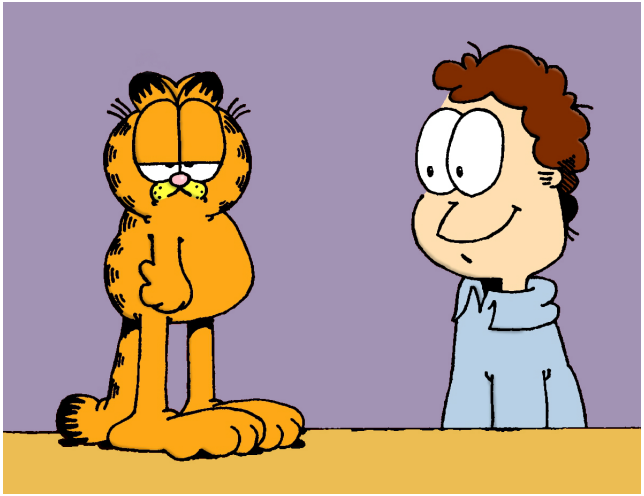
Машинное обучение

Н.С. Медянкин

НИУ ВШЭ, Москва

14 июля 2015, Дубна, Ратмино

Гарфильд (19 июля 1978 —) и Джон



Гарфильд ненавидит понедельники

...и обожает кофе по утрам

Дневник Джона: 19 июля 1978 — 13 июля 2015 (13539 дней)

- **day:** {*Sun, Mon, Tue, Wed, Thu, Fri, Sat*} — день недели;
- **weather:** {*raw, chilly, mild, hot, scorching*} — погода;
- **cups:** [0, 3] — сколько чашек кофе уже выпито;
- **more:** {*Yes, No*} — хочет ли Гарфильд ещё одну чашку.

Mon	mild	0	Yes
Tue	mild	0	Yes
Wed	raw	2	Yes
Thu	chilly	2	Yes
Fri	mild	3	No
Sat	raw	0	Yes
Sun	hot	1	No

Джон собирается в отпуск без Гарфильда

Как по заданному набору *day*, *weather* и *cups* понять, хочет ли Гарфильд ещё чашку?

Mon	scorching	1	?
Sun	raw	3	?
Tue	chilly	3	?
Wed	hot	0	?

В дело вступает машинное обучение

Немного терминологии

- Дневник Джона — **данные** (*data set*);
- *day*, *weather* и *cups* — **признаки** (*features*);
- *more* — **классы** (*classes*).

Что мы собираемся сделать?

На этих данных обучить машину предсказывать класс по набору значений признаков.

Хорошо ли машина научилась?

Обучающая и тестовая выборки

1. Отделить 10% данных — **тестовую выборку** (*test set*).
2. Обучить машину на оставшихся 90% — **обучающей выборке** (*training set*).
3. Предсказать классы для тестовой выборки.
4. Сверить результаты предсказаний с тем, что мы и так знаем про тестовую выборку, — получить **оценку качества**.

Хорошо ли машина научилась?

Скользящий контроль (*cross-validation*)

1. Разбить данные на 10 равных частей.
2. Обучиться десять раз, каждый раз проверяя качество:
   - Каждая часть по одному разу работает тестовой выборкой.
   - Все остальные данные при этом работают обучающей выборкой.
3. Посчитать среднее арифметическое оценки качества за 10 раз.

Поехали!

Алгоритм One Rule

Выбирает один наиболее важный признак.

weather

raw	$\rightarrow$	Yes
chilly	$\rightarrow$	Yes
mild	$\rightarrow$	Yes
hot	$\rightarrow$	No
scorching	$\rightarrow$	No

Качество работы One Rule

	Yes	No	Total
Yes	6112	746	6858
No	2012	4669	6681
Total	8124	5415	

Оценки качества для каждого класса

Точность (*precision*)

$$\text{Точность} = \frac{\text{Про сколько элементов класса предсказали правильно}}{\text{Про сколько элементов сказали, что они принадлежат этому классу}}$$

Полнота (*recall*)

$$\text{Полнота} = \frac{\text{Про сколько элементов класса предсказали правильно}}{\text{Сколько в действительности принадлежит этому классу}}$$

Качество работы One Rule

	Yes	No	Total
Yes	6112	746	6858
No	2012	4669	6681
Total	8124	5415	

Class	Precision	Recall
Yes	0.752	0.891
No	0.862	0.699

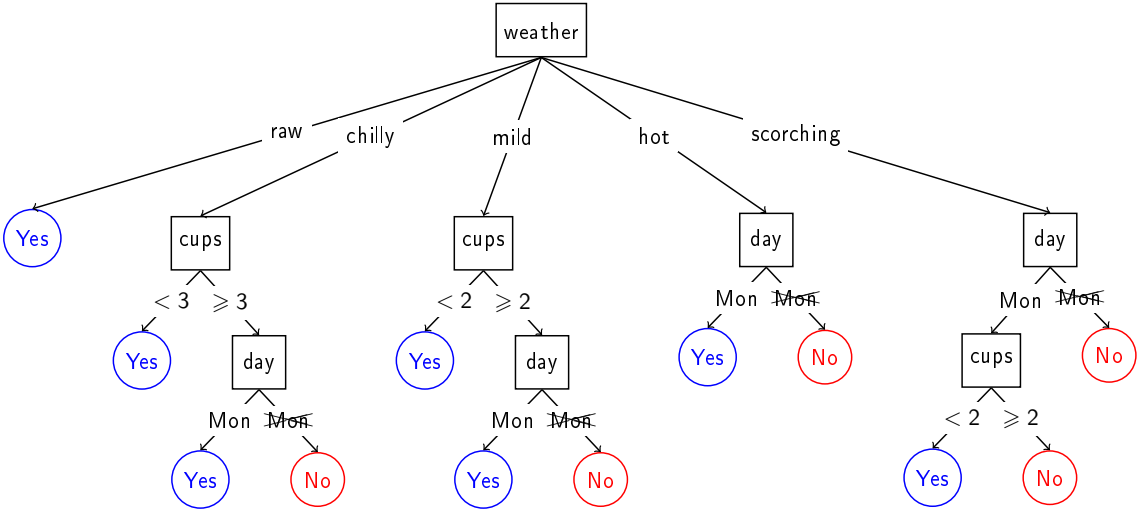
Оценка качества для всех классов

Общая точность (*accuracy*)

$$\text{Общая точность} = \frac{\text{Про сколько элементов всех классов предсказали правильно}}{\text{Общее количество элементов}}$$

Correctly Classified Instances	10781
Total Instances	13539
Accuracy	79.6%

Дерево решений



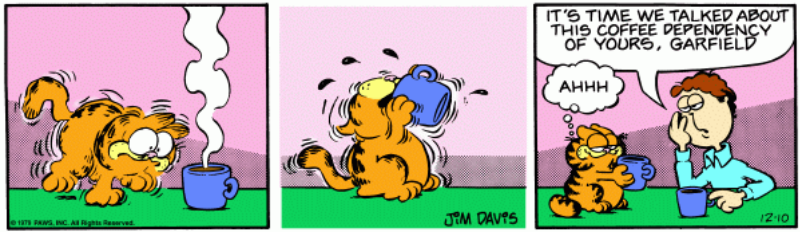
Качество работы дерева решений

Correctly Classified Instances	12898
Total Instances	13539
Accuracy	95.3%

Почему не 100%?

В целом Гарфильд ведёт себя довольно предсказуемо. Но примерно в 5% случаев он принимает решение исключительно по зову левой пятки. Ошибка предсказания заложена на этапе сбора данных.

Спасибо за внимание!



P.S. Что это было?

Частная задача

Машинное обучение с учителем: задача бинарной классификации.

С учителем (supervized)

Есть размеченные данные для обучения (с заранее проставленными ответами).

Классификация (classification)

Множество возможных ответов дискретно (делится на классы).

Бинарная (binary)

Классов два.