Методы машинного обучения с локальной интерпретируемостью результатов

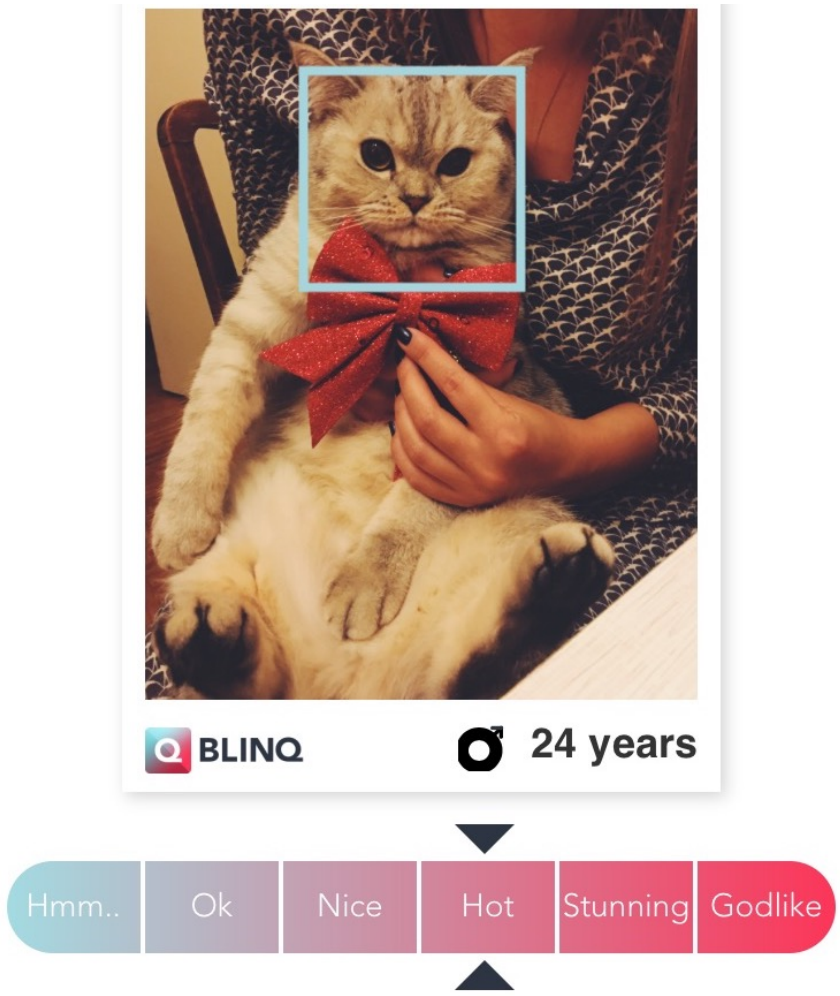
Кашницкий Юрий Савельевич

Высшая Школа Экономики

30 ноября 2016 г.

Black Box в машинном обучении

Почему нейронная сеть говорит, что коту 24 года?



Локальная интерпретируемость

Мотивация

- Странные ошибки модели
- Проверка, что вообще алгоритм “выучил”
- Отладка алгоритмов
- Правовые вопросы машинного обучения

Определение

Способность алгоритма выводить правила, объясняющие ту или иную классификацию любого тестового примера, назовем локальной интерпретируемостью этого алгоритма.

Сразу немного скепсиса

Пост “The Myth of Model Interpretability” Kdnuggets

- Четкого определения интерпретируемости нет
- Если признаков много, веса линейной модели мало о чем говорят
- Веса могут быть “не того знака”
- TF-IDF, производные признаки, PCA и т.д. ухудшают картину
- ...

Существующие методы классификации

Методов классификации в машинном обучении очень много, рассмотрим основные подходы, обратив внимание на:

- локальную интрепретируемость результатов классификации
- качество классификации
- вычислительную сложность

Существующие методы классификации II

- Деревья решений
  - \+ Хорошая интерпретируемость
  - – Чаще всего невысокое качество классификации
  - \+ Очень низкая вычислительная сложность
- Методы, основанные на правилах
  - \+ Хорошая интерпретируемость
  - – Качество классификации лучше, чем у деревьев, но все еще невысокое
  - – Как правило, вычислительно сложны

Существующие методы классификации III

- Метрические методы
  - \+ Относительно хорошая интерпретируемость
  - – Низкое качество классификации
  - \+ Вычислительная сложность невысокая
- Композиции алгоритмов (случайный лес, бустинг, стекинг и блендинг моделей, нейронные сети)
  - – Плохо интерпретируемы
  - \+ Высокое качество классификации в сложных задачах (опыт соревнований Kaggle)
  - – Очень большая вычислительная сложность

Существующие методы классификации IV

- Линейные методы
  - \+ Относительно хорошая интерпретируемость (если признаков не много)
  - – Как правило, низкое качество классификации (кроме случая большого числа разреженных признаков)
  - \+ Низкая вычислительная сложность
- Метод опорных векторов (нелинейный)
  - – Плохая интерпретируемость
  - \+ Высокое качество классификации
  - – Большая вычислительная сложность

Предлагаемое решение (далее в слайдах)

Предлагаемое решение

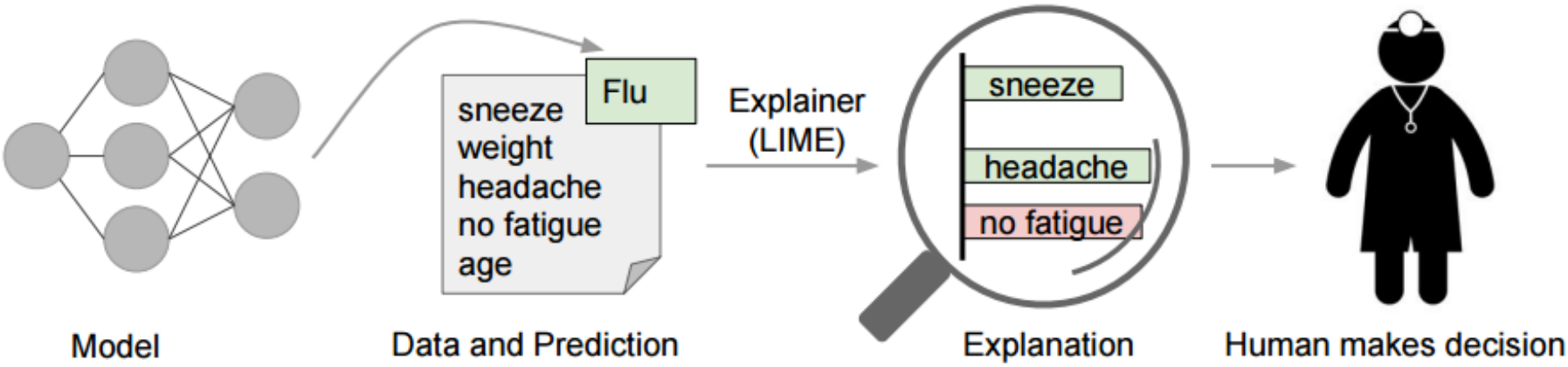
классификация на основе правил для данных со сложной структурой

+ Хорошая интерпретируемость
+ Работа со сложно структурированными данными
+ Высокое качество классификации
– Большая вычислительная сложность

Попытка универсальной интерпретации

Алгоритм LIME

Local Interpretable Model-Agnostic Explanations



+ Локальное линейное приближение
+ Любые (Python) модели
+ Те же проблемы с интерпретацией, если признаков много

За лесом деревьев не видно

Деревья решений:

+ Легко интерпретируются (и то если неглубокие) Одно (короткое) правило для каждого тестового примера
– Низкое качество классификации/регрессии

Случайный лес:

+ Высокое качество классификации/регрессии
– Плохая интерпретируемость Набор (длинных) правил для каждого тестового примера или оценка важности признаков

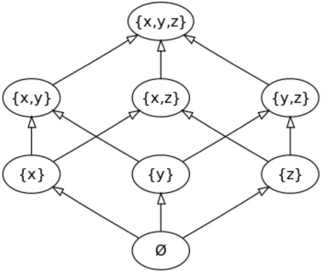
Идея: Что-то между ними?

Идея нахождения лучших правил

В деревьях жадно оптимизируется некоторый функционал (прирост информации, неопределенность Джини, дисперсия вокруг среднего).

Идея: Что если явно искать лучшие правила для каждого тестового примера?

Надо ли перебирать все $\approx 2^{|num\_features|}$ правил?



Нет!

Частые замкнутые множества признаков

Поиск частых замкнутых множеств признаков – известное направление в Data Mining.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Множество {Beer, Bread, Diaper} – частое, если ограничение на поддержку t = 0.4 (наблюдается в 2 из 5 транзакциях).

Частые замкнутые множества признаков II

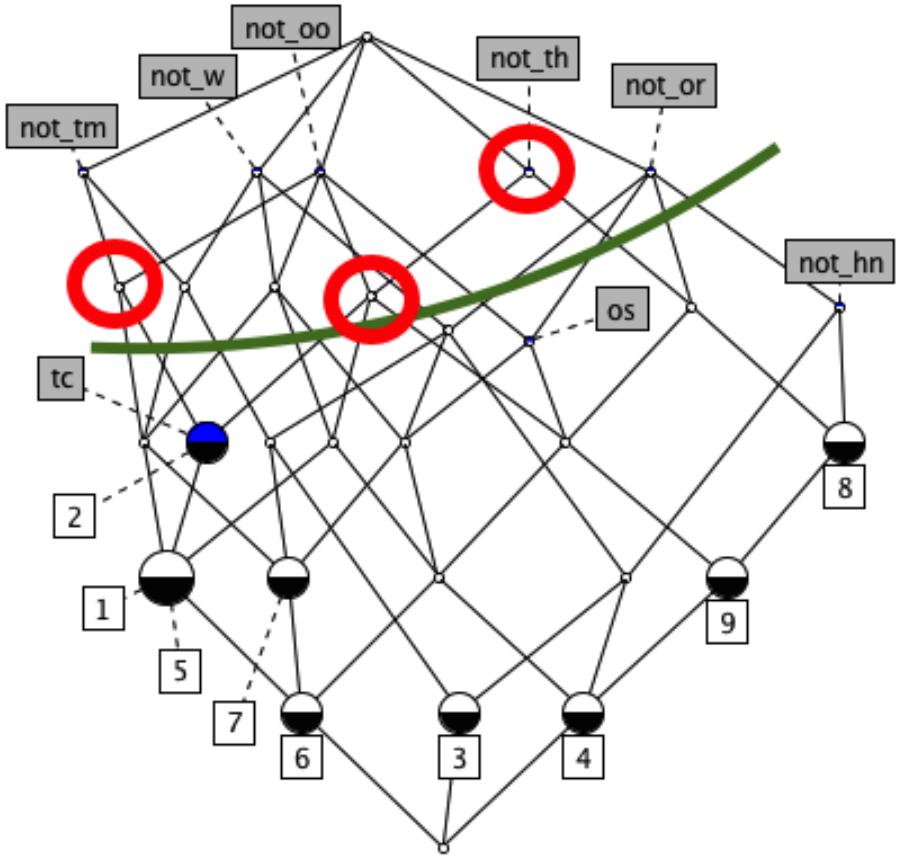
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Набор признаков называется замкнутым, если никакое из его непосредственных надмножеств не имеет ту же поддержку.
Набор {Beer, Bread, Diaper} замкнут.

Идея алгоритма

Поиск правил среди частых замкнутых множеств признаков.
Под (локальной) интерпретируемостью будем понимать среднее число признаков в посылке правила.



Пример

Классический игрушечный пример задачи классификации (Mitchell)

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	Normal	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	High	Strong	Yes
D8	Sunny	Mild	Normal	Weak	No
D9	Sunny	Hot	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Cool	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

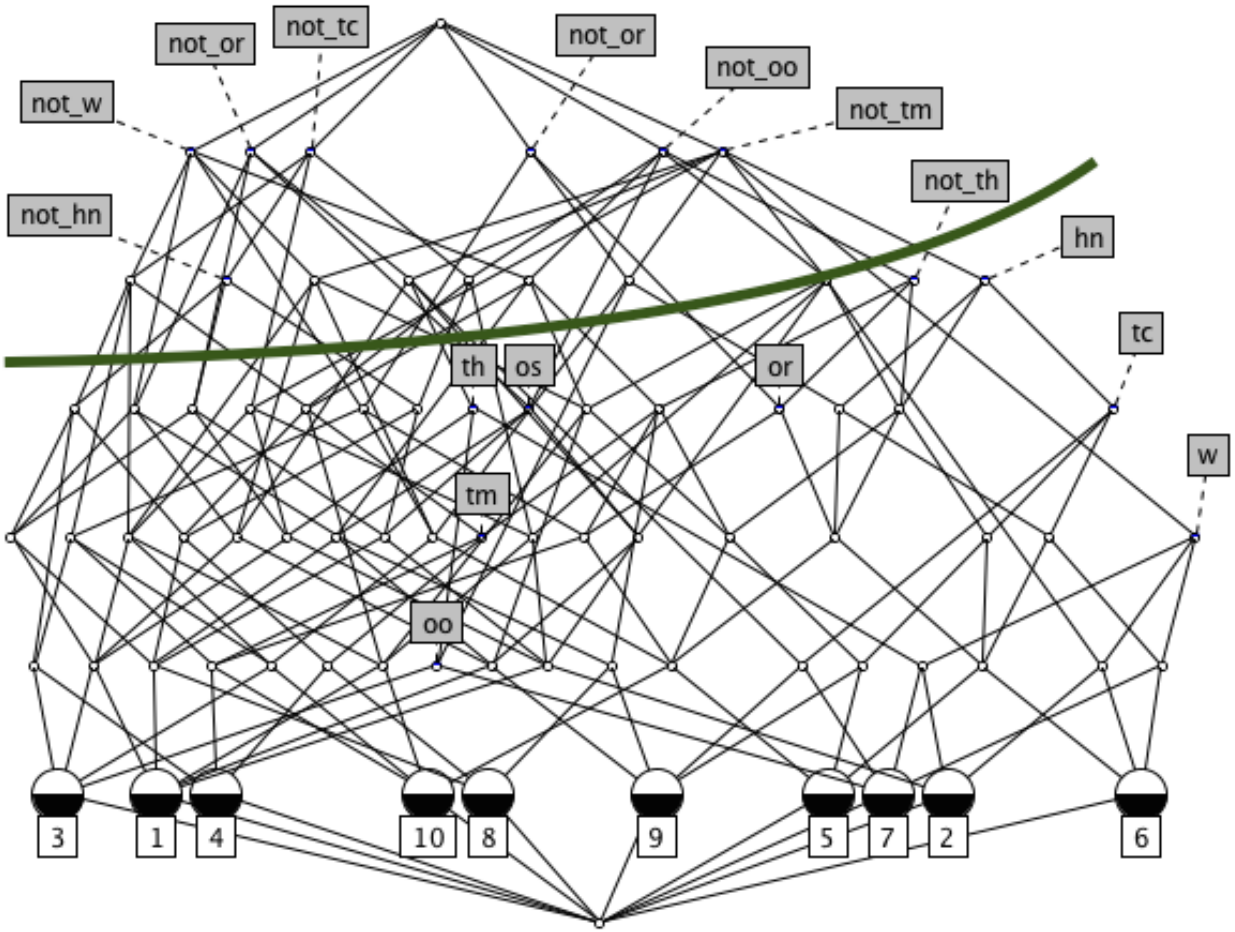
Пример II

То же с бинаризованными признаками и их отрицаниями.
Доля положительных объектов – 0.6.

G\M	os	¬os	oo	¬oo	or	¬or	th	¬th	tm	¬tm	tc	¬tc	¬hn	hn	w	¬w	play
1	×			×		×	×			×		×	×			×	
2	×			×		×	×			×		×	×		×		
3		×	×			×	×			×		×	×			×	×
4		×		×	×			×	×			×	×			×	×
5		×		×	×			×		×	×			×		×	×
6		×		×	×			×		×	×			×	×		
7		×	×			×		×		×	×			×	×		×
8	×			×		×		×	×			×	×			×	
9	×			×		×		×		×	×			×		×	×
10		×		×	×			×	×			×		×		×	×
11	×			×		×		×	×			×		×	×		?
12		×	×			×		×	×			×	×		×		?
13		×	×			×	×			×		×		×		×	?
14		×		×	×			×	×			×	×		×		?

Пример III

Решетка формальных понятий и ограничение на минимальную относительную поддержку 0.4 (гиперпараметр).



Пример IV

10 “лучших” классифицирующих правил. В заключениях правил – доли положительных примеров, под них “попадающих”.

	Правило	Неопределенность Джини
1	$os, \neg tc, \neg hn \to 0$	0.171
2	$\neg os, \neg w \to 1$	0.267
3	$\neg oo, \neg tm, w \to 0$	0.3
4	$os, \neg tc, \neg hn, \neg w \to 0$	0.3
5	$os, th, \neg hn \to 0$	0.3
6	$os \to 0.25$	0.317
7	$\neg oo, \neg tc, \neg hn \to 0.25$	0.317
8	$\neg or, \neg tc, \neg hn \to 0.25$	0.317
9	$\neg os \to 0.83$	0.317
10	$or, \neg th, \neg w \to 1$	0.343

Пример V

3 лучших (по неопределенности Джини) правила для классификации тестового примера Outlook=sunny, Temperature=mild, Humidity=normal, Windy=true

$os \rightarrow 0.25$	0.317
$\neg oo \rightarrow 0.5$	0.4
$\neg th, hn \rightarrow 0.5$	0.4

Усредняя заключения правил, получаем, что пример классифицируется отрицательно: $\frac{1}{3}(0.25 + 0.5 + 0.5) \approx 0.4 < 0.6$

Результаты на данных репозитория UCI II

Данные	DT acc	RF acc	kNN acc	CoLiBRi acc	DT F1	RF F1	kNN F1	CoLiBRi F1
audiology	0.75	0.8	0.63	0.79*	0.71	0.74	0.58	0.74
breast-cancer	0.63	0.66	0.76	0.65	0.58	0.63	0.75	0.61
breast-wisc	0.7	0.74	0.73	0.76	0.45	0.42	0.38	0.44*
car	0.75	0.78*	0.71	0.79	0.75	0.76	0.71	0.76
hayses-roth	0.84*	0.83*	0.49	0.86	0.84*	0.82	0.49	0.85
lymph	0.8	0.83	0.86	0.83	0.77	0.85	0.84*	0.84*
mol-bio-prom	0.78	0.83	0.83	0.82*	0.78	0.84	0.8	0.83*
nursery	0.64	0.65	0.72	0.65	0.62	0.62	0.7	0.62
primary-tumor	0.41	0.46	0.41	0.45*	0.37	0.41	0.37	0.4*
solar-flare	0.7*	0.7*	0.63	0.72	0.67	0.69*	0.6	0.71
soybean	0.91*	0.91*	0.92	0.91*	0.91*	0.93	0.92*	0.91*
spect-train	0.61	0.69	0.68	0.7	0.34	0.36	0.23	0.38
tic-tac-toe	0.79	0.79	0.85	0.78	0.82	0.86	0.89	0.85

Неопределенность Джини и длины посылок правил

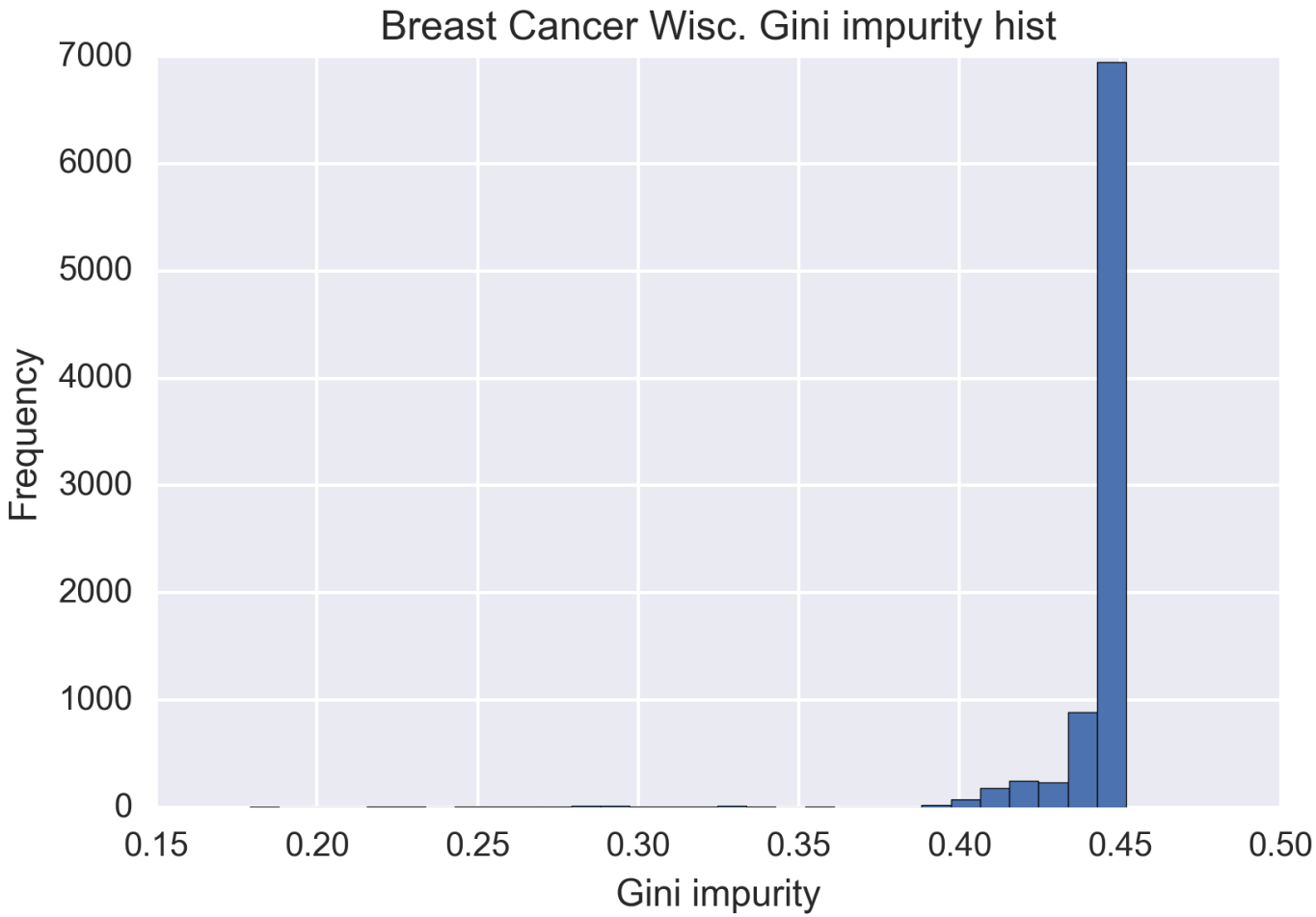
Набор Breast Cancer Wisconsin

Принцип Оккама

Распределение неопределенности Джини

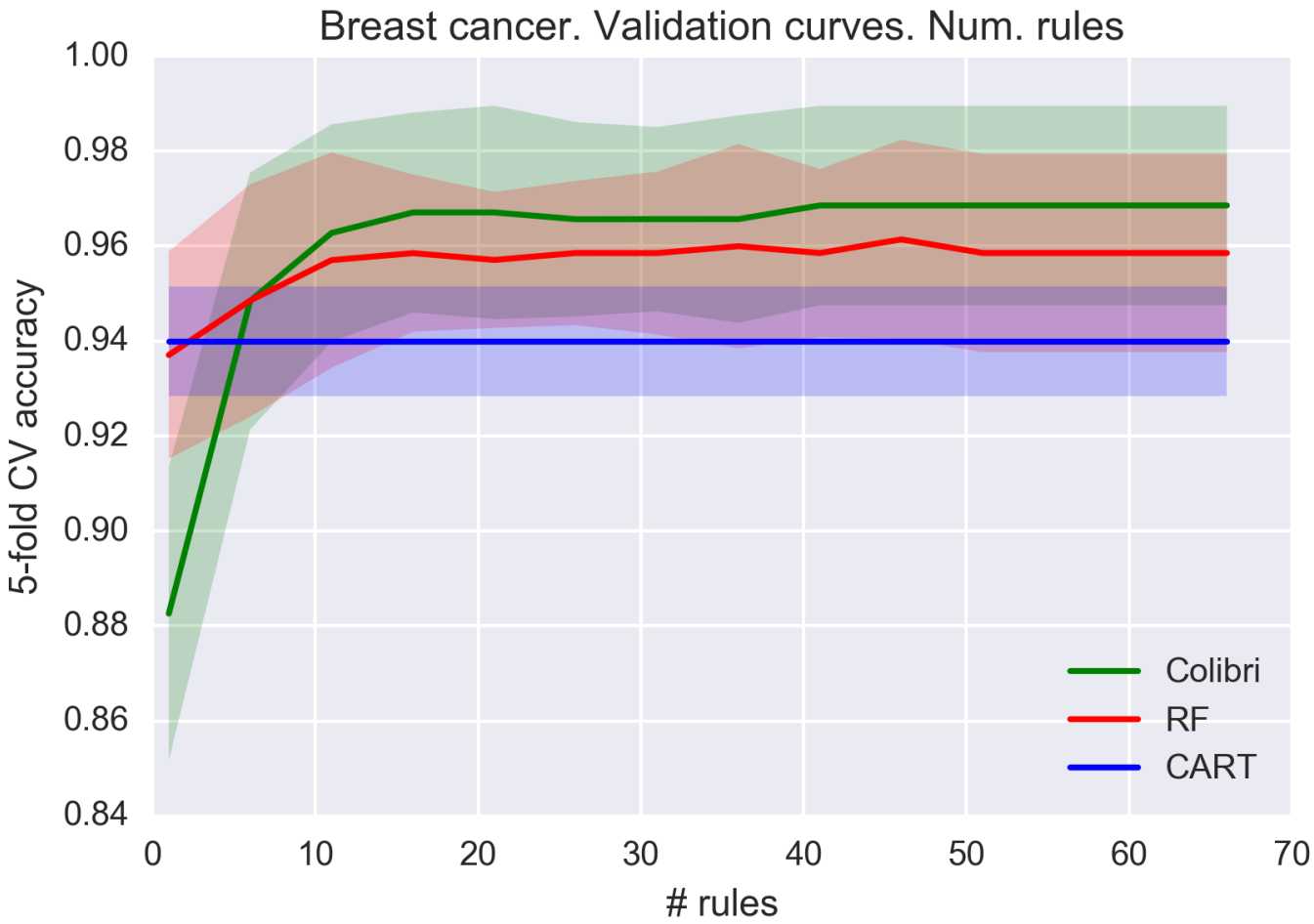
Набор Breast Cancer Wisconsin

Плохих правил больше



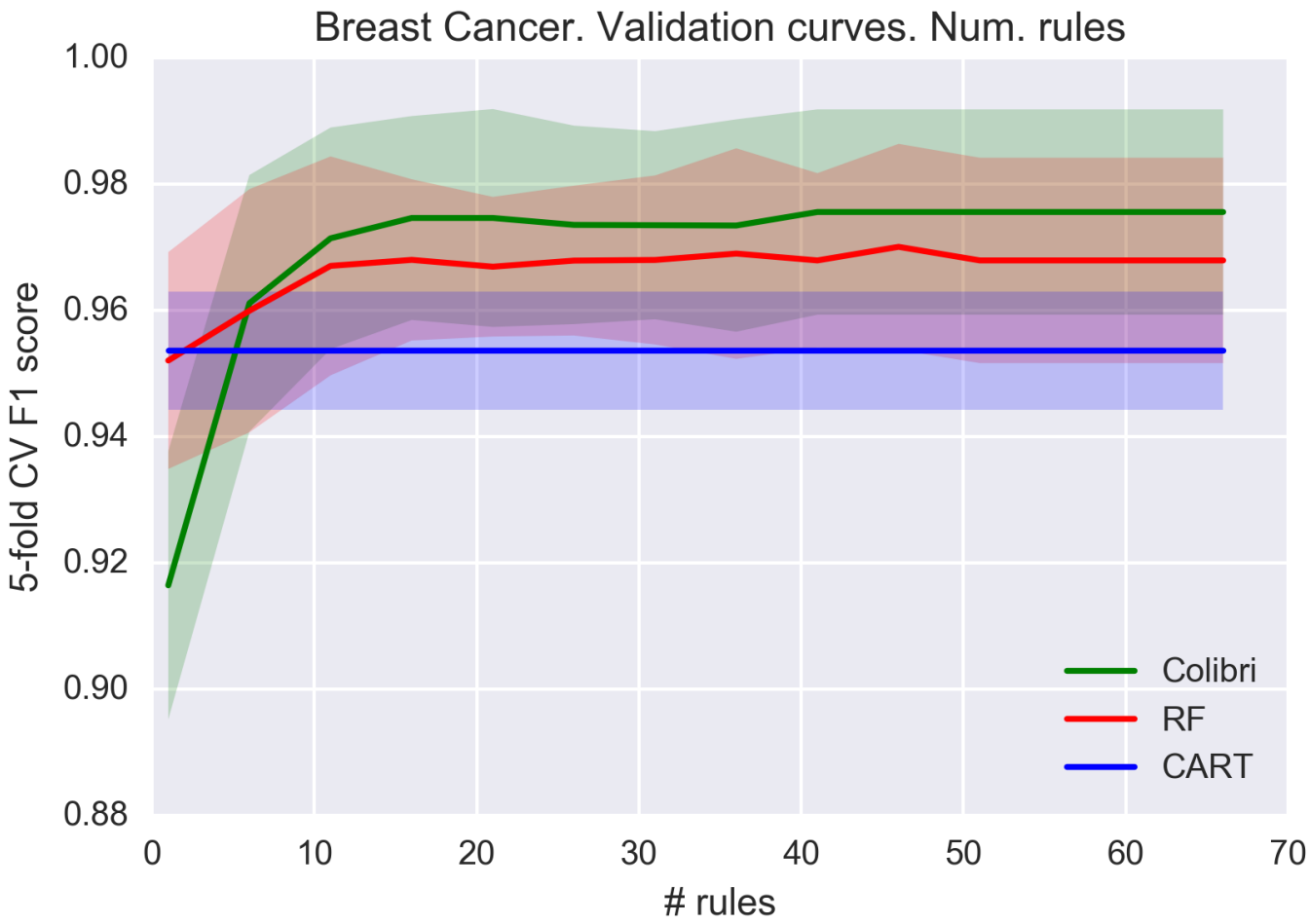
Кривые валидации. Доля верных ответов и число правил

Набор Breast Cancer Wisconsin



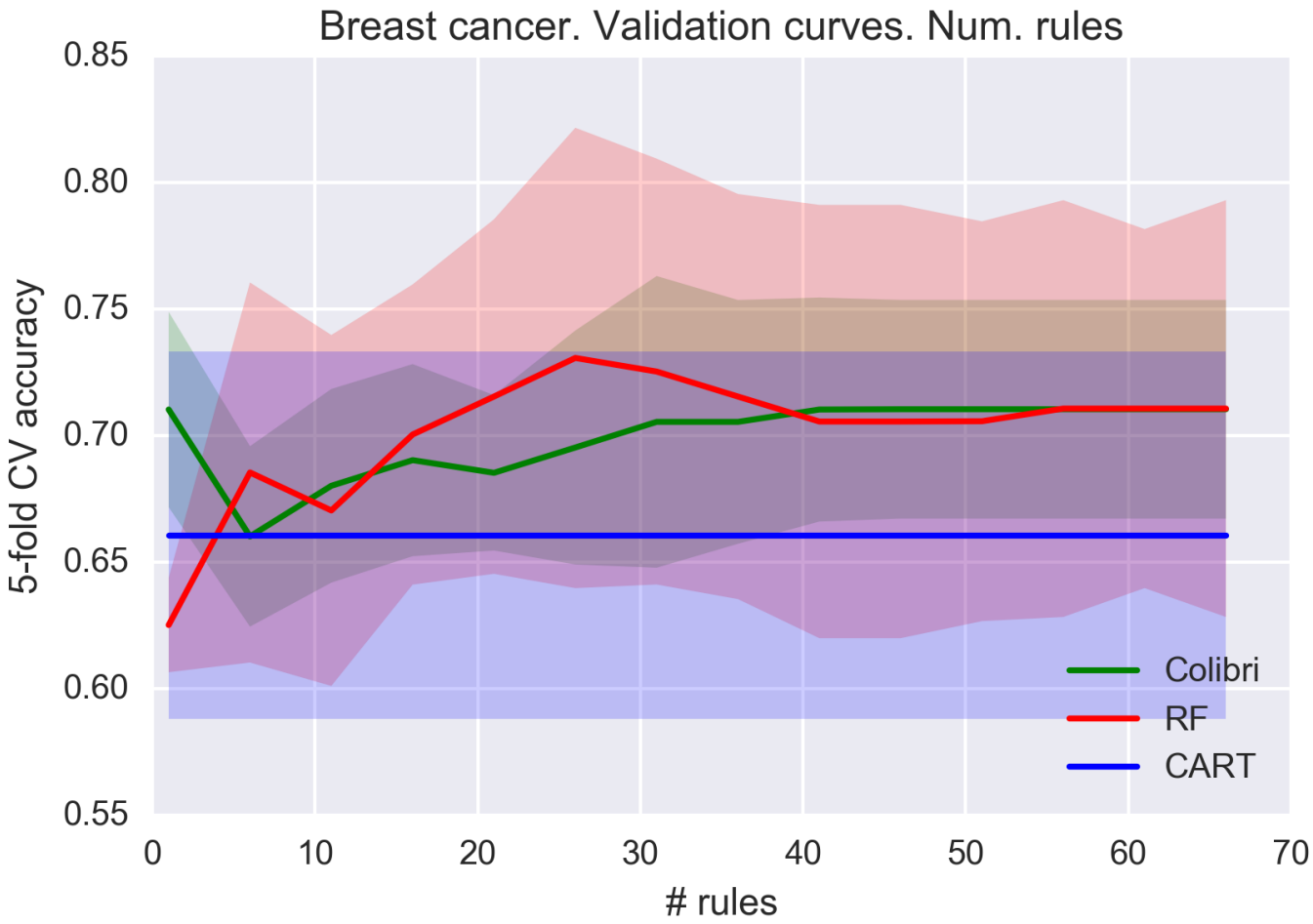
Кривые валидации. F1 и число правил

Набор Breast Cancer Wisconsin



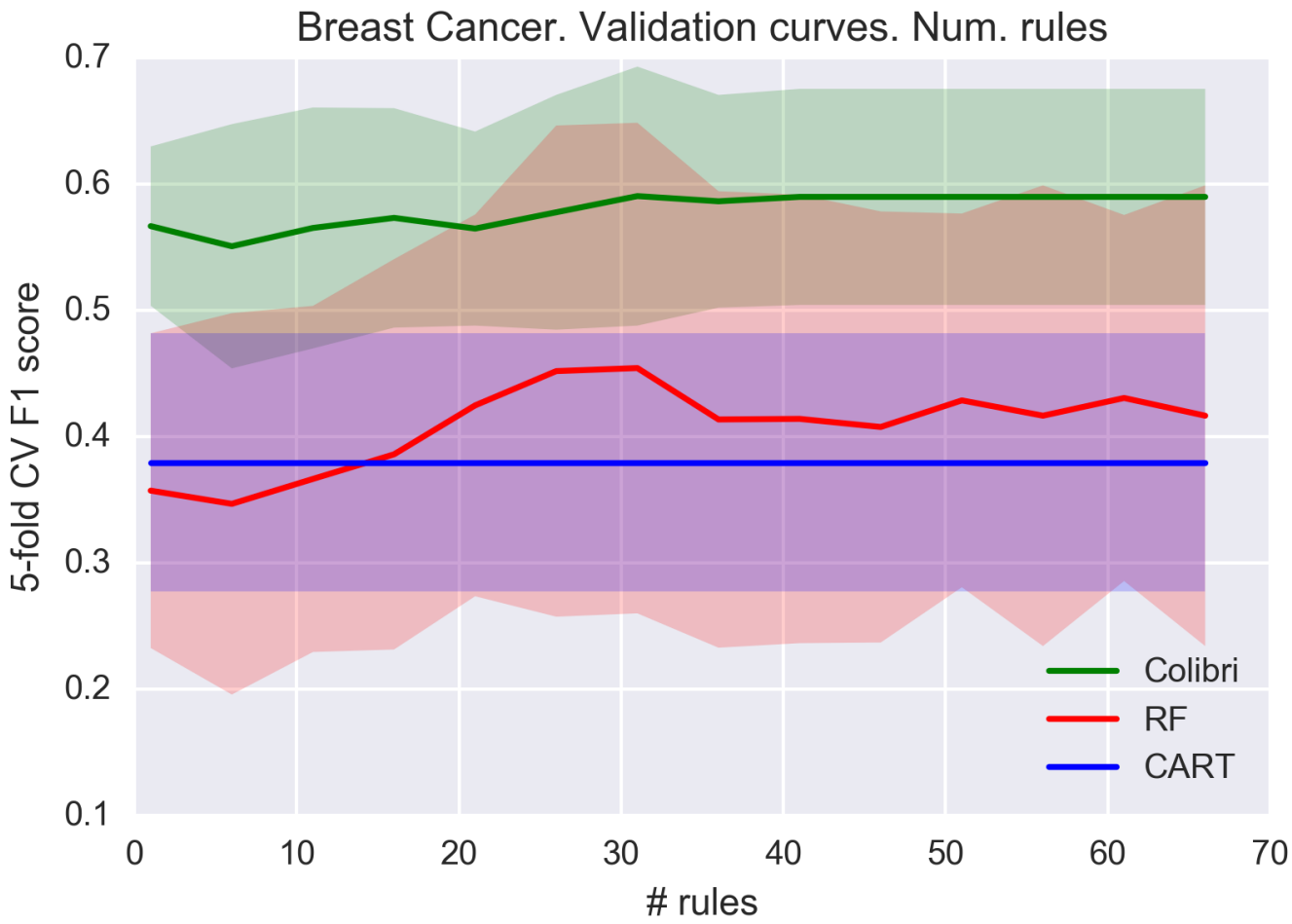
Кривые валидации. Доля верных ответов и число правил

Набор Breast Cancer



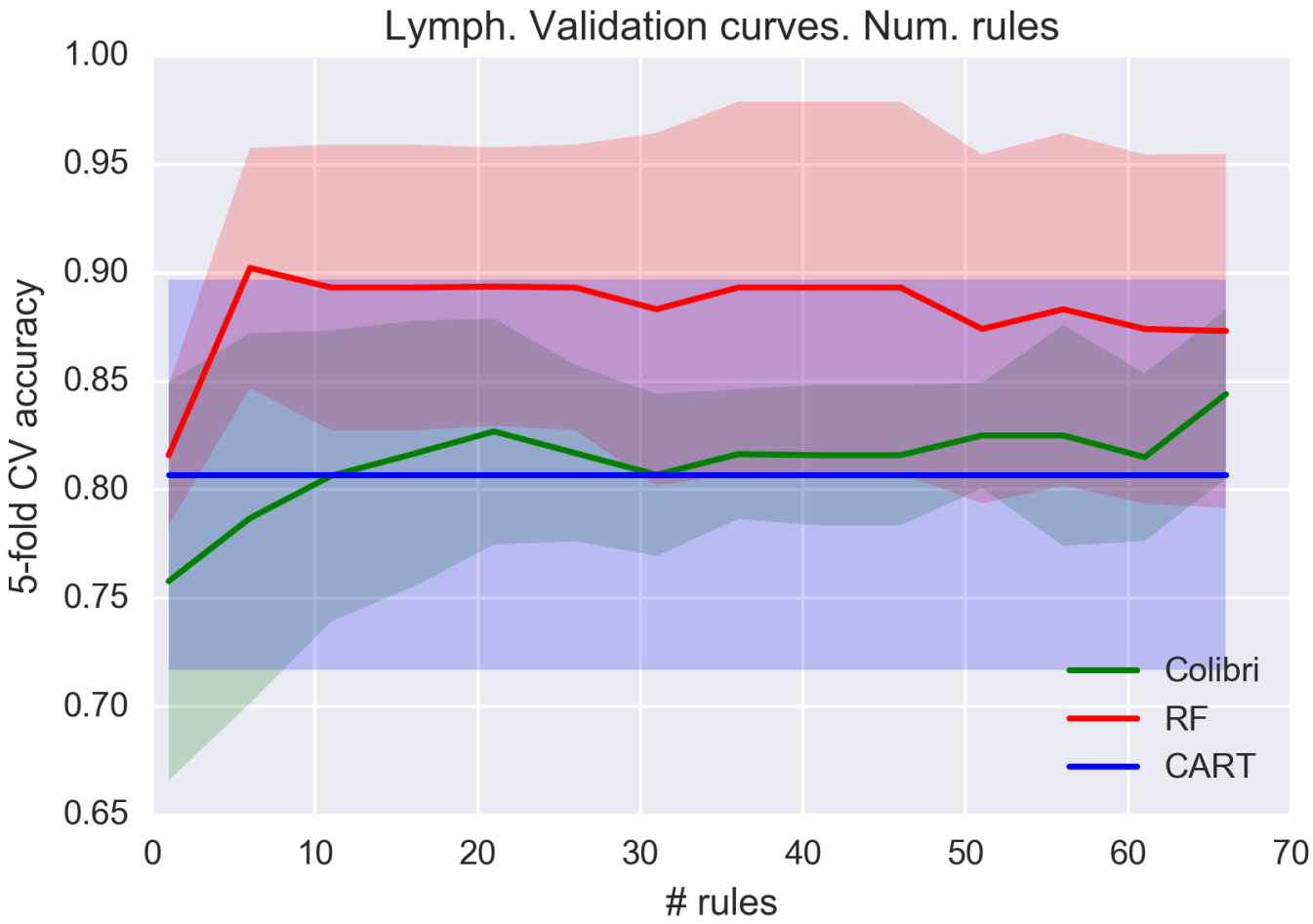
Кривые валидации. F1 и число правил

Набор Breast Cancer



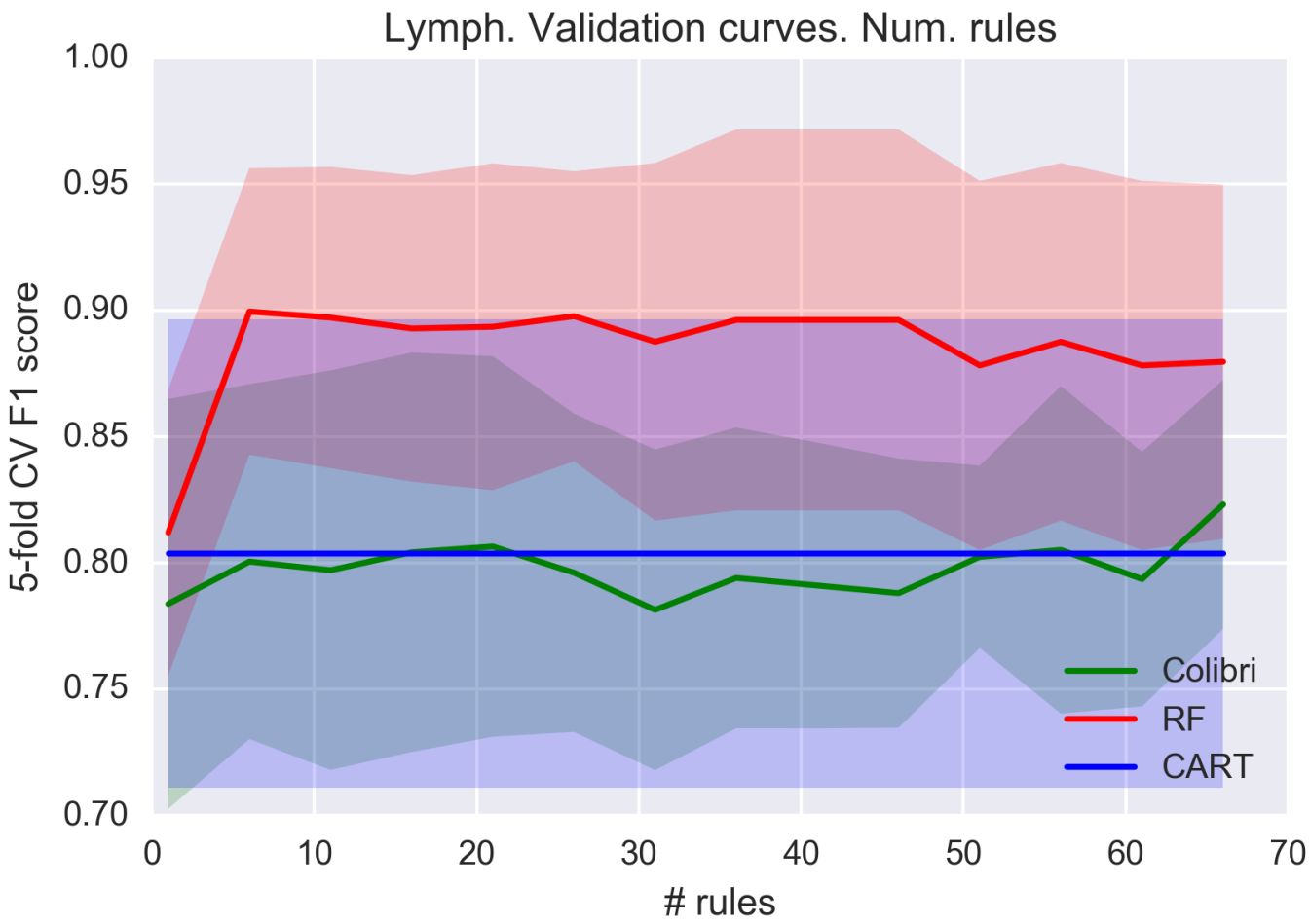
Кривые валидации. Доля верных ответов и число правил

Набор данных репозитория UCI Lymph



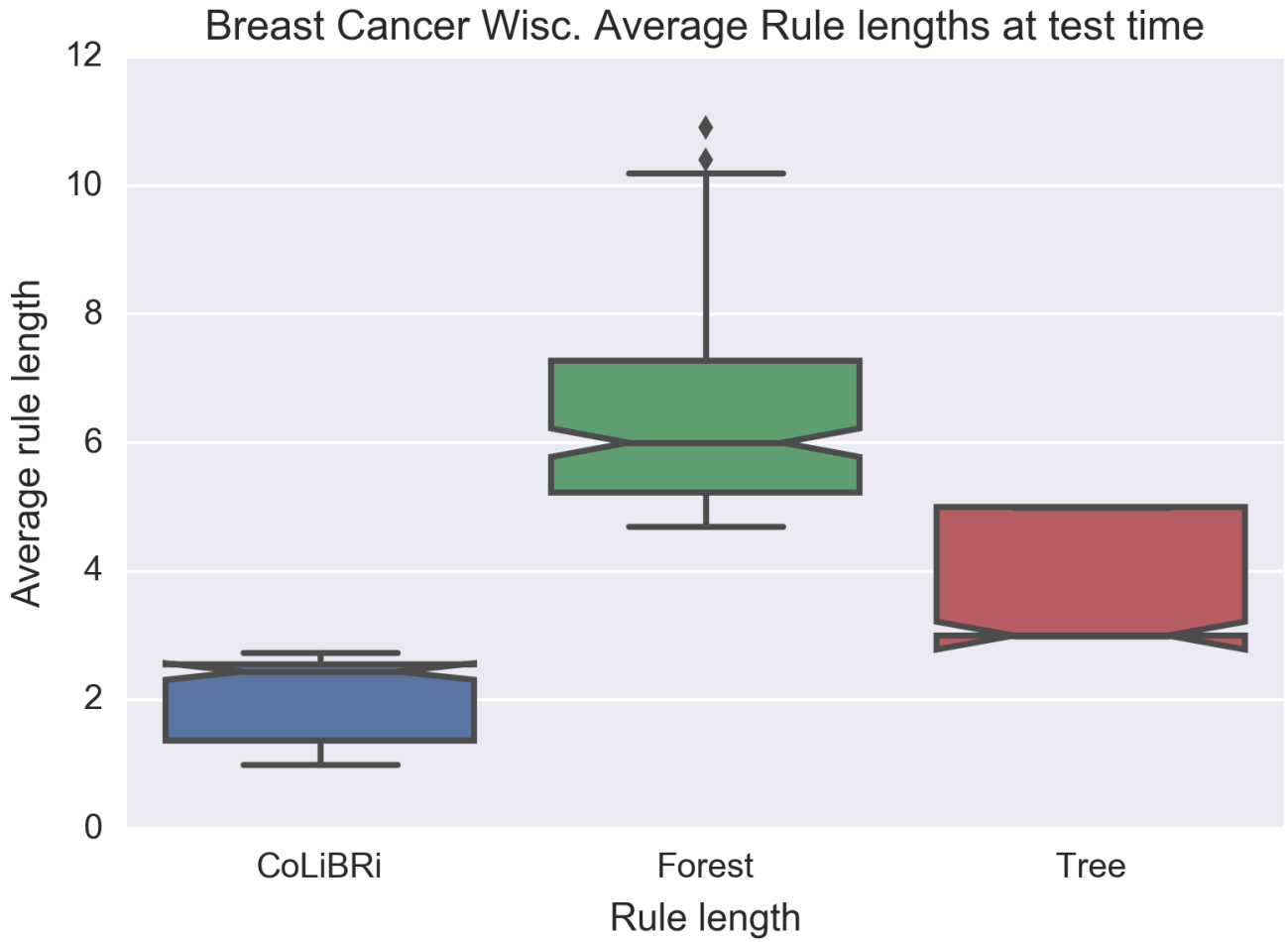
Кривые валидации. F1 и число правил

Набор данных репозитория UCI Lymph



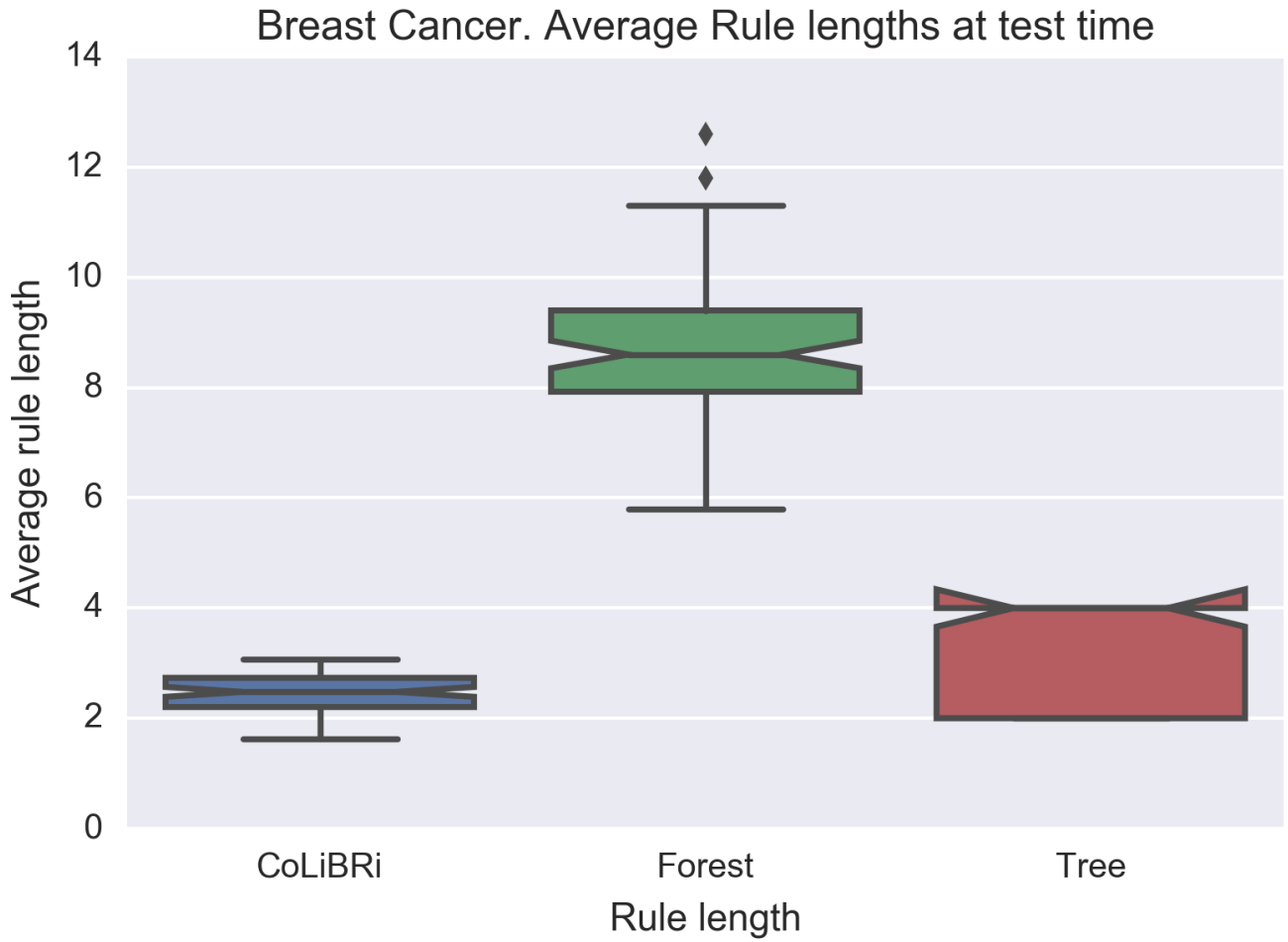
Длины посылок правил, участвовавших в классификации тестовых примеров

Набор Breast Cancer Wisconsin



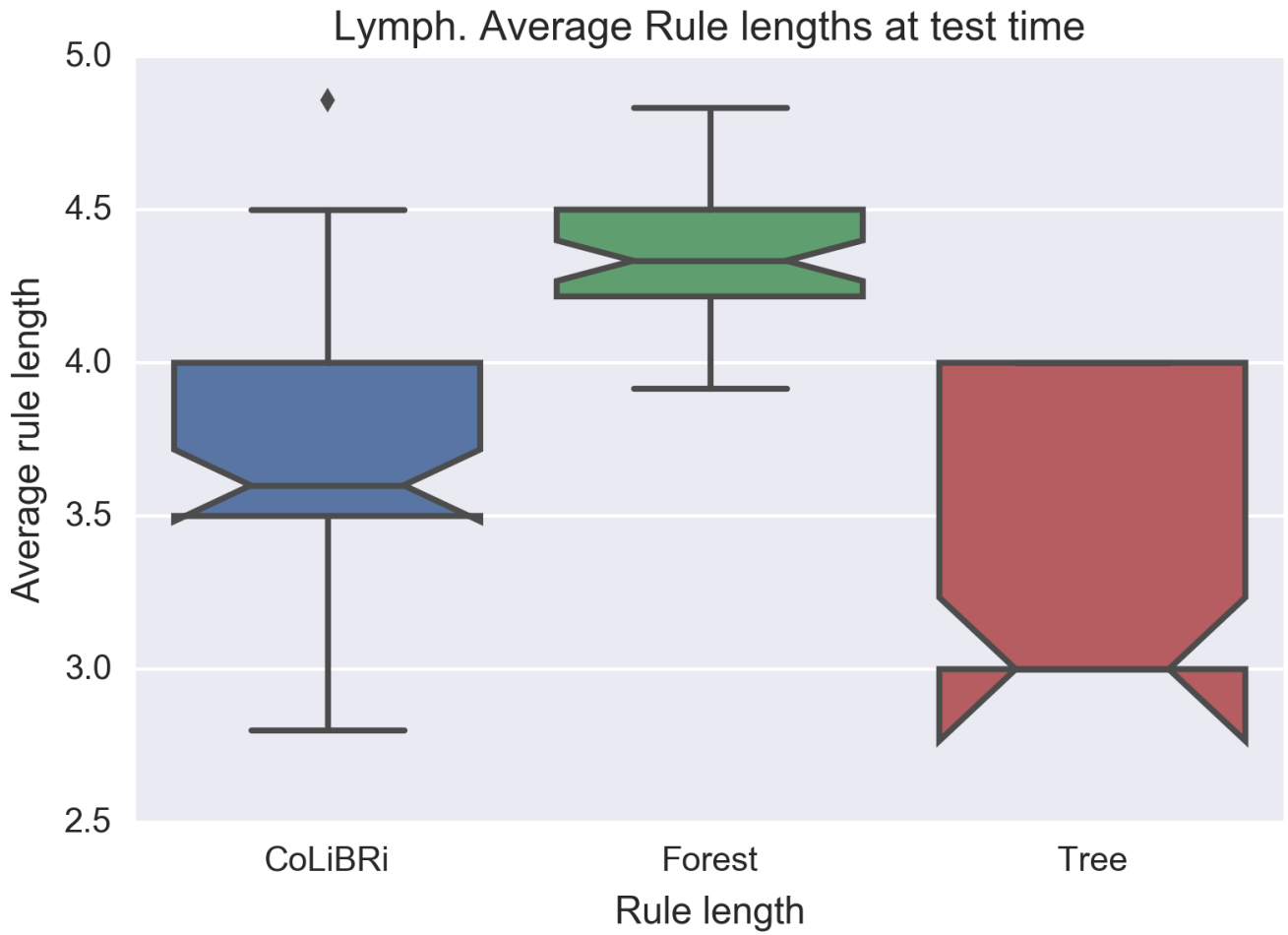
Длины посылок правил, участвовавших в классификации тестовых примеров

Набор Breast Cancer



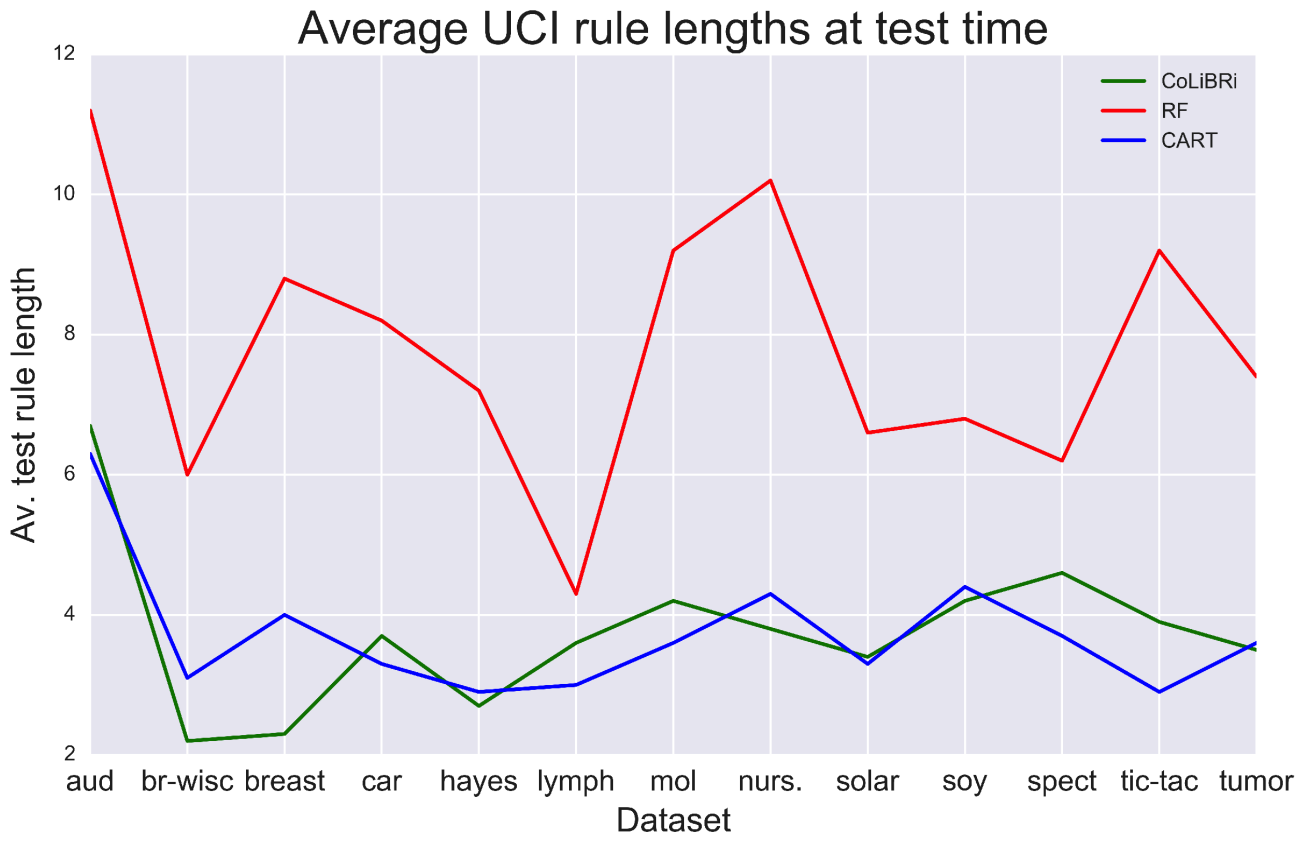
Длины посылок правил, участвовавших в классификации тестовых примеров

Набор Breast Cancer



Средние длины посылок правил

Средние мощности посылок правил, которыми были классифицированы тестовые объекты, для 3 алгоритмов и 13 наборов данных репозитория UCI.



Выводы про алгоритм CoLiBRi

- Качество классификации выше, чем у деревьев и ниже, чем у случайного леса
- Локальная интерпретируемость лучше, чем у случайного леса и хуже, чем у деревьев
- Может обобщаться на данные со сложной структурой
- Минус – в худшем случае экспоненциальная сложность $\rightarrow$ для малых данных

Спасибо! Вопросы?