

Введение в анализ данных

17.01.2018

Задачи машинного обучения

Все машинное обучение можно условно разделить на три категории:

- 1 Обучение с учителем
- 2 Обучение без учителя
- 3 Обучение с подкреплением (не рассматриваем в данном курсе)

Дано

X – множество объектов;

Y – множество целевых значений;

$X^\ell = \{x_1, x_2, \dots, x_\ell\}$ – обучающая выборка;

$Y^\ell = \{y_1, y_2, \dots, y_\ell\}$ – целевые значения;

Предполагается, что между объектами и ответами имеется зависимость

Необходимо

Выработать решающее правило, позволяющее для объектов X спрогнозировать целевую переменную Y

Примеры

- Классификация
- Регрессия

Обучение без учителя

Дано

X – множество объектов;

Y – множество целевых значений;

$X^\ell = \{x_1, x_2, \dots, x_\ell\}$ – обучающая выборка;

Необходимо

выработать решающее правило, позволяющее для объектов X спрогнозировать целевую переменную Y

Примеры

- Кластеризация

Классификация

- Кредитный скоринг (выдавать или нет кредит).
 x_i – информация о человеке; y_i – вернул кредит или нет
- Медицинская диагностика.
По набору медицинских характеристик требуется поставить диагноз
- Борьба со спамом.
Классификация текста на два класса

Регрессия

- Предсказание стоимости страховки.
 x_i – информация о клиенте и о страхуемом объекте; y_i – стоимость страховки
- Предсказание продаж

Кластеризация

- По результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества.

Впечатляющие примеры ML

- cs.stanford.edu/people/karpathy/deepimagesent/generationdemo
Примеры сгенерированных предложений по изображению
- <https://www.imageidentify.com>
Классификация загруженного изображения
- <https://cs.stanford.edu/people/jcjohns/fake-math/4.pdf>
Документ .tex сгенерированный нейронной сетью (подробнее тут <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>)