Decision Trees

Логическая закономерность (Rule)

$$R_y \colon \mathrm{X} \to \{0, 1\}$$

Виды закономерностей

Пороговые условия:

$$R(\mathrm{x}) = [a \leq x_i \leq b]$$

Синдром пороговых условий:

$$R(\mathrm{x}) = \left[\sum_{j \in J} [a_j \leq x_j \leq b_j] \geq d\right]$$

При $d = |J|$ - конъюнкция условий.

Оценки правил (классификаторов)

		$\boldsymbol{y_i}$		
		$= \boldsymbol{y}$	$\neq \boldsymbol{y}$	
$\mathbf{R_y(x_i)}$	**1**	True positive (TP)	False positive (FP)	Positive predictive value, **Precision** $\frac{TP}{TP+FP}$
	0	False negative (FN)	True negative (TN)	
		True positive rate, **Recall** $\frac{TP}{TP+FN}$		**Accuracy** $\frac{TP+TN}{TP+FP+TN+FN}$

$$\boldsymbol{F_1 score} = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2\frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

Пример

		y_i		
		$= y$	$\neq y$	
$\mathbf{R_y(x_i)}$	**1**	0 (TP)	0 (FP)	Positive predictive value, **Precision** 0
	0	50 (FN)	950 (TN)	
		True positive rate, **Recall** 0		**Accuracy** 95%

$$\boldsymbol{F_1 score} = 0$$

Пример #2

<table>
<tr><td></td><td></td><td colspan="2">$\boldsymbol{y_i}$</td><td></td></tr>
<tr><td></td><td></td><td>$= \boldsymbol{y}$</td><td>$\neq \boldsymbol{y}$</td><td></td></tr>
<tr><td rowspan="2">$\mathbf{R_y(x_i)}$</td><td>1</td><td>50 (TP)</td><td>950 (FP)</td><td>Positive predictive value, Precision

5%</td></tr>
<tr><td>0</td><td>0 (FN)</td><td>0 (TN)</td><td></td></tr>
<tr><td></td><td></td><td>True positive rate, Recall

100%</td><td></td><td>Accuracy

5%</td></tr>
</table>

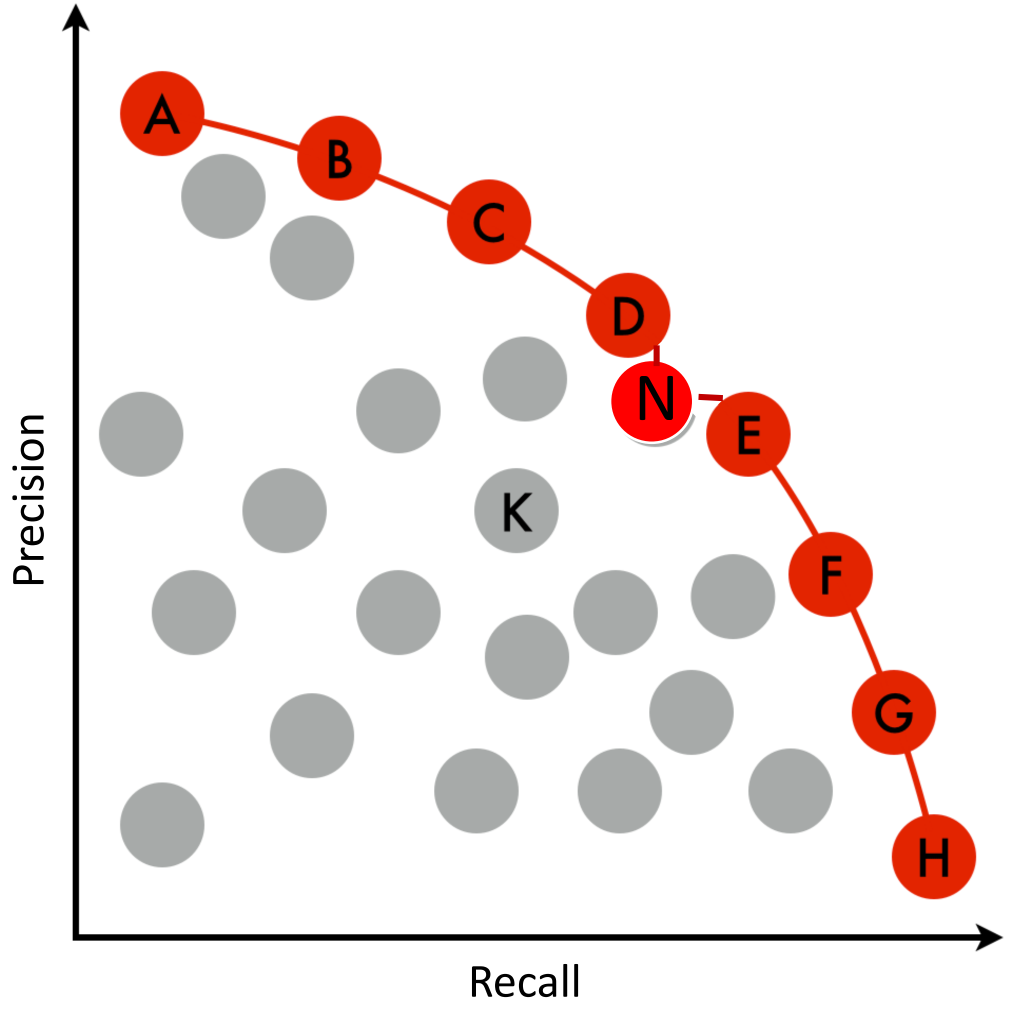
$$\boldsymbol{F_1 score} = 10\%$$

Пример #3

		y_i		
		$= y$	$\neq y$	
$\mathbf{R_y(x_i)}$	**1**	25 (TP)	475 (FP)	Positive predictive value, **Precision** 5%
	0	25 (FN)	475 (TN)	
		True positive rate, **Recall** 50%		**Accuracy** 50%

$$\boldsymbol{F_1 score} = 9\%$$

Эффективность по Парето



- Парето фронт, 1-я линия

N – тоже 1-я линия, K – 2-я

Эффективное по Парето правило – такое, что нет другого правила, у которого одновременно выше и precision и recall.

Оценки правил (классификаторов)

		$\boldsymbol{y_i}$		
		$= \boldsymbol{y}$	$\neq \boldsymbol{y}$	
$\mathbf{R_y(x_i)}$	**1**	True positive (TP)	False positive (FP)	Positive predictive value, **Precision** $\frac{TP}{TP+FP}$
	0	False negative (FN)	True negative (TN)	
		True positive rate, **Recall** $\frac{TP}{TP+FN}$	False positive rate $\frac{FP}{FP+TN}$	**Accuracy** $\frac{TP+TN}{TP+FP+TN+FN}$

$$\boldsymbol{F_1 score} = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2\,\frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

Пример #1

		$\boldsymbol{y_i}$		
		$= \boldsymbol{y}$	$\neq \boldsymbol{y}$	
$\mathbf{R_y(x_i)}$	**1**	0 (TP)	0 (FP)	Positive predictive value, **Precision** 0%
	0	50 (FN)	950 (TN)	
		True positive rate, **Recall** 0%	False positive rate 0%	**Accuracy** 95%

$$\boldsymbol{F_1 score} = 0$$

Пример #2

		y_i		
		$= y$	$\neq y$	
$R_y(x_i)$	**1**	50 (TP)	950 (FP)	Positive predictive value, **Precision** 5%
	0	0 (FN)	0 (TN)	
		True positive rate, **Recall** 100%	False positive rate 100%	**Accuracy** 5%

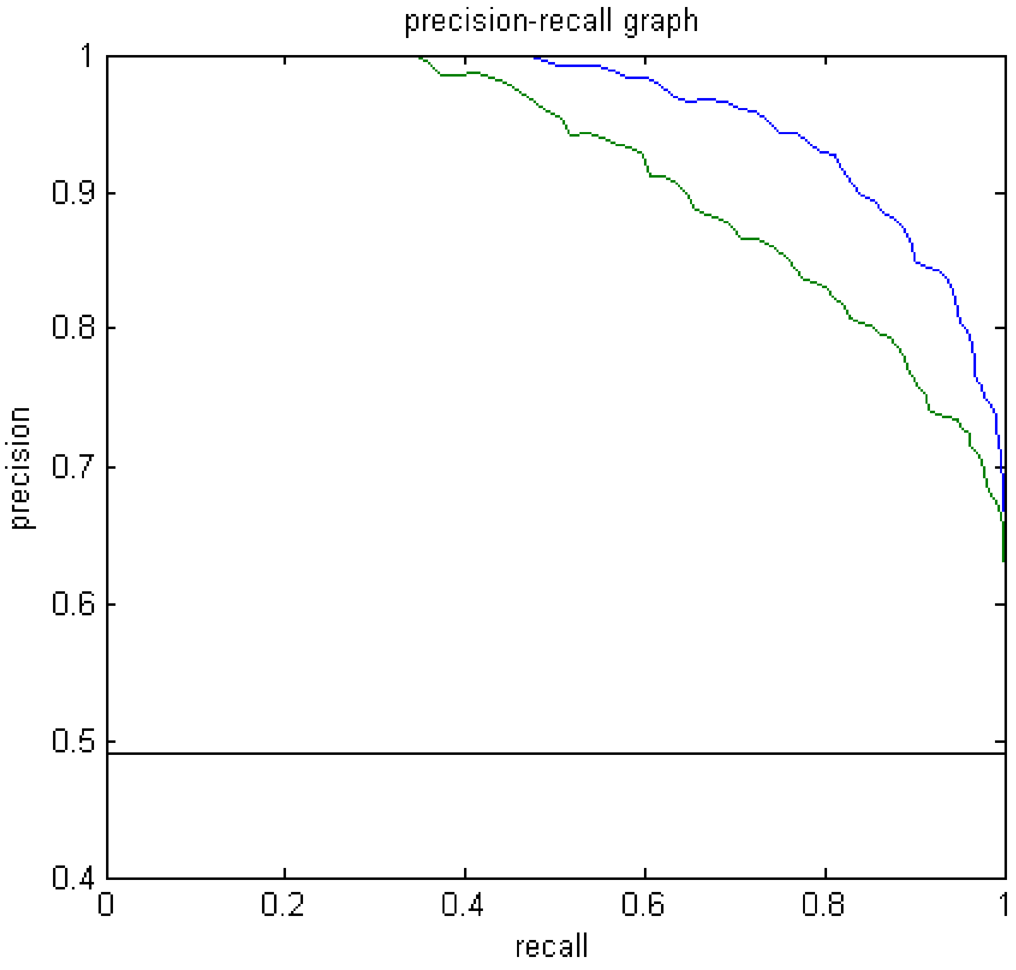
$$\boldsymbol{F_1 score} = 10\%$$

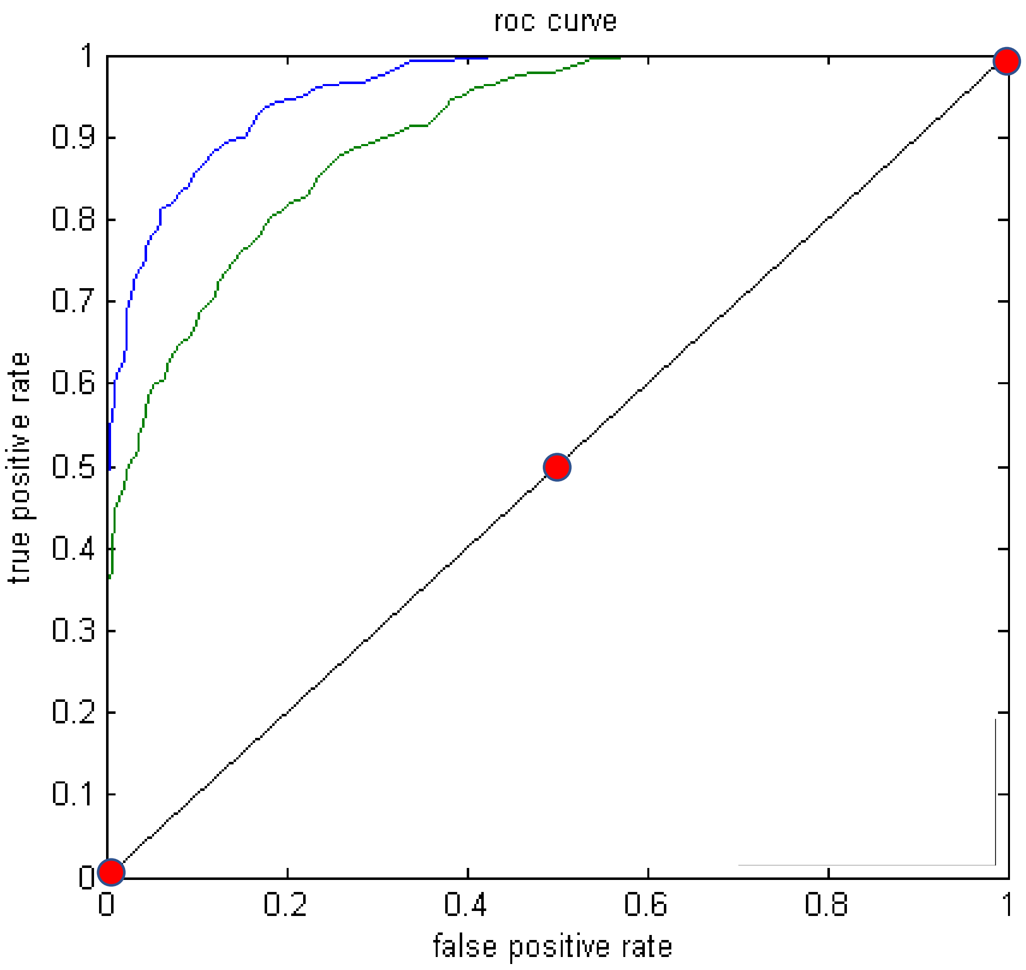
Пример #3

		y_i		
		$= y$	$\neq y$	
$\mathbf{R_y(x_i)}$	**1**	25 (TP)	475 (FP)	Positive predictive value, **Precision** 5%
	0	25 (FN)	475 (TN)	
		True positive rate, **Recall** 50%	False positive rate 50%	**Accuracy** 50%

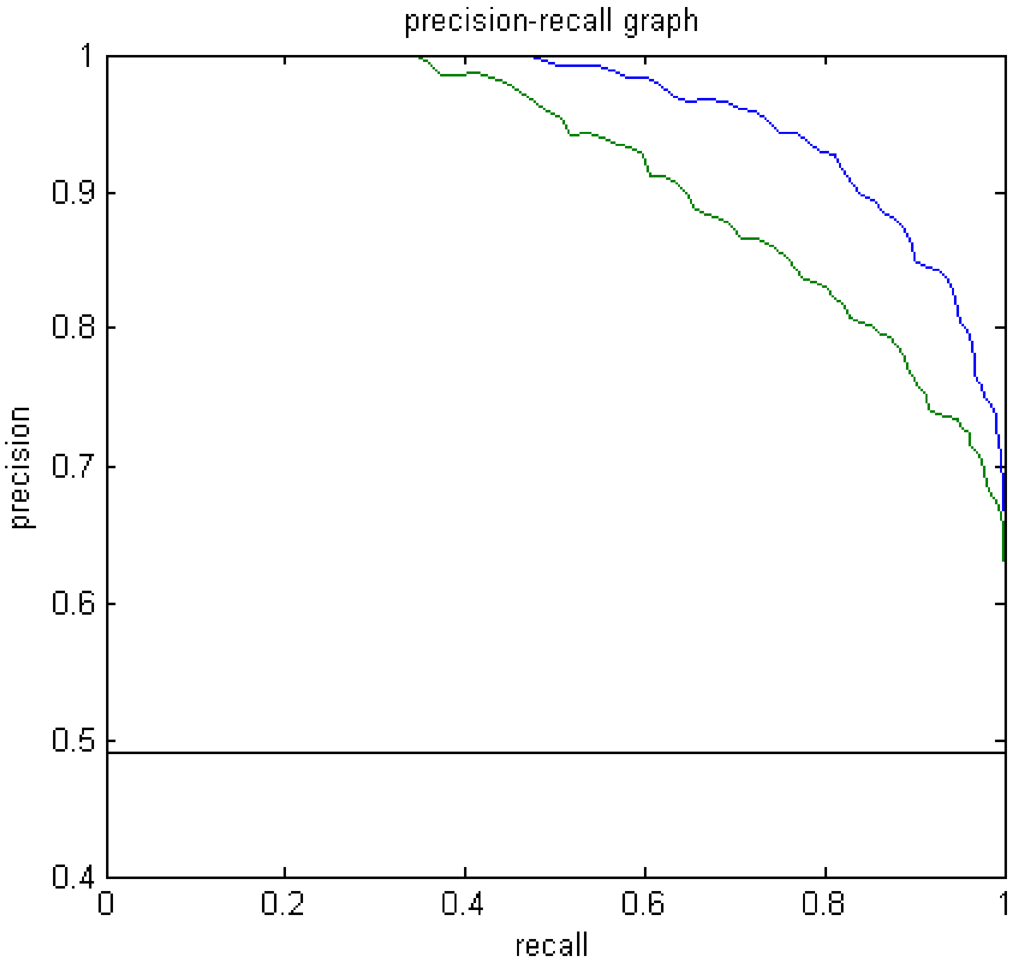
$$\boldsymbol{F_1 score} = 9\%$$

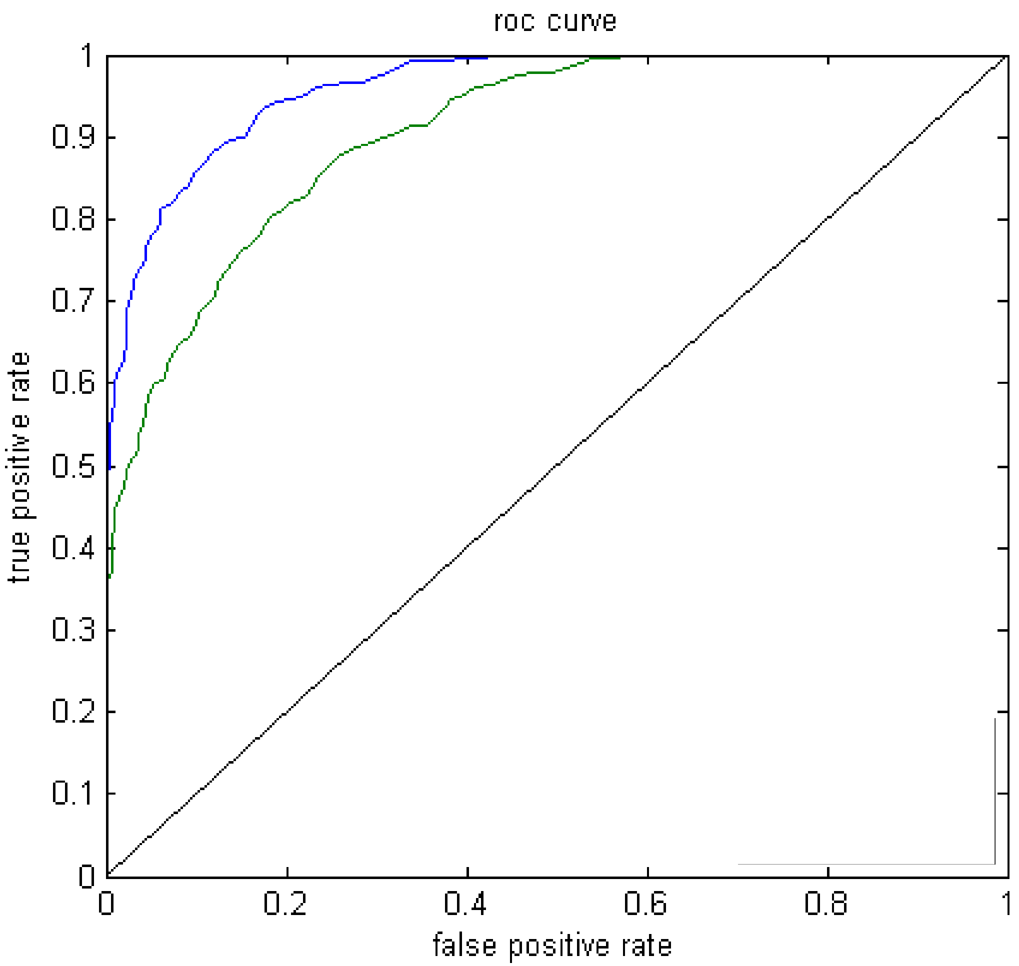
ROC (**R**eceiver **O**perating **C**haracteristic)





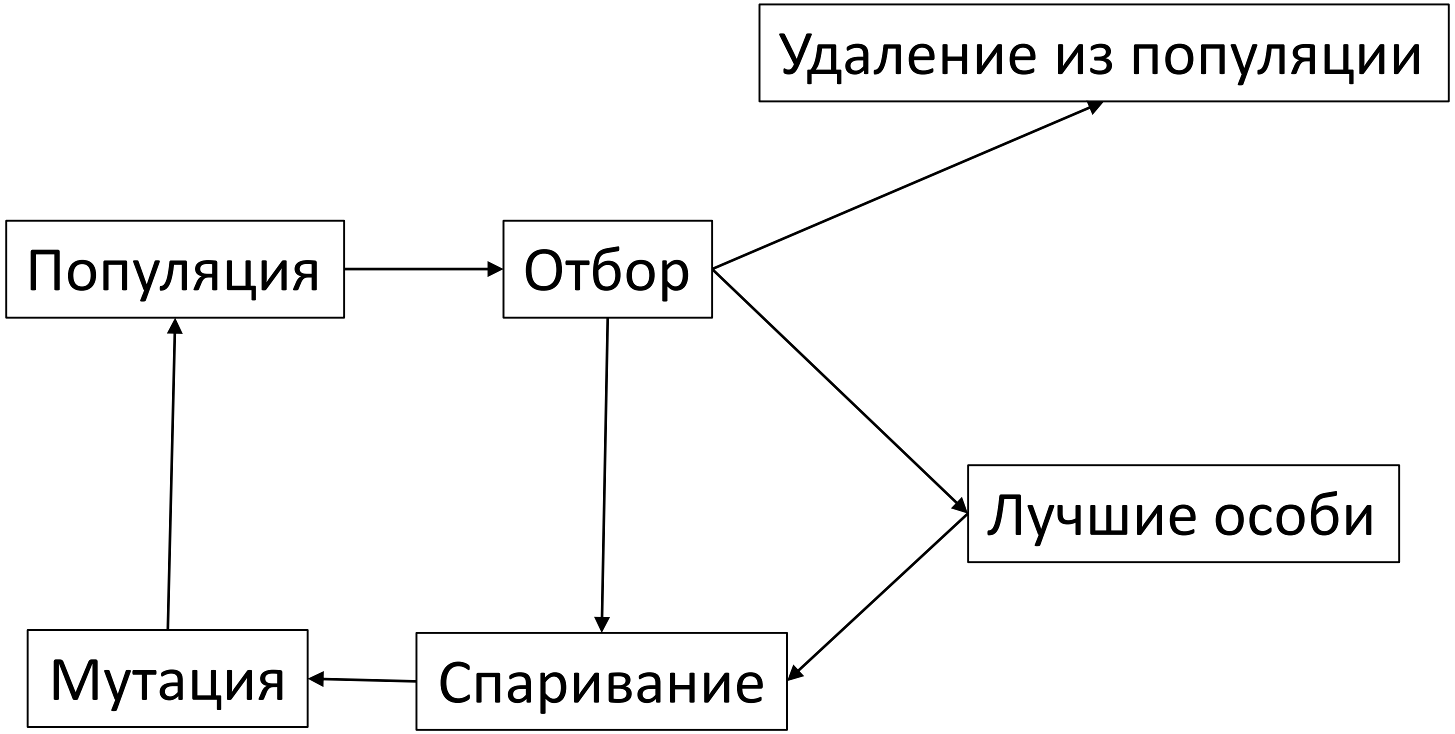
ROC (**R**eceiver **O**perating **C**haracteristic)



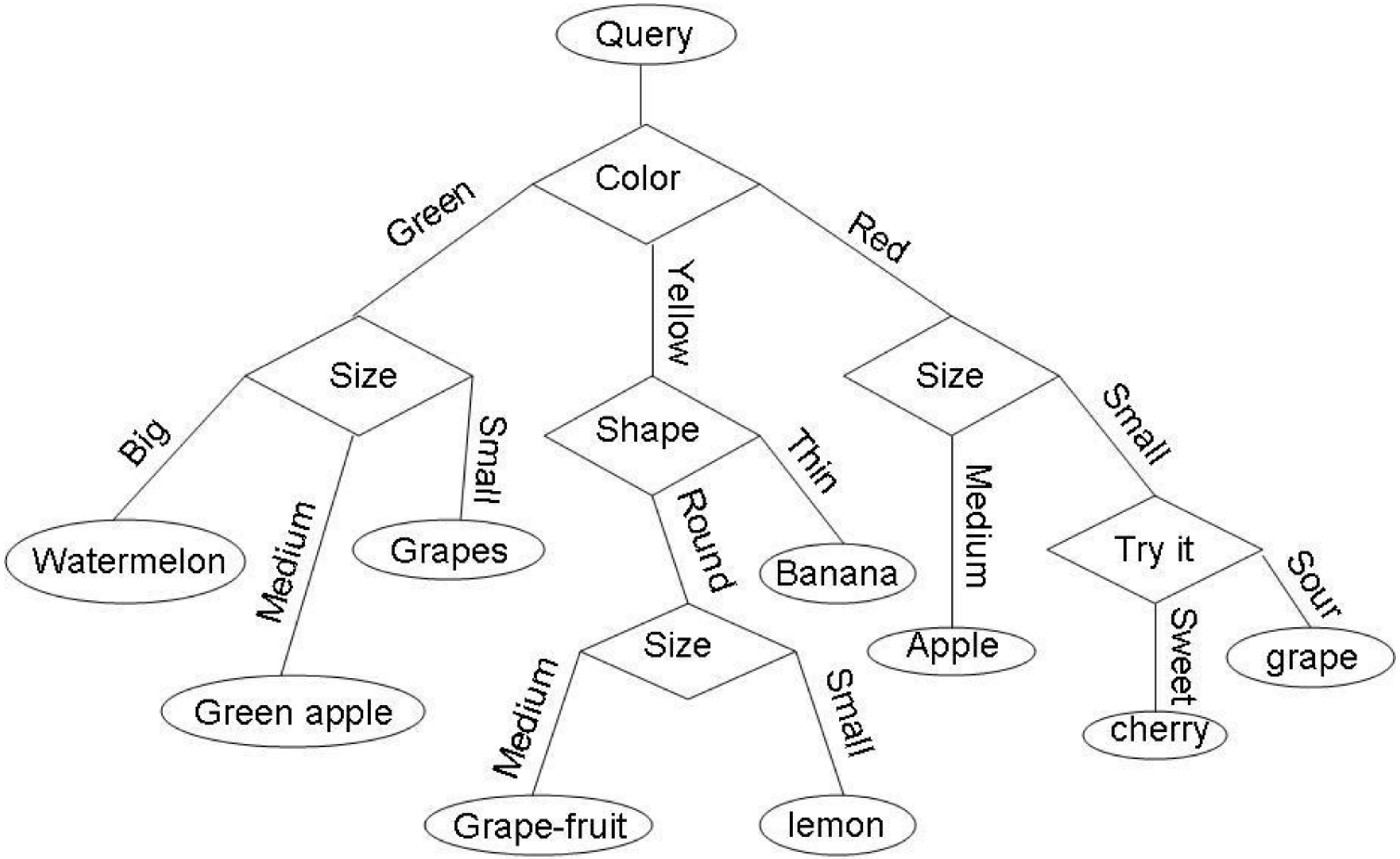


AUC – **A**rea **U**nder the ROC **C**urve

Отбор и генерация правил
Генетический алгоритм



Деревья решений



Оценки логических закономерностей

Information Gain (критерий информационного выигрыша):

$$IG(R) = H(\mathrm{X}) - \frac{|R^1|}{|\mathrm{X}|} H(R^1) - \frac{|R^0|}{|\mathrm{X}|} H(R^0)$$

$$H(\mathrm{X}) = -\sum_{y \in \mathrm{Y}} \frac{|\mathrm{x}_i : y_i = y|}{|\mathrm{X}|} \times \log_2 \frac{|\mathrm{x}_i : y_i = y|}{|\mathrm{X}|} \quad - \quad \text{энтропия}$$

Gini impurity (критерий Джини)

$$I_g(\mathrm{X}) = \sum_{y \in \mathrm{Y}:} \frac{|\mathrm{x}_i : y_i = y|}{\mathrm{X}} \frac{|\mathrm{x}_i : y_i \neq y|}{\mathrm{X}}$$

Вероятность того, что мы ошибемся, если будем элементам из множества приписывать класс согласно распределению классов в множестве.

Алгоритмы ID3 и C4.5
(Iterative Dichotomizer 3 и C4.5)

Автор: John Ross Quinlan

1. Если все объекты лежат в одном классе –> вернуть лист с этим классом.

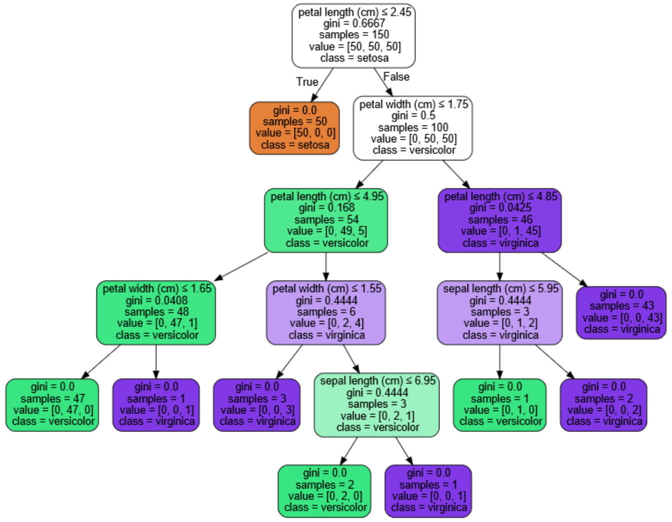
2. Найти правило с лучшим нормализованным информационным выигрышем или по Джини.

3. Правила с информационным выигрышем нет –> вернуть лист с мажоритарным классом.

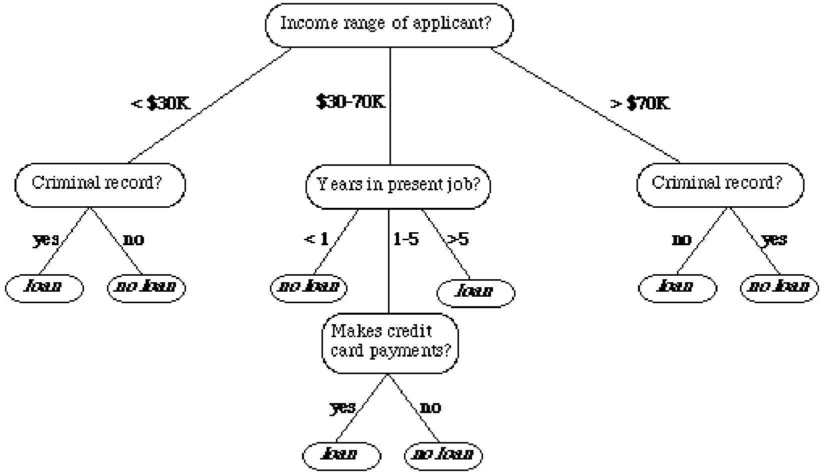
4. Разделить по этому правилу объекты.

5. Вызвать 1. для каждого листа, на который разделены объекты.

ID3



C4.5

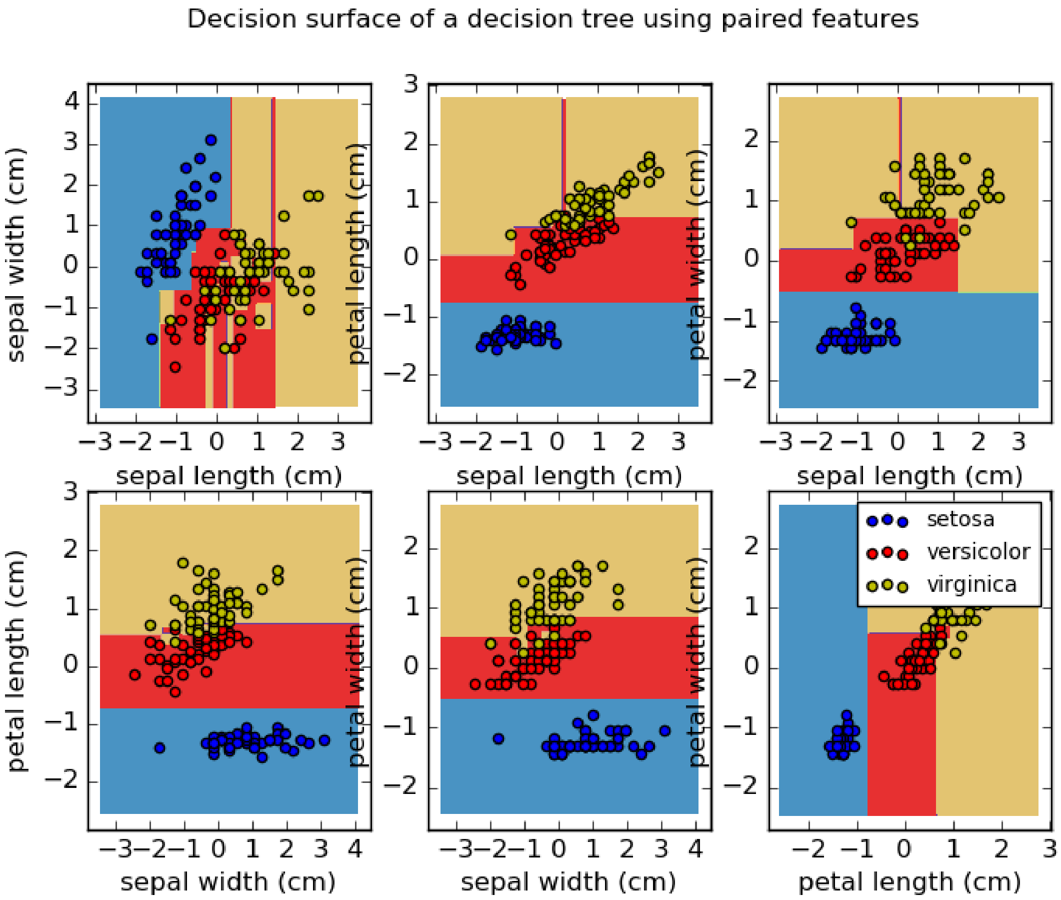


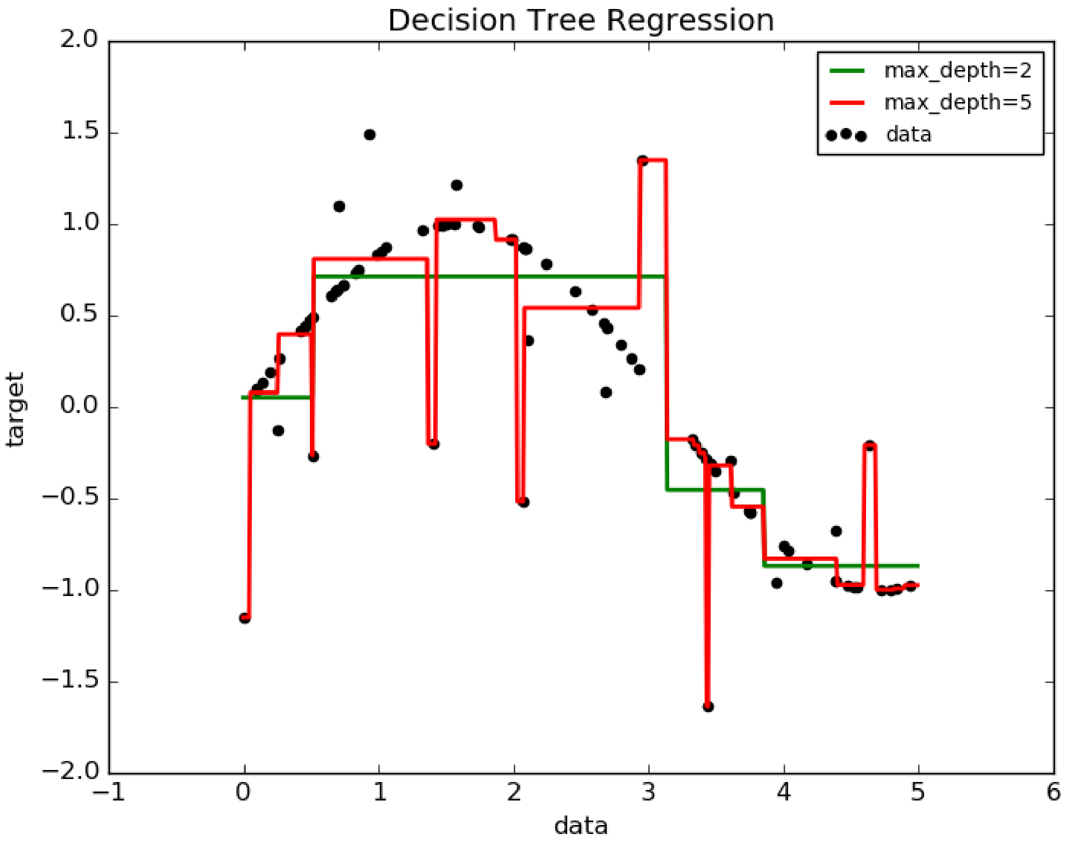
Алгоритм CART
(**C**lassification **a**nd **R**egression **T**rees)

Дисперсия:

$$\mathrm{V}(\mathrm{X}) = \frac{1}{|\mathrm{X}|^2} \sum_{\mathrm{x}_i \in \mathrm{X}} \sum_{\mathrm{x}_j \in \mathrm{X}} \frac{1}{2} \left(y_i - y_j\right)^2$$

Пример работы





Regularization and Train\Test

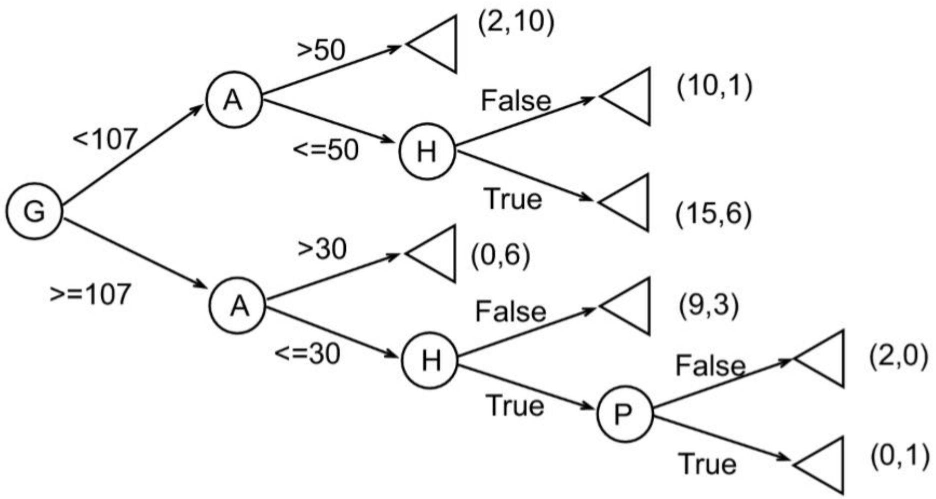
Выделим долю точек ($\approx 10 - 20\%$) в отдельную выборку - Test.

Обучать будем на оставшихся – Train.

По метрике на Test можно оптимизировать параметры модели (дерева), например – максимальную глубину.

Oblivious Decision Trees
(Небрежные решающие деревья)

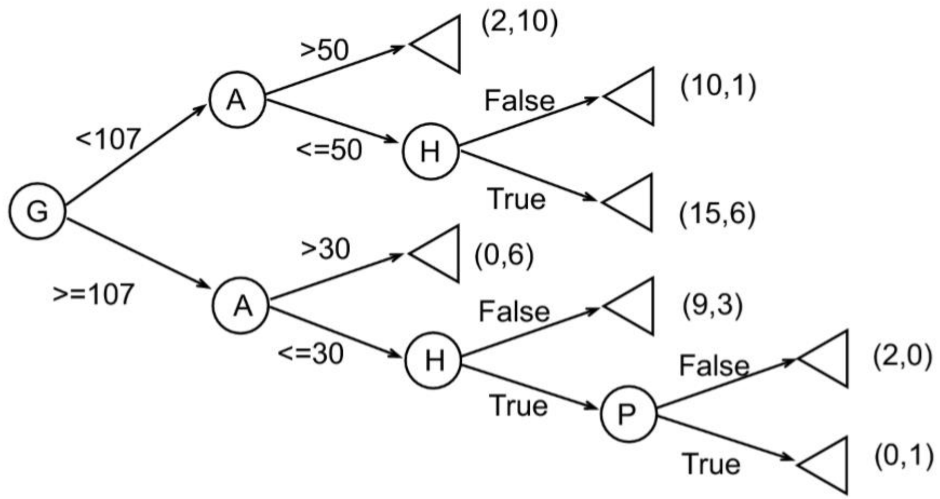
На одном уровне – одинаковые признаки.



(G)lucose level
(A)ge
(H)ypertension
(P)regnancy

Oblivious Decision Trees
(Небрежные решающие деревья)

На одном уровне – одинаковые признаки.



(G)lucose level
(A)ge
(H)ypertension
(P)regnancy

MatrixNet (Yandex) – Oblivious Decision Trees, где правило для разделения тоже одинаковое на каждом уровне. В этом случае порядок правил не важен и оценку и удаление признаков легко организовать.

Random Forests (Случайные леса)

Обучим много деревьев на «случайных данных»

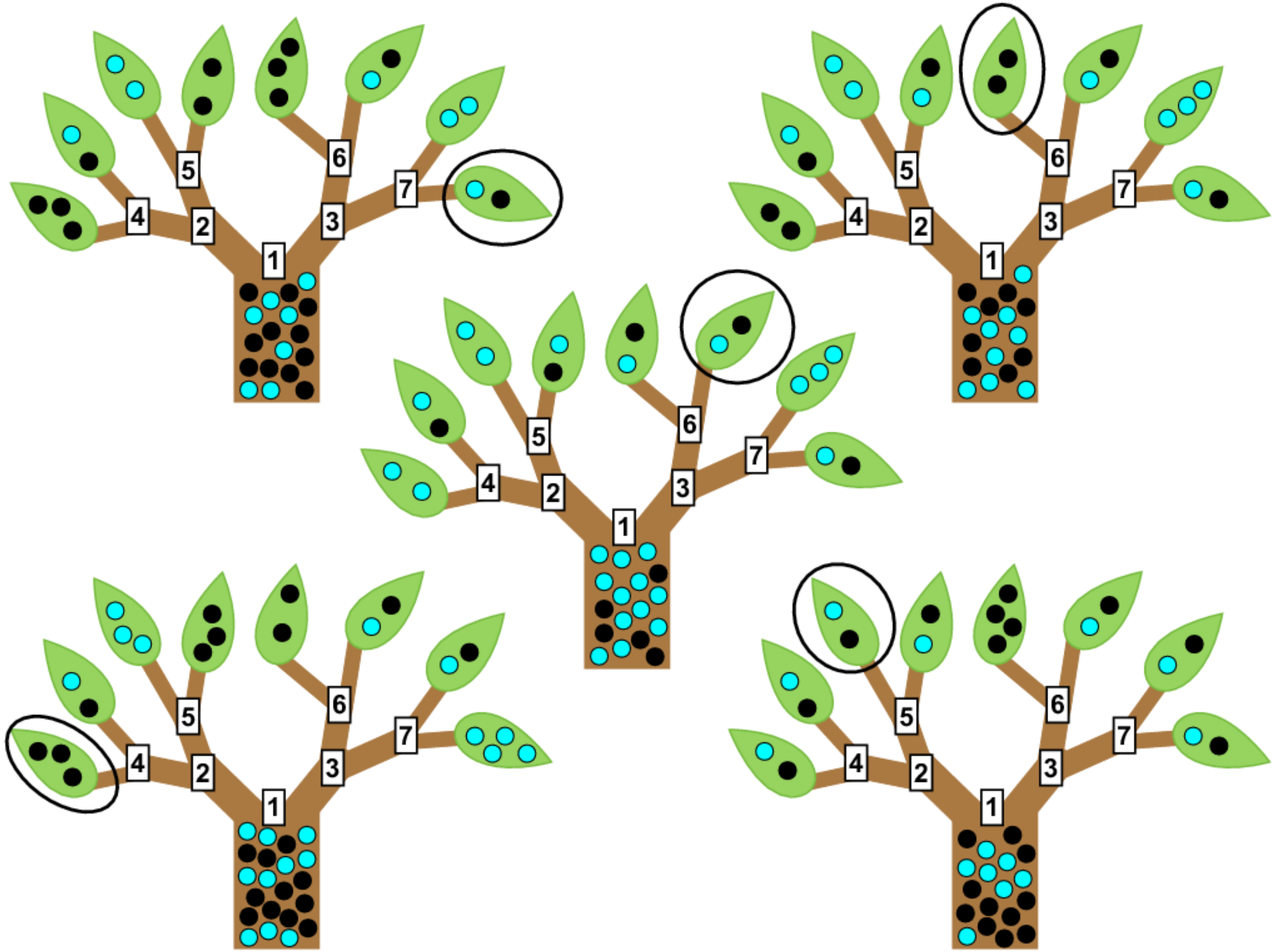
Случайные данные:

1) Датасеты с помощью **bagging (b**ootstrap **agg**regat**ing)** – вытаскиваем из исходного датасета столько же данных, но с повторениями.

2) Берем случайное подмножество признаков.

Результат – голосование, вероятность, логарифм вероятностей.

Параметры – количество используемых примеров и признаков для построения каждого дерева.



Немножко истории статистики

Sampling Случайная выборка

Jackknife resampling
(John Tukey, 1958)
«like a Boy Scout's jackknife, it is a "rough and ready" tool that can solve a variety of problems even though specific problems may be more efficiently solved with a purpose-designed tool»

Выкидываем поочередно по одному примеру

Bootstrapping (Bootstrap resampling)
(Bradley Efron, 1979)
«to pull oneself up by one's bootstraps»

Набираем выборку того же размера с повторениями из исходной

Bagging (**B**ootstrap **AGG**regat**ING**)
(Leo Breiman, 1994)

Используем разные bootstrap* выборки для обучения