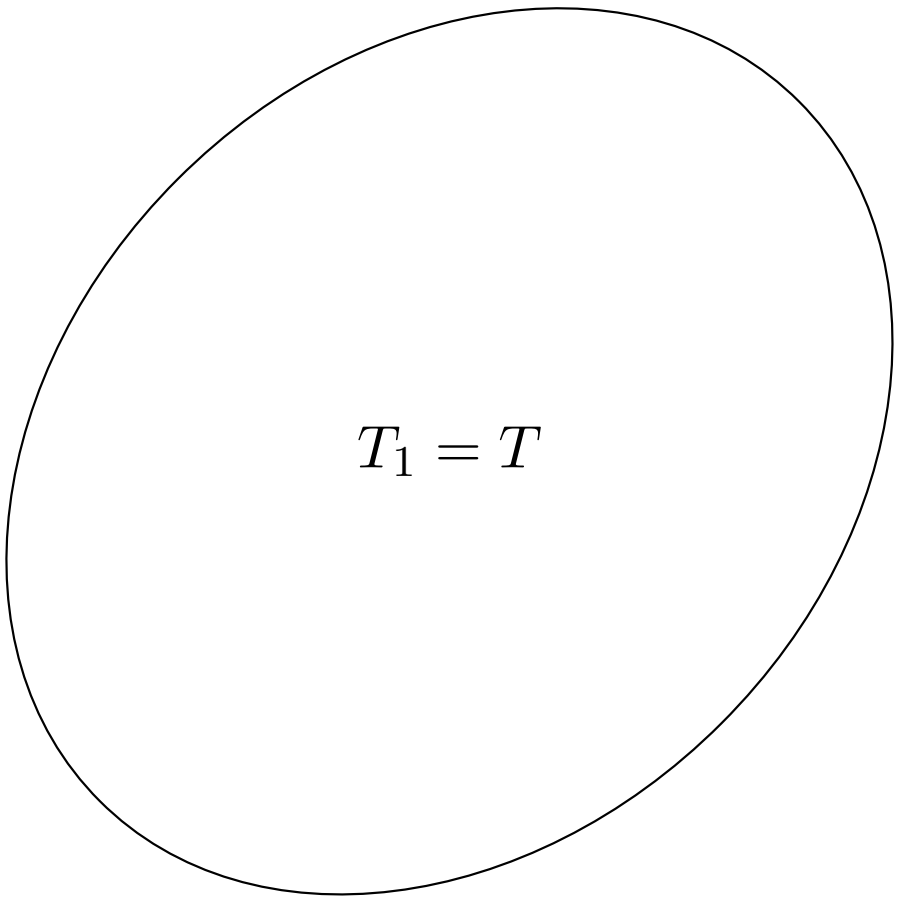
3. $\Delta(T, I) = \log P(T^+ \mid \Theta[T^+]) + \log P(T^- \mid \Theta[T^-]) - \log P(T \mid \Theta[T])$

   (функционал качества).

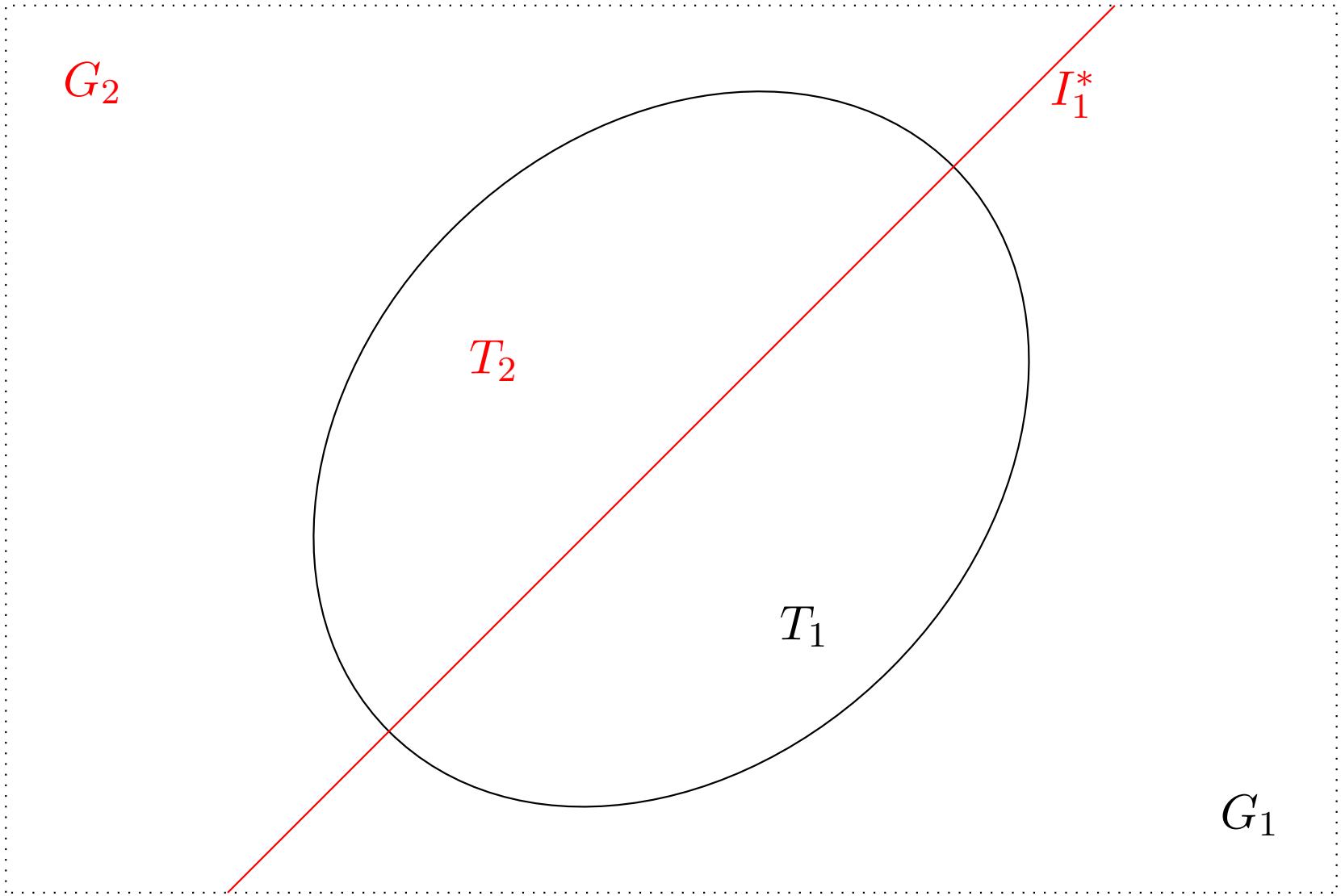
Наблюдение

Сумма $\Delta(T^*, I^*)$, полученных на каждой итерации алгоритма построения разбиения, равна $\log P(T \mid \Theta[T_1], \ldots, \Theta[T_k])$ с точностью до постоянной.

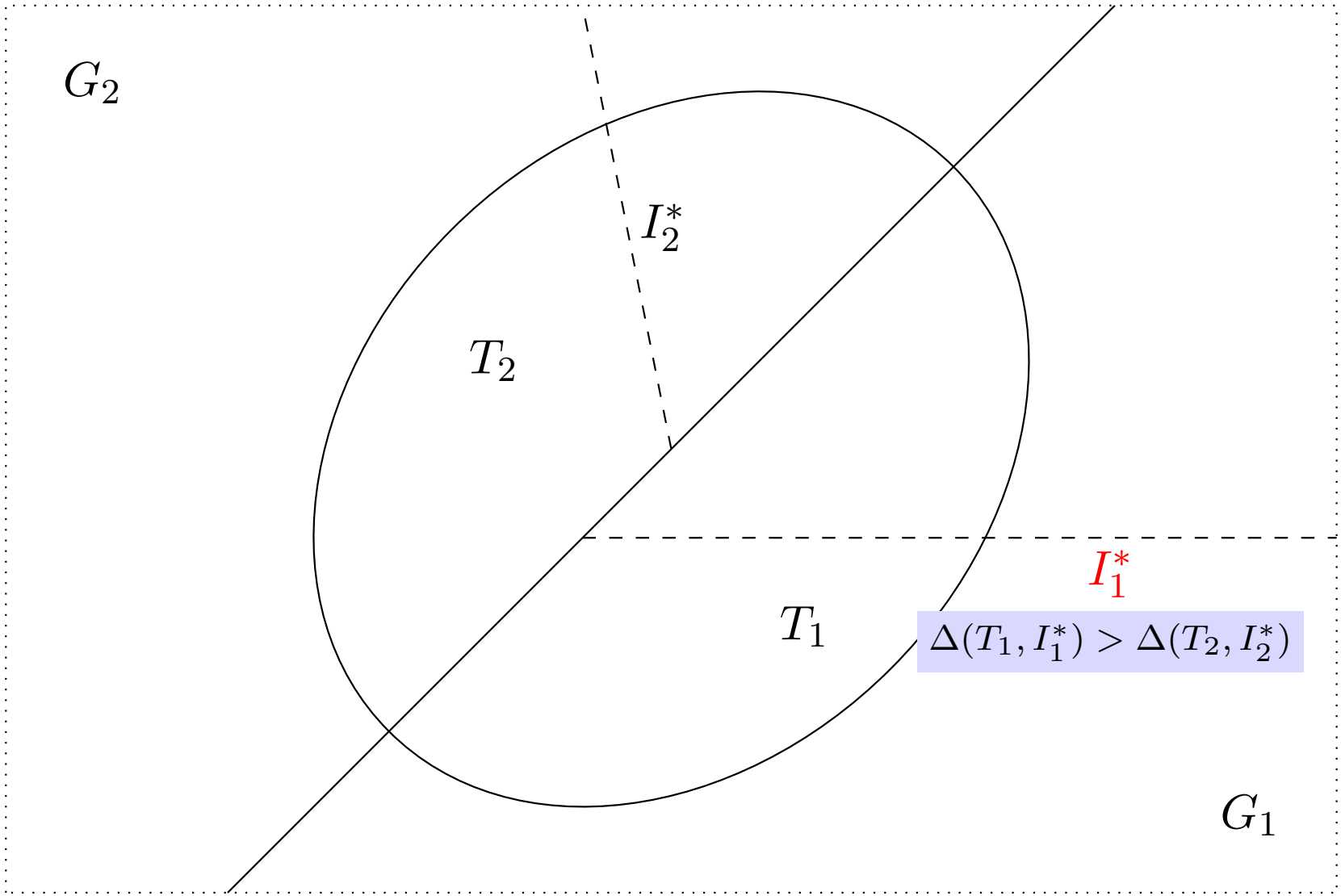
Пример построения разбиения



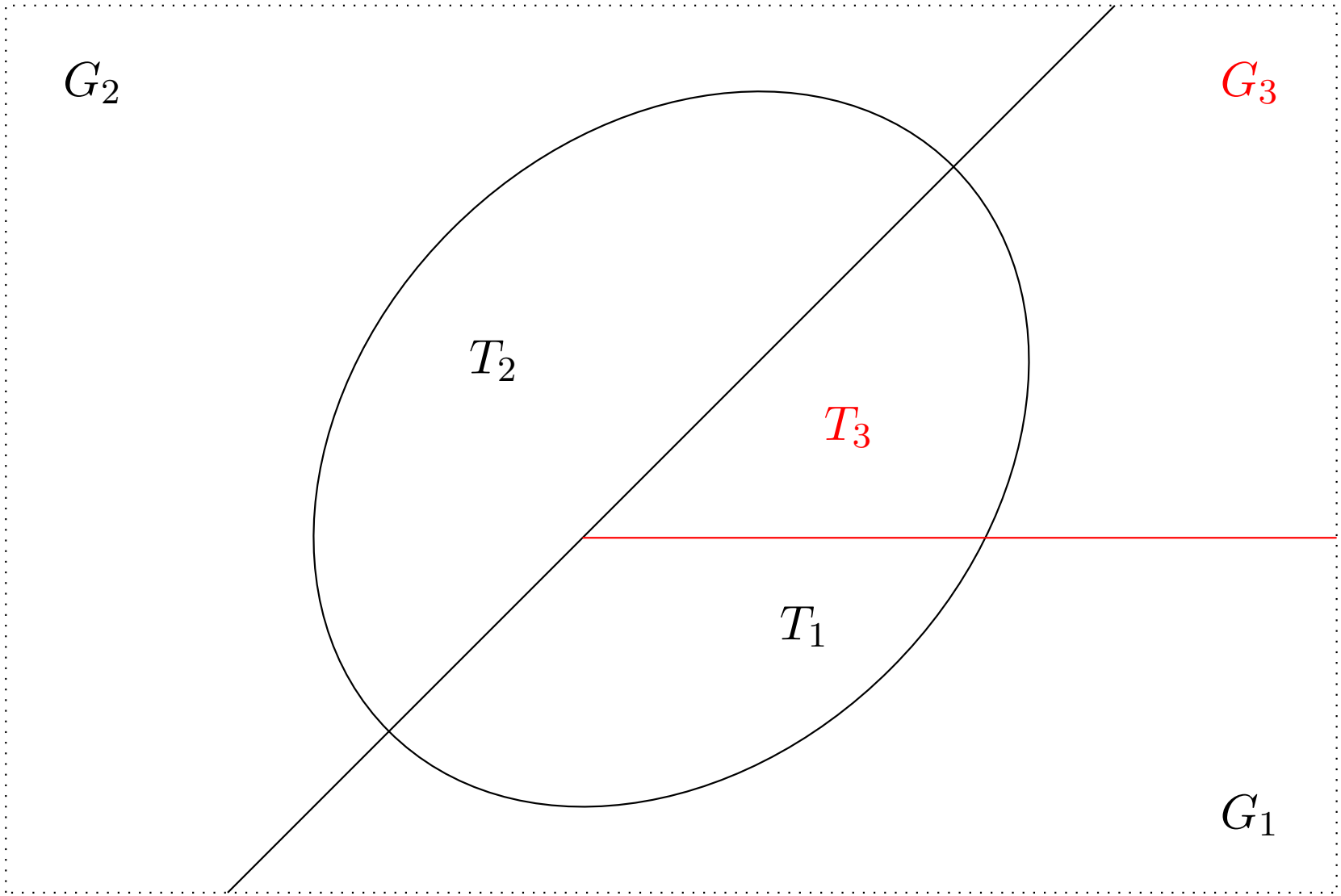
Пример построения разбиения



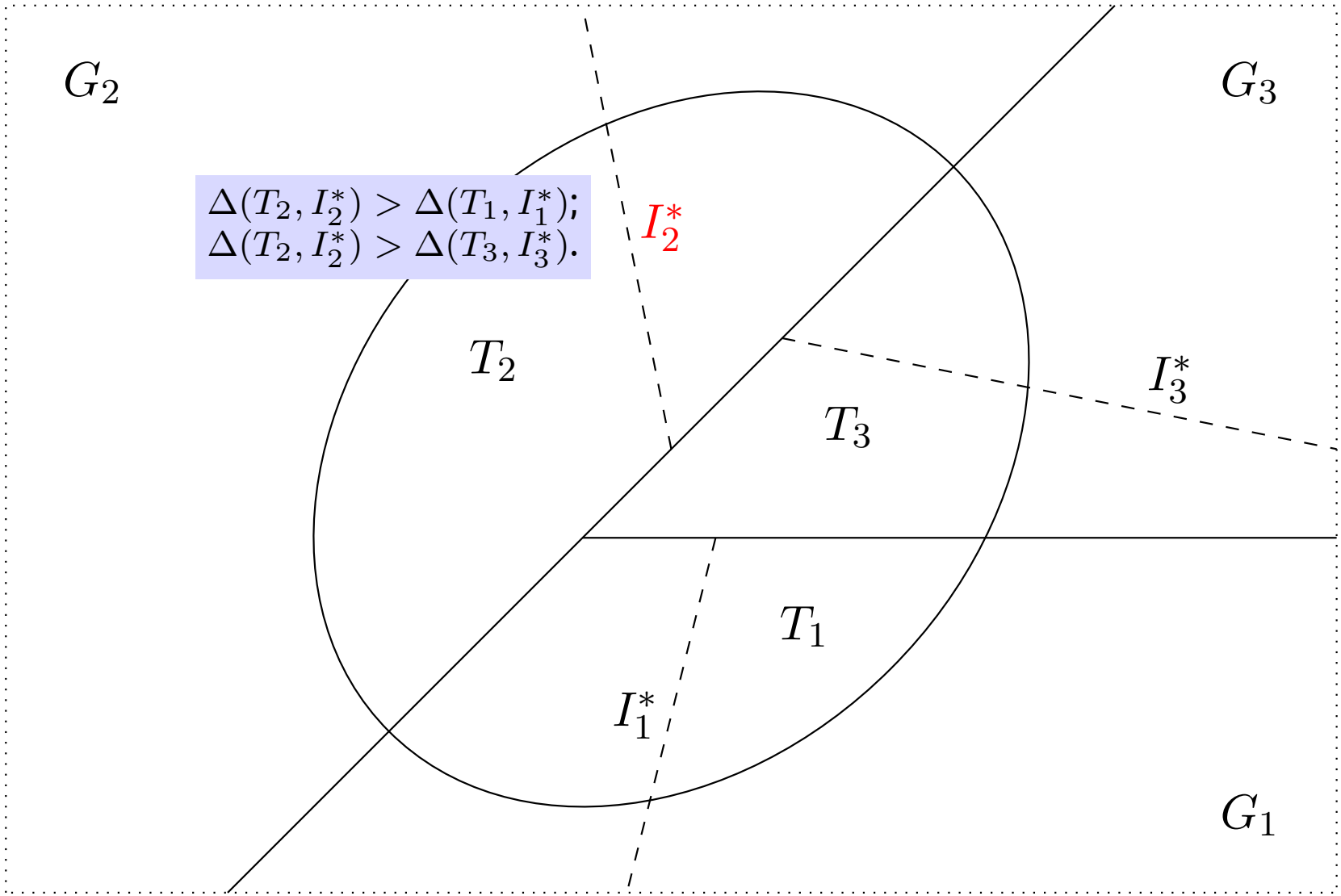
Пример построения разбиения



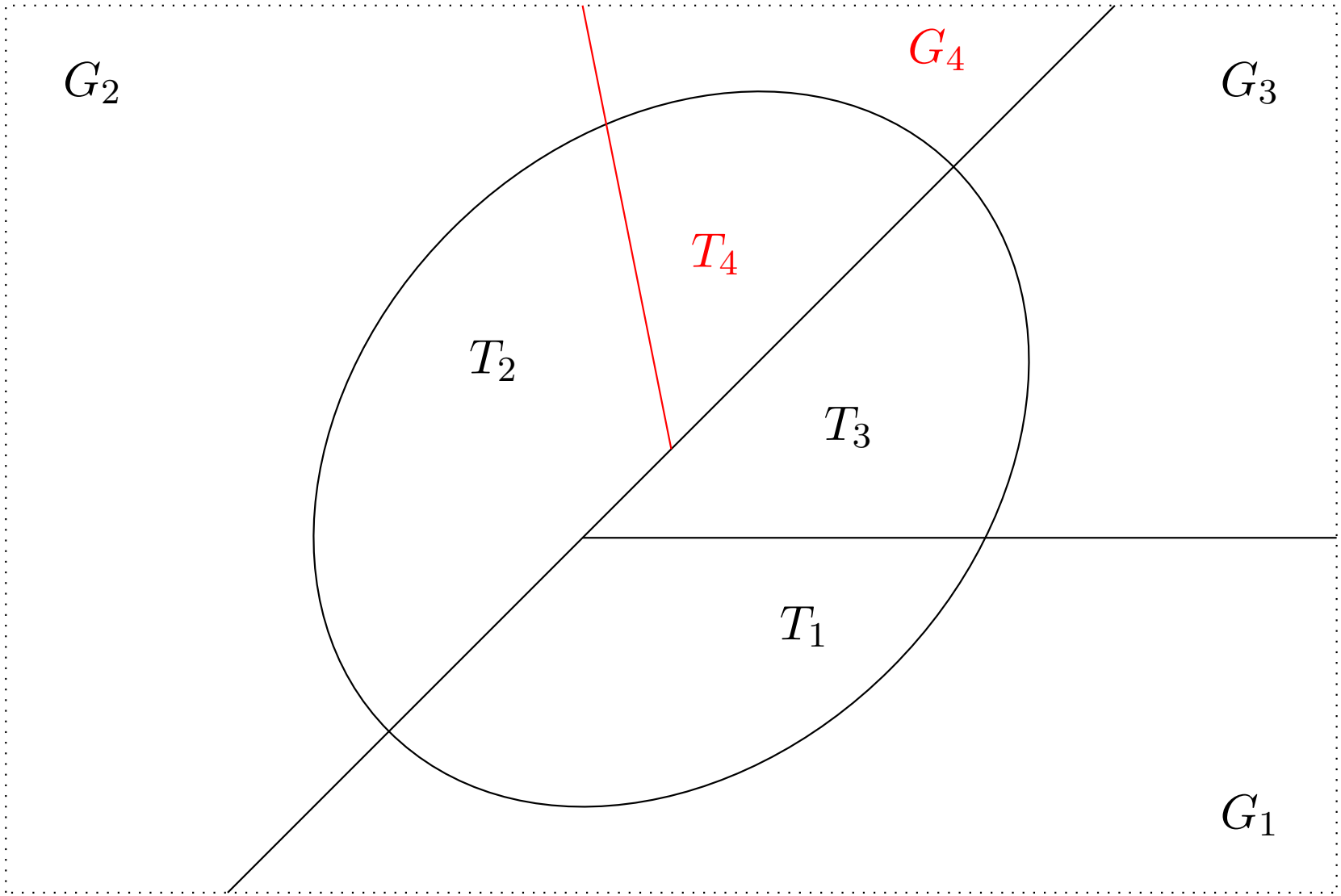
Пример построения разбиения



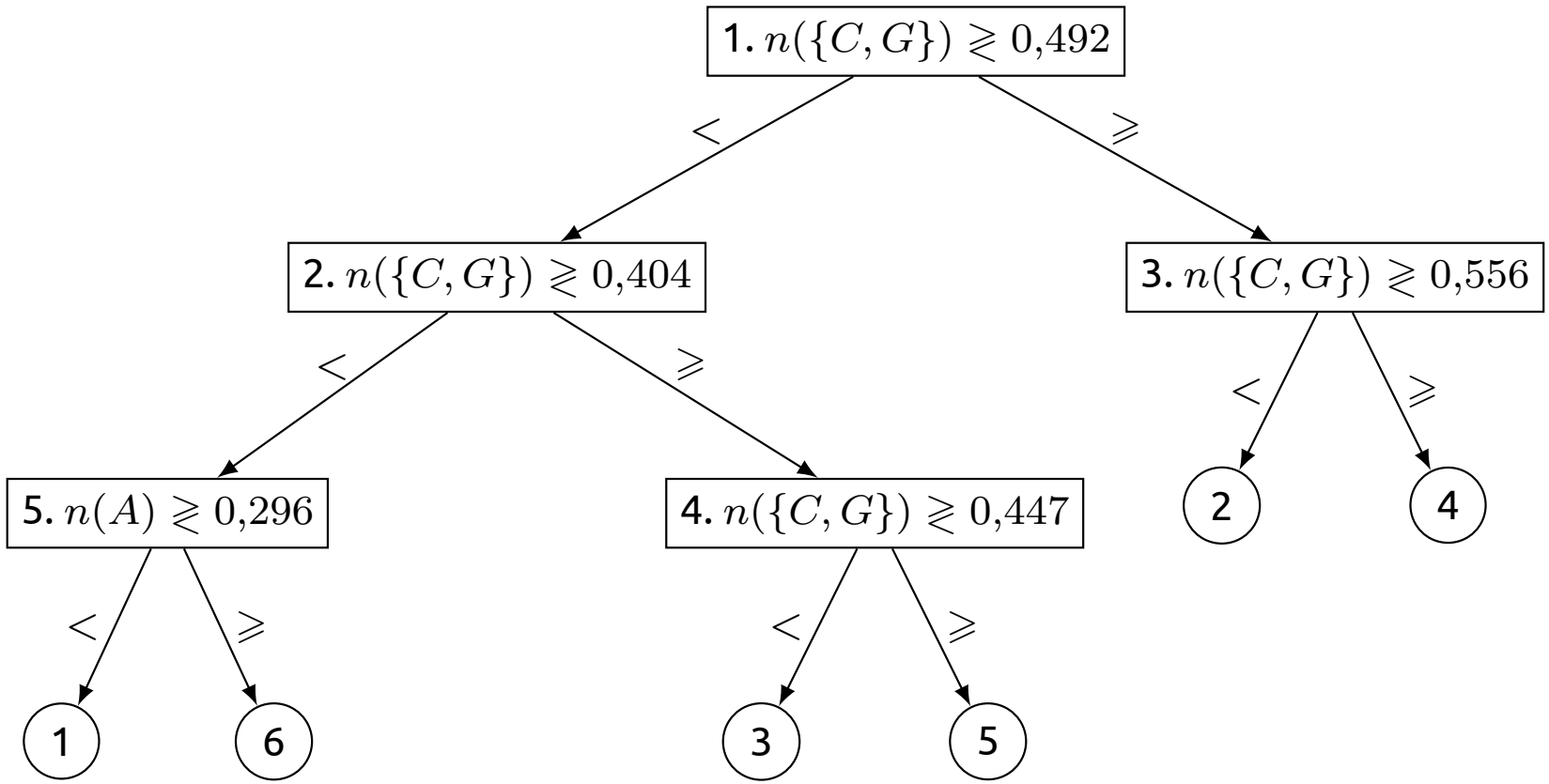
Пример построения разбиения



Пример построения разбиения



Пример разбиения



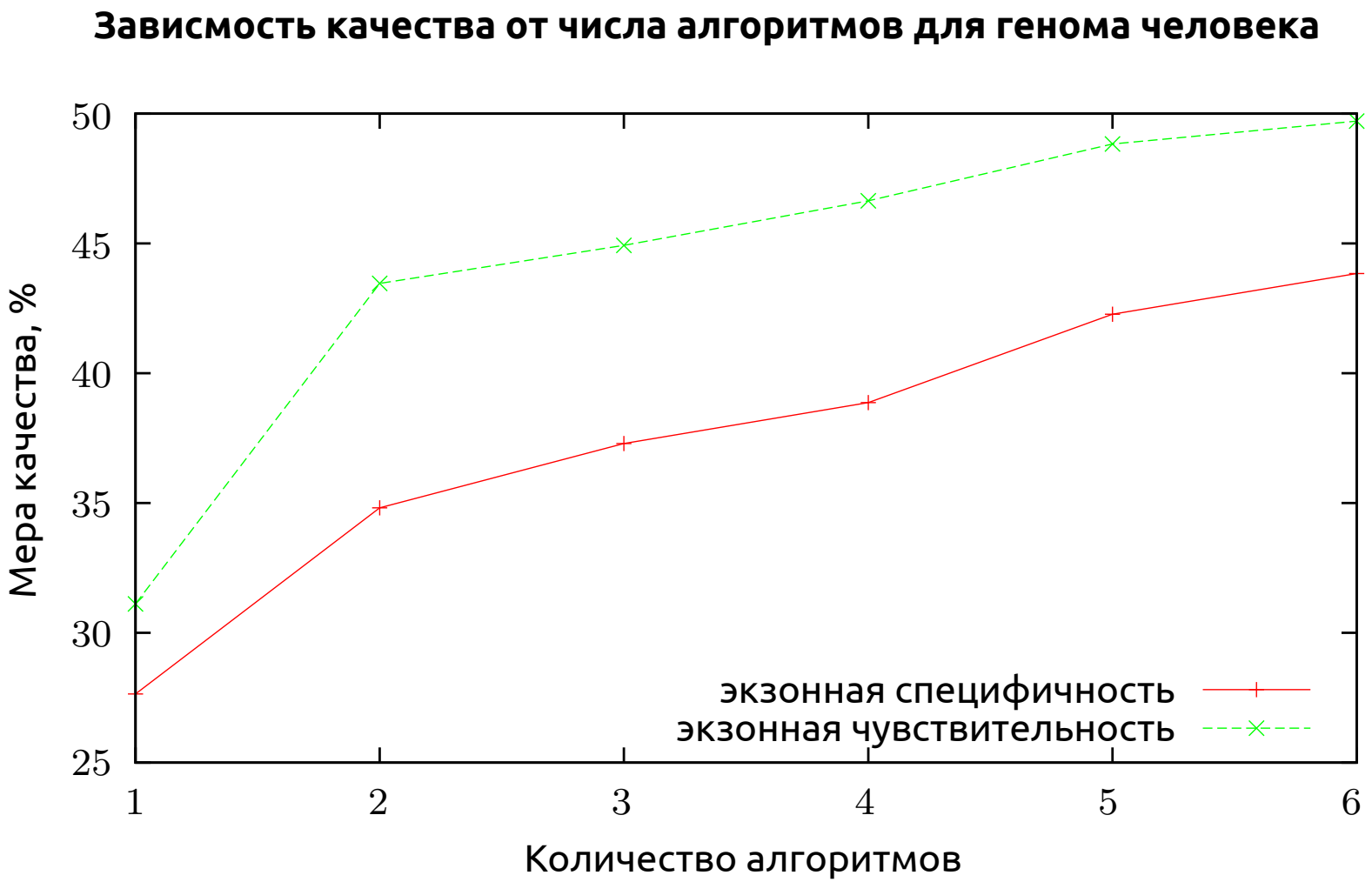
Дерево разбиения для генома человека

Пример разбиения

Разбиение для генома человека:

Часть разбиения	Описание	Кол-во генов
G_1	$n(\{C,G\}) < 0{,}404; n(A) < 0{,}296$	1727
G_2	$0{,}492 \leqslant n(\{C,G\}) < 0{,}556$	5549
G_3	$0{,}404 \leqslant n(\{C,G\}) < 0{,}447$	3731
G_4	$n(\{C,G\}) > 0{,}556$	5525
G_5	$0{,}447 \leqslant n(\{C,G\}) < 0{,}492$	3641
G_6	$n(\{C,G\}) < 0{,}404; n(A) \geqslant 0{,}296$	2077

Качество композиций



Качество композиций

Вид	k	Мера качества, %					
		отдельные состояния				границы	
		SSp	SSn	CC	ACP	ESp	ESn
Homo sapiens	1	35,58	89,56	49,48	76,45	27,64	31,11
(человек)	6	54,43	87,53	64,71	83,02	43,84	49,71
Sus scrofa	1	33,24	85,66	47,60	75,75	24,87	26,64
(свинья)	4	47,54	81,02	58,04	79,83	36,21	40,86
Mus musculus	1	59,97	85,32	67,20	83,97	42,22	40,39
(мышь)	4	71,96	83,03	74,08	87,11	50,90	47,73
Rattus norvegicus	1	61,73	83,59	67,47	84,01	40,75	36,49
(крыса)	6	76,23	75,32	72,22	86,11	47,34	38,78
Papio anubis	1	39,91	86,65	52,28	77,45	30,41	31,42
(павиан)	6	65,00	79,75	68,56	84,42	50,54	51,18

Итоги вычислительного эксперимента

- Качество при использовании композиций повышается на 10–15 %.
- Оптимальное количество алгоритмов в композиции — $k \in \{4, 5, 6\}$.
- Большое количество предикатов на основе концентрации $n(\{C, G\})$.
- Близким видам соответствуют похожие деревья предикатов.
- Для задачи распознавания вторичной структуры белков композиции неэффективны.

Спасибо за внимание!

Сайт: http://sestudy.edu-ua.net/ru/bio

E-mail: ostrovski.alex@gmail.com