использование для выработки прогнозов

Основные проблемы машинного обучения

Данные плохого качества

Ниже перечислены возможные меры:

- Если некоторые примеры являются несомненными выбросами, то может помочь простое их отбрасывание или попытка вручную исправить ошибки.
- Если в некоторых примерах отсутствуют какие-то признаки (скажем, 5% ваших заказчиков не указали свой возраст), тогда вам потребуется решить, что делать: игнорировать такие атрибуты; игнорировать примеры с недостающими атрибутами; заполнить отсутствующие значения (возможно, средним возрастом); обучать одну модель с признаками и одну модель без признаков; и т.д.

Такой процесс, называемый конструированием признаков (feature engineering), включает в себя:

- выбор признаков (feature selection) - выбор самых полезных признаков
- выделение признаков (feature extraction) - сочетание существующих признаков для выпуска более полезного признака (как было показано ранее, помочь может алгоритм понижения размерности);
- создание новых признаков путем сбора новых данных.

Переобучение (overfitting)

Переобучение происходит, когда модель слишком сложна относительно объема и зашумленности обучающих данных.

Ниже перечислены возможные решения:

- Упростить модель, выбрав вариант с меньшим числом параметров (например, линейную модель вместо полиномиальной модели высокого порядка), сократив количество атрибутов в обучающих данных или ограничив модель
- Накопить больше обучающих данных.
- Понизить шум в обучающих данных (например, исправить ошибки данных и устранить выбросы).

Ограничение модели с целью ее упрощения и снижения риска переобучения называется **регуляризацией (regularization)**.

Объемом регуляризации, подлежащим применению во время обучения, можно управлять с помощью **гиперпараметра.**

Гиперпараметр - это параметр обучающего алгоритма (не модели). По существу сам обучающий алгоритм на него не влияет; гиперпараметр должен быть установлен перед обучением, а во время обучения оставаться постоянным. Если вы установите гиперпараметр регуляризации в очень большую величину, то получите почти плоскую модель (с наклоном, близким к нулю); обучающий алгоритм почти наверняка не допустит переобучения обучающих данных, но с меньшей вероятностью найдет хорошее решение. Регулировка гиперпараметров - важная часть построения системы МО.

Недообучение (underfitting)