Как научиться решать соревнования по анализу данных

ЛКШ.2017.Август

Липкина Анна

7 августа 2017 г.

Машинное обучение

Определение

Машинное обучение — это наука, изучающая способы извлечения закономерностей из ограниченного количества примеров

Пример задачи

Задача

Предсказать, возьмут ли школьника в ЛКШ

Формат предсказания

- Бинарный ответ: 1 — возьмут, 0 — не возьмут
- Вещественный ответ $p \in [0, 1]$ — вероятность того, что возьмут

Данные

Откуда брать данные?

Данные

- У нас есть данные по достаточно большому количеству зачислений/незачислений школьников за предыдущие года

Обучающая выборка

Каждый пример, на котором известен ответ, называется **обучающим**, а совокупность всех таких примеров — **обучающей выборкой**. Обозначение: $\mathbb{X}$

Ответы

Величина, которую мы хотим предсказывать, называется **целевой переменной**. Обозначение: $\mathbb{Y}$

Как описывать школьника

Объекты

В данном случае, каждый отдельный школьник называется **объектом**

Объекты

Как хранить объект в компьютере?

Признаковое описание объекта

- Объект — некоторая абстрактная сущность, с которым компьютер не умеет работать напрямую

Признаки

Будем описывать каждый объект некоторым набором характеристик, называемых **признаками**

Объекты

Какие признаки можно выделить для объектов из нашей задачи?

Матрица объекты-признаки

Город	Пол	Класс	Был ли в ЛКШ	Параллель
Сосновый Бор	М	7	1	С
Москва	М	8	1	С’
Москва	Ж	8	1	С’
СПб	М	8	1	С
Волгоград	Ж	8	1	С’
Железногорск	Ж	6	0	
Москва	М	10	1	А’

Типы признаков

- **Бинарные** — принимают 2 значения
- **Вещественные** — принимают значения из $\mathbb{R}$
- **Категориальные** — принимают значения из неупорядоченного множества

Какие типы у признаков, выделенных в этой задаче?

А что дальше?

- Проходит год, объявляется новый набор в ЛКШ
- Появляются новые данные
- Для каждого школьника хотим предсказать, возьмут ли его в ЛКШ

Тестовая выборка

Каждый объект, для которого ответ неизвестен (и мы хотим предсказать его), называется **тестовым объектом**. Совокупность тестовых объектов называется **тестовой выборкой**

Построение модели

- Чтобы предсказывать ответы на тестовой выборке, необходимо иметь некоторый алгоритм a этого предсказания.

Модель

Функция $a : \mathbb{X} \to \mathbb{Y}$, которая будет предсказывать ответ на тестовой выборке, называется **алгоритмом (моделью).**

Выбор модели

Как выбирать модель?

Выбор модели

Как выбирать модель?

Функционал качества

Чтобы модель соответствовала нашим ожиданиям и корректно отображала зависимости между данными и ответами, необходимо ввести **функционал качества (ошибки)**, измеряющий качество работы алгоритма. Функция, измеряющая ошибку одного предсказания, называется **функцией потерь**. Как правило, функционалы ошибки являются суммой функций потерь на всех объектах тестовой выборки

Выбор модели

Как выбирать модель?

Функционал качества

Чтобы модель соответствовала нашим ожиданиям и корректно отображала зависимости между данными и ответами, необходимо ввести **функционал качества (ошибки)**, измеряющий качество работы алгоритма. Функция, измеряющая ошибку одного предсказания, называется **функцией потерь**. Как правило, функционалы ошибки являются суммой функций потерь на всех объектах тестовой выборки

Какие функции потерь можно использовать в данной задаче?

Выбор алгоритма

- Для выбора алгоритма фиксируется некоторое параметрическое семейство алгоритмов $\mathcal{A}$
- Из этого семейства выбирается алгоритм, наилучший с точки зрения выбранного функционала

Линейная регрессия

- Пусть в обучающей выборке n объектов и d признаков
- $$\mathcal{A} = \{a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d | w_i \in \mathbb{R} \quad \forall 0 \leqslant i \leqslant d\}$$
- x_i — значение i-го признака у объекта x
- $\frac{1}{d}(w_0 + \sum_{i=1}^{d} w_j x_{ij} - y_i)^2 \to \min_{w_k}$

Обучение

Обучение

Обучение — процесс поиска оптимального алгоритма из заданного семейства относительно функционала ошибки

И это всё?

- Перед обучением очень часто возникает потребность в **предобработки данных**

Предобработка данных

- Удаление неинформативных признаков
- Удаление выбросов
- Нормализация данных

Переобучение

Переобучение

Переобучение — процесс, при котором алгоритм слишком сильно подгоняется под обучающую выборку, но не выявляется никаких закономерностей в ней

Борьба с переобучением

- Контроль сложности семейства алгоритмов — чем меньше данных для обучения, тем более простые семейства следует выбирать
- Контроль качества обучения. Кросс-валидация

Кросс-валидация

- Перемешиваем данные
- Разбиваем обучающую выборку на n примерно равных частей
- n раз обучаем алгоритм на подвыборке, состоящей из $n - 1$ кусков.
- Считаем качество обучения на каждой из подвыборок, усредняем их
- Итоговый результат показывает, насколько хорошо алгоритм приближает истинную зависимость

Кросс-валидация

- Кросс-валидацию можно использовать и не только для контроля переобучения
- С ее помощью подбираются оптимальные для данного семейства алгоритмов параметры

Итоги

- Постановка задачи
- Выделение признаков
- Формирование выборки
- Выбор метрики качества
- Предобработка данных
- Построение модели
- Кросс-валидирование, оценивание качества модели