Национальный исследовательский Томский государственный университет

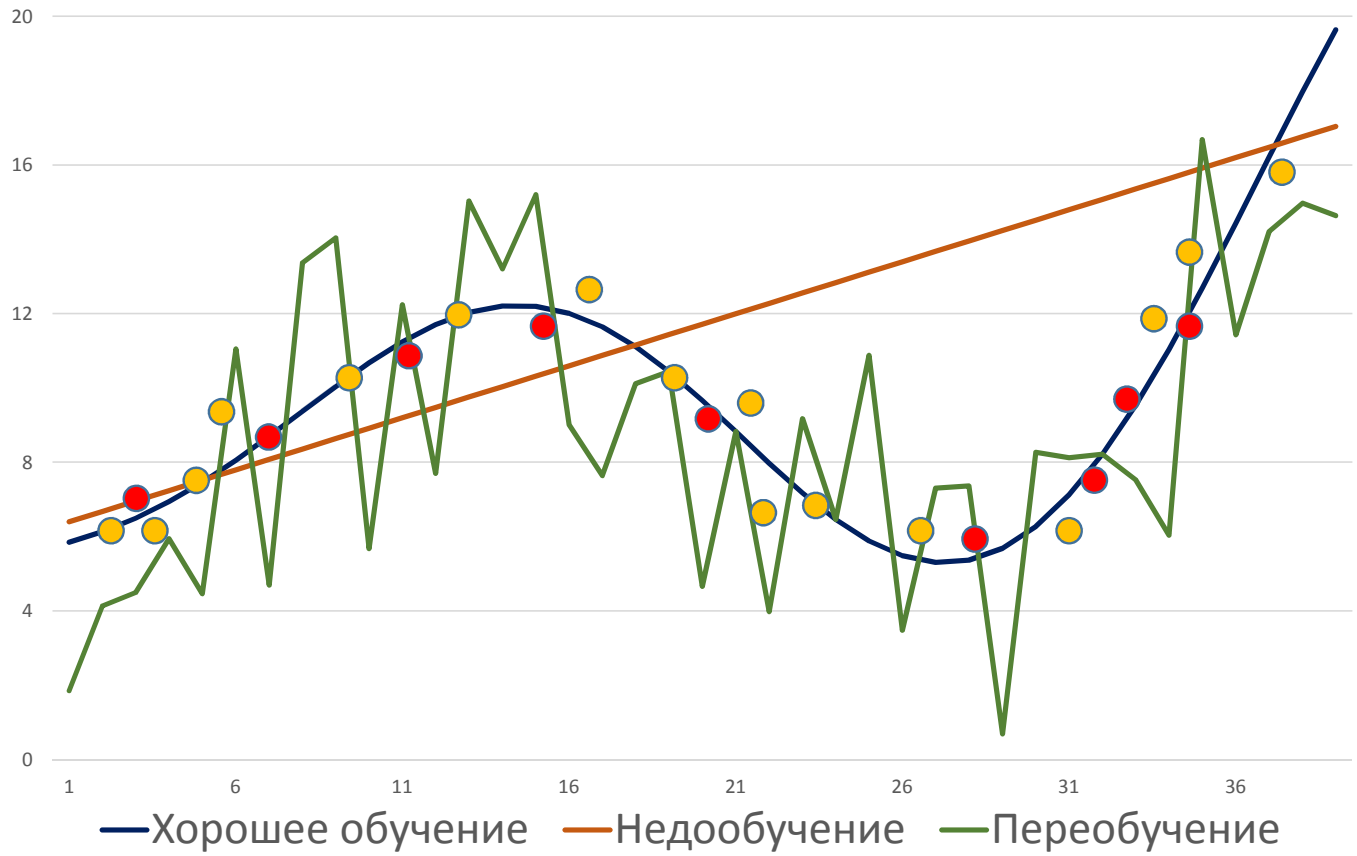
world skills Russia

Подбор переменных и алгоритмов для модели

Сергей Аксёнов, к.т.н., доцент кафедры

Теоретических основ информатики ТГУ

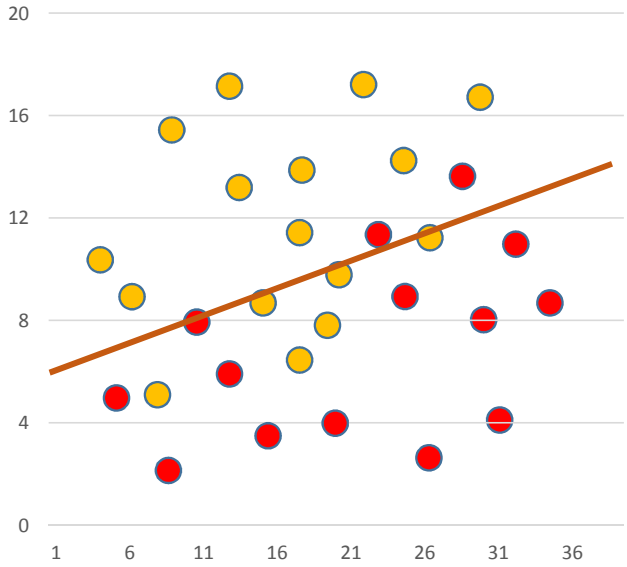
Плохое и хорошее обучение (регрессия)



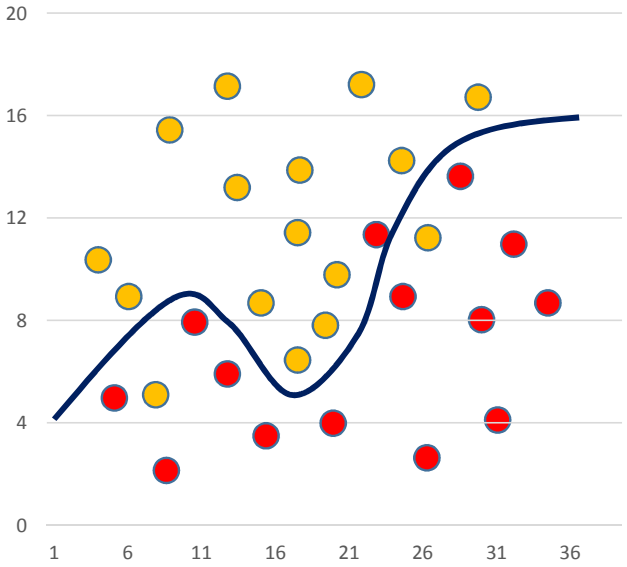
Обучающие примеры

Тестовые примеры

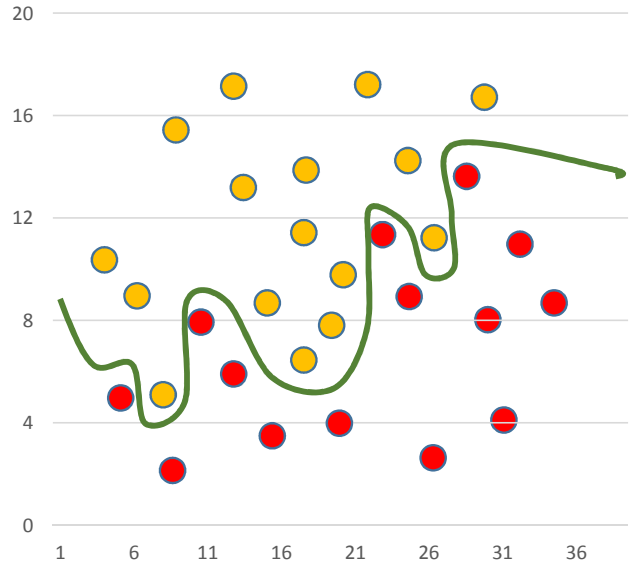
Плохое и хорошее обучение (Классификация)



Недообучение



Хорошее обучение



Переобучение

Класс A Класс B

Линейно разделимые и линейно неразделимые классы

Бинарный линейный классификатор

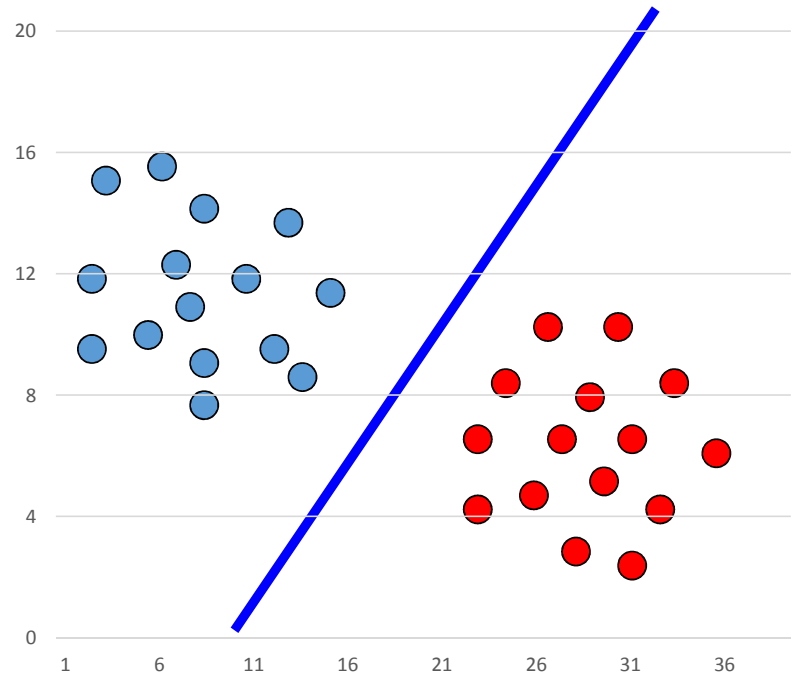
Результат обучения: входной вектор относится либо к положительному ($\hat{y}$=+1), либо отрицательному ($\hat{y}$=-1) классу

Вектор признаков:

$$x = (x_1, x_2, x_3, \ldots, x_N)$$

Выход модели:

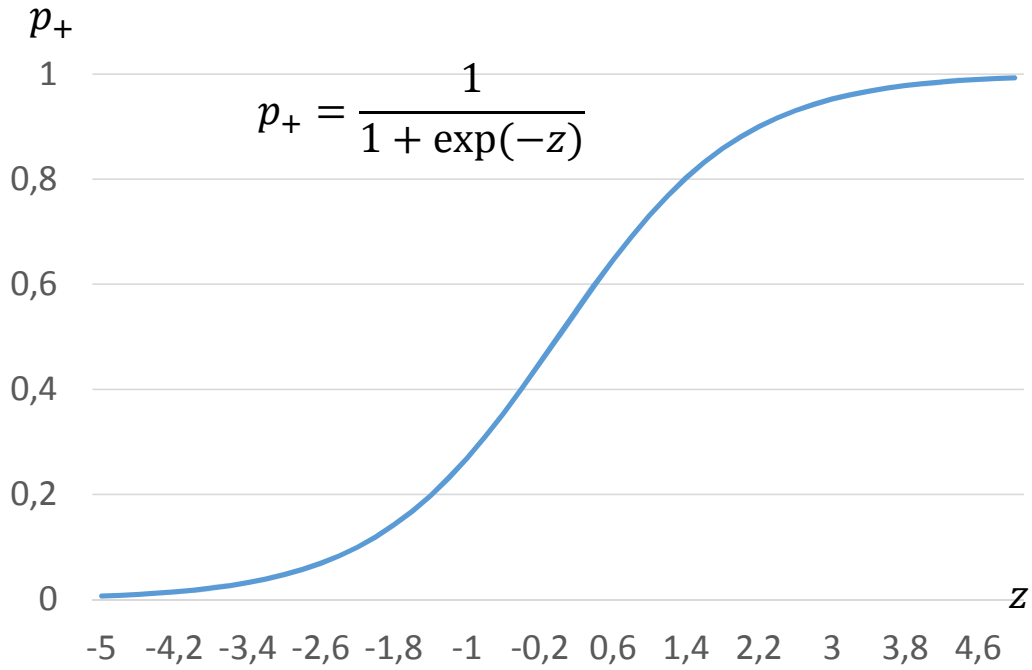
$$\hat{y} = \hat{y}(x, w) = sign\left(w_0 + \sum_i^N w_i x_i\right) = sign(w^T x)$$



Логистическая регрессия

Прогнозируют вероятность p_+ отнесения примера x r к классу +1

$z=\sum_i^N w_i x_i$



Функция стоимости

$$J(w)=\sum_i \frac{1}{2}\left(\phi(z^{(i)})-y^{(i)}\right)^2$$

Функция правдоподобия

$$L(w)=P(y\,|\,x;w)=\prod_{i=1}^{n} P\left(y^{(i)}\,|\,x^{(i)};w\right)=\prod_{i=1}^{n}\left(\phi(z^{(i)})\right)^{y^{(i)}}\left(1-\phi(z^{(i)})\right)^{1-y^{(i)}}$$

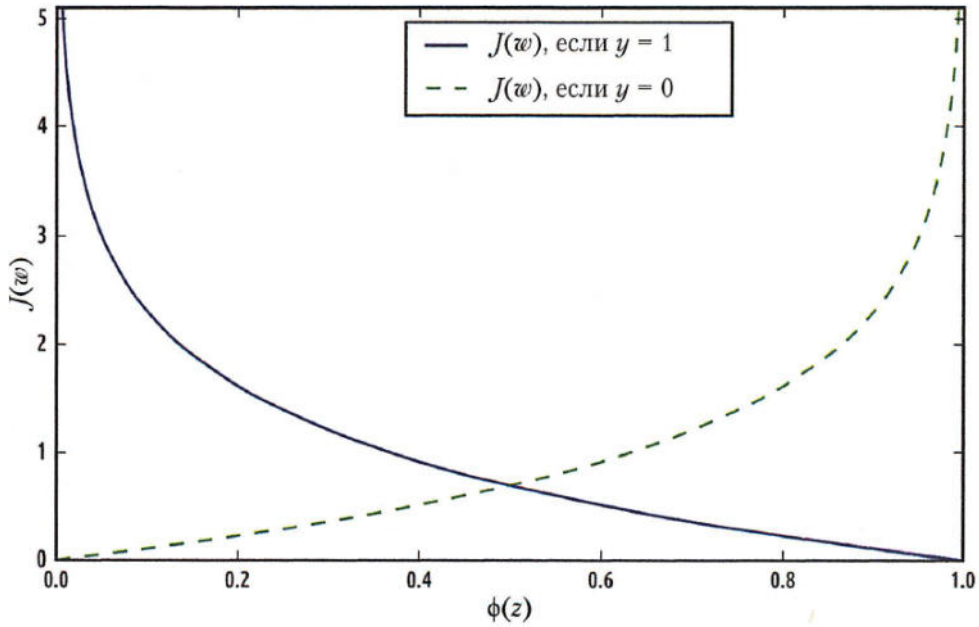
Логарифмическая функция правдоподобия

$$l(w)=\log L(w)=\sum_{i=1}^{n}\left[y^{(i)}\log\left(\phi(z^{(i)})\right)+\left(1-y^{(i)}\right)\log\left(1-\phi(z^{(i)})\right)\right]$$

Регуляризация в логистической регрессии

Функция, использующаяся для поиска параметров модели

$$J(w)=\sum_{i=1}^{n}\left[-y^{(i)}\log\left(\phi(z^{(i)})\right)-\left(1-y^{(i)}\right)\log\left(1-\phi(z^{(i)})\right)\right]$$



Регуляция переобучения выполняется при помощи регуляризации.
Наложение штрафов на экстремальные значения параметров.

$$\frac{\lambda}{2}\|w\|^2=\frac{\lambda}{2}\sum_{j=1}^{m}w_j^2$$ L2 - регуляризация

Новая функция, учитывающая штрафы

$$J(w)=\sum_{i=1}^{n}\left[-y^{(i)}\log(\phi(z^{(i)}))-(1-y^{(i)})\log(1-\phi(z^{(i)}))\right]+\frac{\lambda}{2}\|w\|^2$$

Дерево решений

Разбиение данных на подмножества, приводящему к самому большому приросту информации (получению однородных регионов решения)

Функция прироста информации:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

Меры неоднородности:

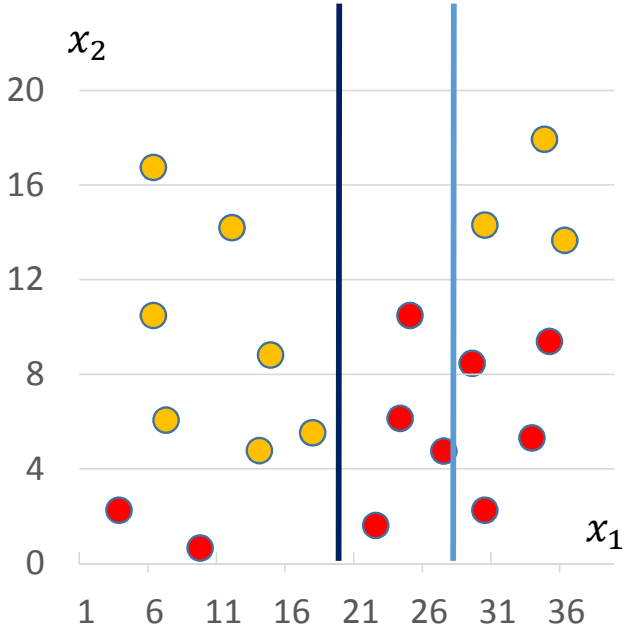
Энтропия: $I_G(t) = 1 - \sum_{i=1}^{c} p(i|t)^2$

Мера неопределенности Джини: $I_H(t) = -\sum_{i=1}^{c} p(i|t) log_2 p(i|t)$

Ошибка классификации: $I_E(t) = 1 - \max(p(i|t))$

$p(i|t)$ -доля образцов, принадлежащая классу i для узла t

Построение деревьев решений. Пример-1



$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

В качестве критерия взята ошибка классификации:

$$I_E(t) = 1 - \max(p(i|t))$$

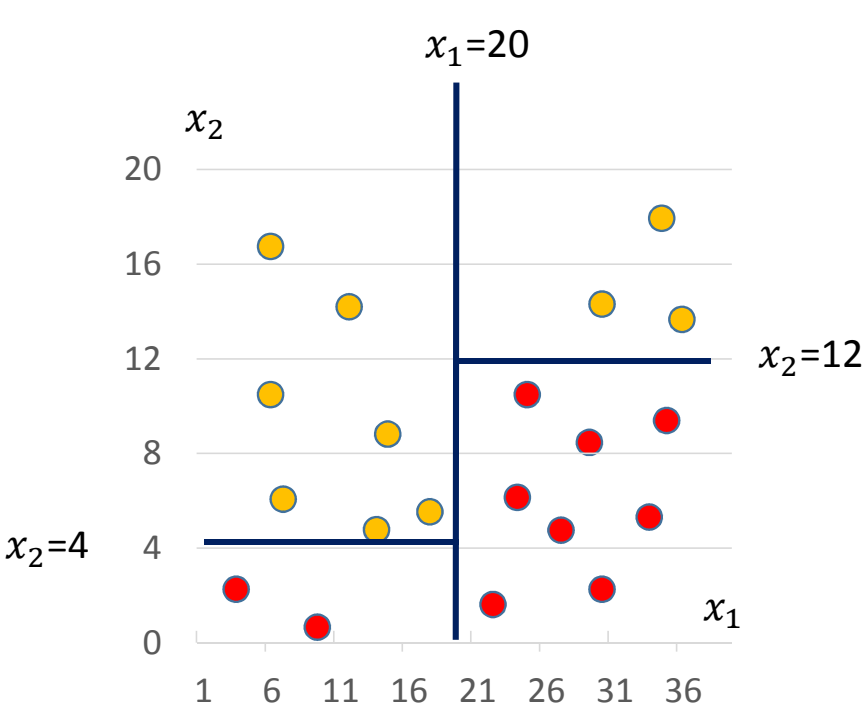
Неоднородность корневого узла:

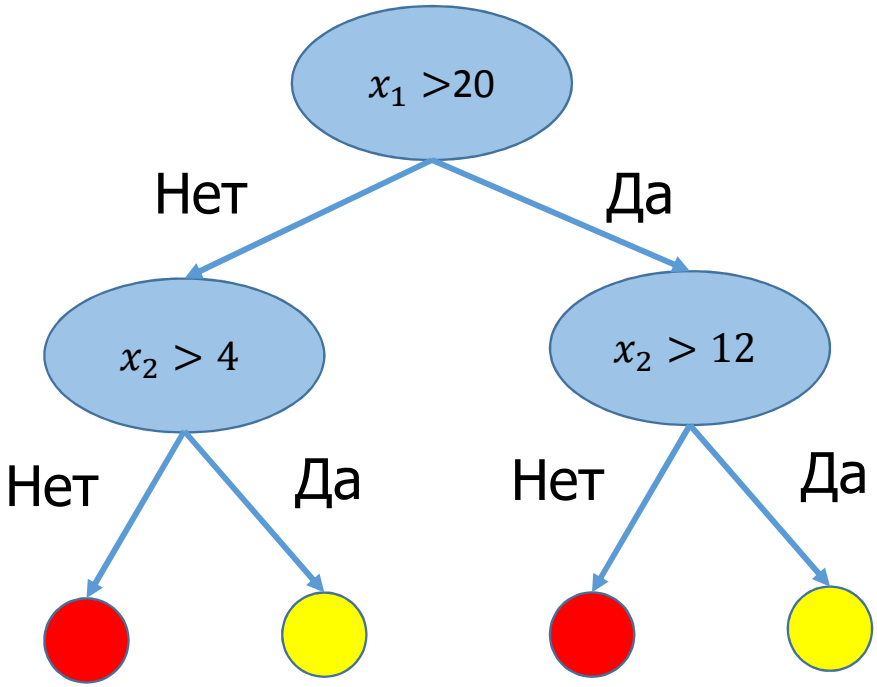
$$I(D_0) = 1 - \max\left(\frac{10}{20}, \frac{10}{20}\right) = 1 - 0.5 = 0.5$$

Для расщепления $x_1 = 20$: $IG(D_0, x_1 = 20) = 0.5 - \frac{9}{20}\left(1 - \frac{7}{9}\right) - \frac{11}{20}\left(1 - \frac{8}{11}\right) = 0.25$

Для расщепления $x_1 = 28$: $IG(D_0, x_1 = 28) = 0.5 - \frac{13}{20}\left(1 - \frac{7}{13}\right) - \frac{7}{20}\left(1 - \frac{4}{7}\right) = 0.05$

Построение деревьев решений. Пример-2





Примеры классов из Scikit-learn. Параметры

class sklearn.tree.DecisionTreeClassifier(criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, presort=False)

class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)

Хорошая модель должна использовать не параметры алгоритмов по умолчанию, а исследование результатов алгоритмов с разными параметрами!!!

Случайный лес. Выборки для обучения

#	A1	A2	A3	A4	A5
1					
2					
3					
4					
5					

Исходный набор: признаки

Выборка 1

#	A2	A4
1		
3		
4		

Выборка 2

#	A1	A3	A4
3			
5			

Выборка 3

#	A1	A5
1		
2		
3		

Выборка 4

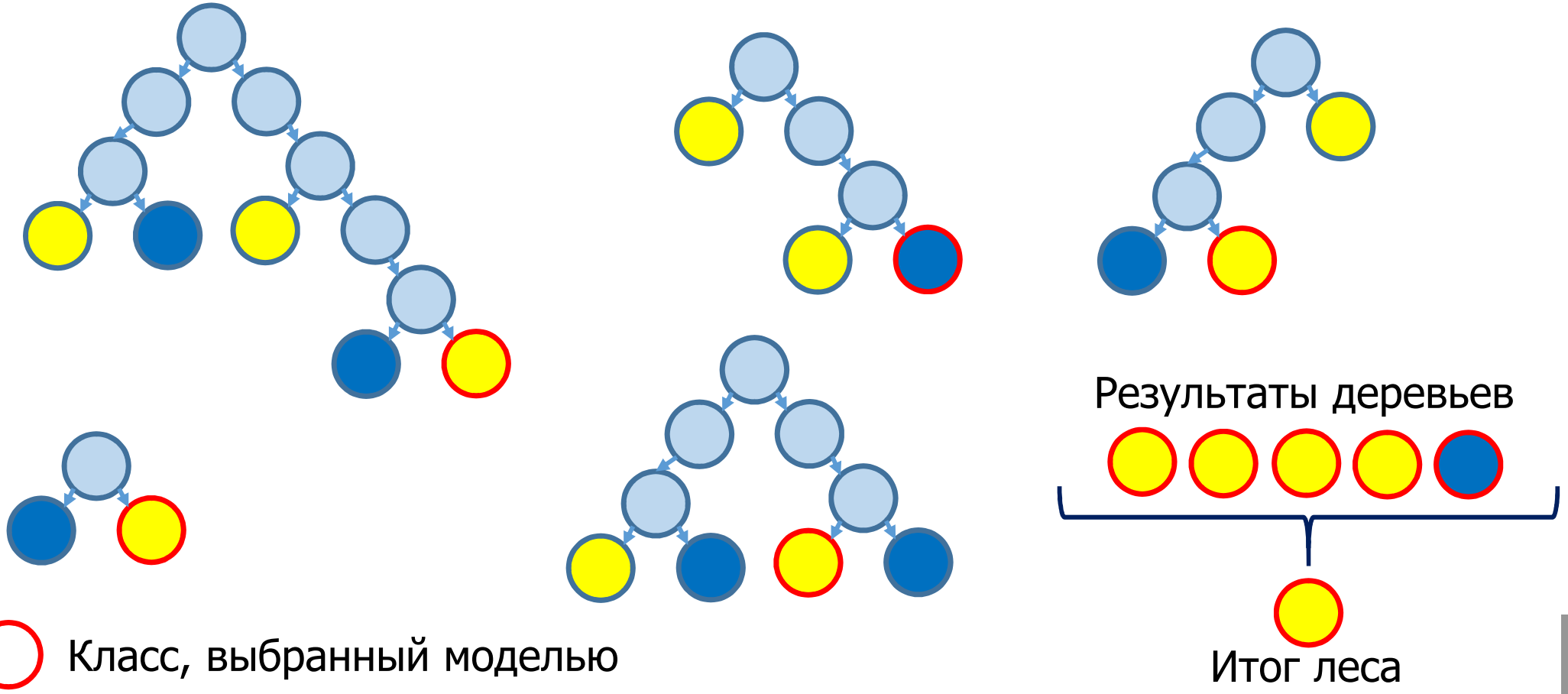
#	A2	A3	A5
2			
4			
5			

Выборка 5

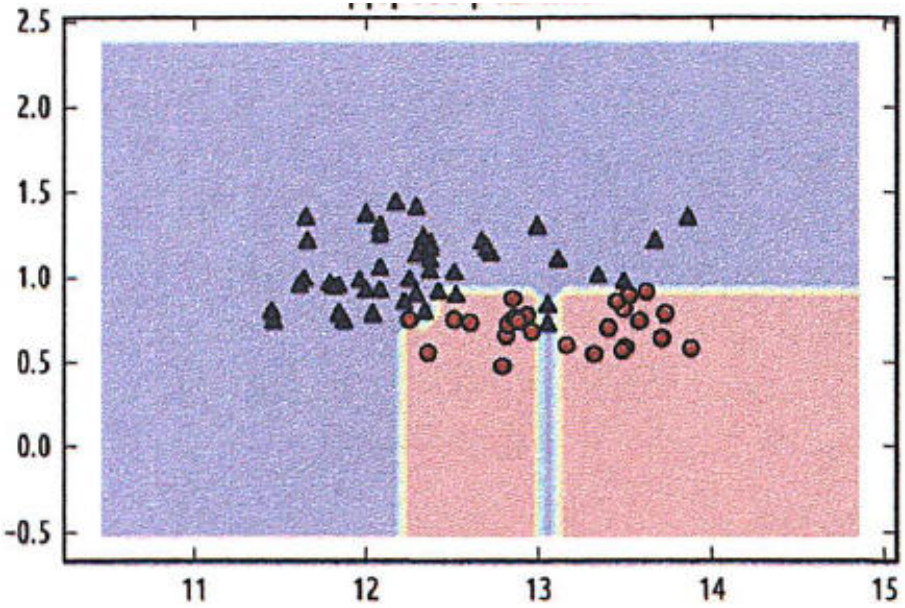
#	A1	A2
1		
2		

Случайный лес

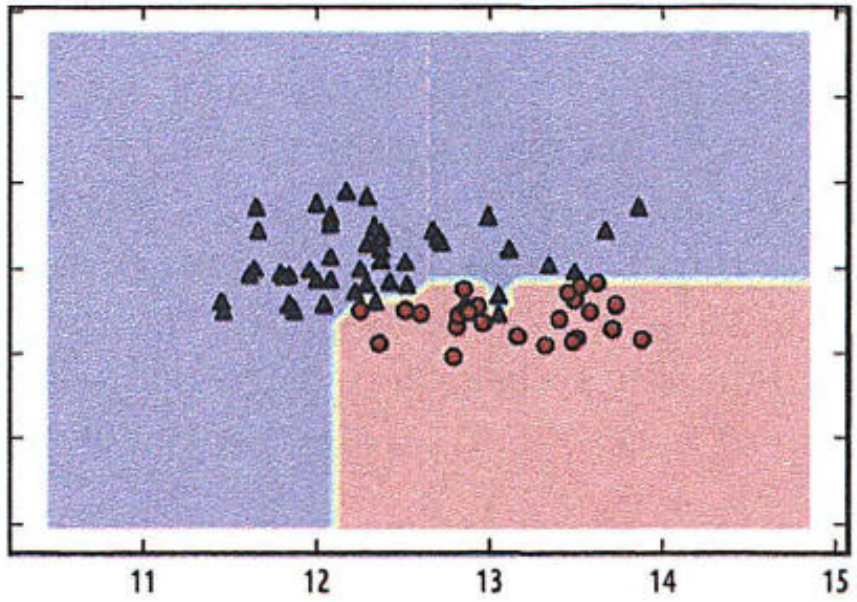
Объединение работы нескольких деревьев.
Мажоритарное голосование



Случайный лес



Дерево решений



Случайный лес

Определение важности признаков на случайном лесе

Основано на критерии прироста информации для всех деревьев в случайном лесе

$$IG(D_p, f)$$

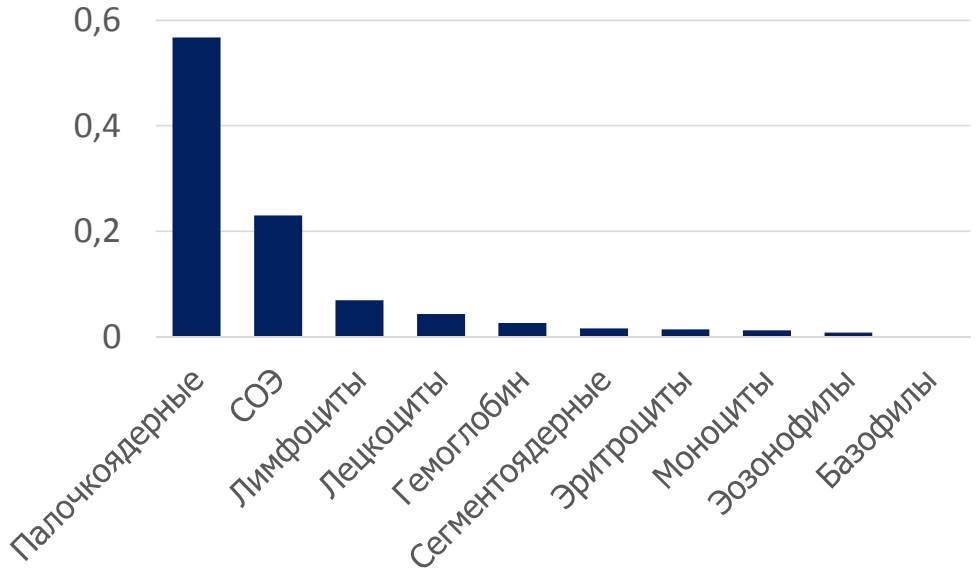
Пример: Оценка важности признаков для задачи оценки состояния тяжести заболевания по результатам общего анализа крови (из собств. практики)

АНАЛИЗ КРОВИ № _____
от 02.05.00
Фамилия, И.О. Л-в В.В.
Возраст 41
Участок
медицинская карта № 2284

	Результат	Норма	Единицы, подлежащие замене
Гемоглобин М Ж	121	130,0-160,0 / 120,0-140,0 г/л	13,0-16,0 / 12,0-14,0 г%
Эритроциты М Ж	3.84	4,0-5,0 / 3,9-4,7 $*10^{12}$/л	млн. в 1 мкл
Цветовой показатель	0.95	0,85-1,05	0,85-1,05
Среднее содержание гемоглобина (в 1 эритроците) МСН	31.51	30-35 пг	30-35 пг
Ретикулоциты	1	2-10 ‰	2-10 ‰
Тромбоциты	196	180,0-320,0 $*10^9$/л	180,0-320,0 тыс. в 1 мкл
Лейкоциты	18.8	4,0-9,0 $*10^9$/л	4,0-9,0 тыс. в 1 мкл
Миелоциты	-	-	-
Метамиелоциты	1	-	-
Палочкоядерные	21	1-6 0,040-0,300	1-6 40-300
Сегментоядерные	59	47-72 2,000-5,500	47-72 2000-5000
Эозинофилы	2	0,5-5 0,020-0,300	0,5-5 20-300
Базофилы	0	0-1 0-0,065	0-1 0-65
Лимфоциты	9	19-37 1,200-3,000	19-37 1200-3000
Моноциты	8	3-11 0,090-0,600	3-11 90-600
Плазматические клетки	-	-	-
Скорость (реакция) М Оседания эритроцитов Ж	61	2-10 мм/ч	2-10 мм/час
Гематокрит	37		

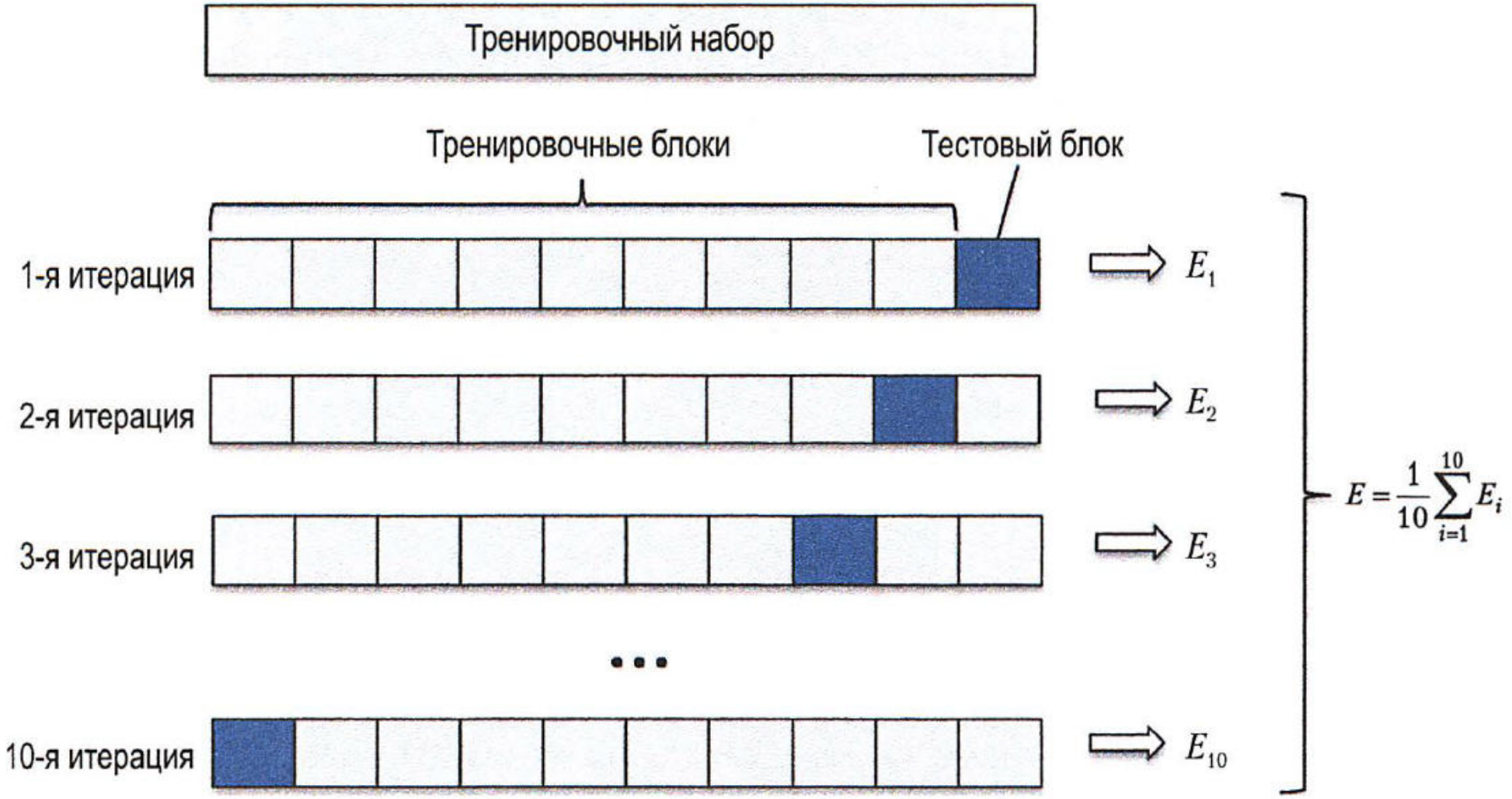
2
АНАЛИЗ КРОВИ № _____
от 03.05.00
Фамилия, И.О. Л-в В.В.
Возраст 41
Участок 6 отд
медицинская карта № 2284

	Результат	Норма	Единицы, подлежащие замене
Гемоглобин М Ж	110	130,0-160,0 / 120,0-140,0 г/л	13,0-16,0 / 12,0-14,0 г%
Эритроциты М Ж	3.75	4,0-5,0 / 3,9-4,7 $*10^{12}$/л	4,0-5,0 млн. в 1 мкл
Цветовой показатель	0.88	0,85-1,05	0,85-1,05
Среднее содержание гемоглобина (в 1 эритроците)	29.33	30-35 пг	30-35 пг
Ретикулоциты	10	2-10 ‰	2-10 ‰
Тромбоциты	210	180,0-320,0 $*10^9$/л	180,0-320,0 тыс. в 1 мкл
Лейкоциты	13	4,0-9,0 $*10^9$/л	4,0-9,0 тыс. в 1 мкл
Миелоциты	0	-	-
Метамиелоциты	0	-	-
Палочкоядерные	33	1-6 0,040-0,300	1-6 40-300
Сегментоядерные	52	47-72 2,000-5,500	47-72 2000-5000
Эозинофилы	1	0,5-5 0,020-0,300	0,5-5 20-300
Базофилы	0	0-1 0-0,065	0-1 0-65
Лимфоциты	3	19-37 1,200-3,000	19-37 1200-3000
Моноциты	11	3-11 0,090-0,600	3-11 90-600
Плазматические клетки	0	-	-
Скорость (реакция) М Оседания эритроцитов Ж	65	2-10 мм/ч	2-10 мм/час
Гематокрит	32		



Как изменилось состояние пациента?

К-блочная перекрестная проверка



Мажоритарное голосование в ансамбле

Номер модели	Пример 1	Пример 2	Пример 3	Пример 4	Пример 5	Количество ошибок
Модель 1	+	-	+	-	+	2
Модель 2	+	+	-	+	-	2
Модель 3	-	-	+	+	-	2
Модель 4	+	-	-	+	+	2
Модель 5	-	+	+	-	+	2
Мажорит. голосование	+	-	+	+	+	1

“+” – Правильно классифицированные примеры
“-” – Ошибочно классифицированные примеры

Пример мажоритарного голосования