Словарь начинающего аналитика

Полезен тем, что:

- Содержит расшифровку основных терминов, которые будут встречаться в курсе
- Всегда будет под рукой, и к нему можно обратиться в любой момент при обучении/работе

Для удобства термины расположены по порядку, исходя из последовательности лекций

Лекция 8

Библиотека Scikit-learn (Sklearn) — это библиотека машинного обучения на языке программирования Python с открытым исходным кодом. Содержит реализации практически всех возможных преобразований, и нередко её одной хватает для полной реализации модели.

Выборка для обучения — это выборка для «обучения» той или иной модели, т. е. для построения математических отношений между некоторой переменной-откликом и предикторами (средством прогнозирования).

Выборка для тестирования — это выборка для получения оценки прогнозных свойств модели на новых данных, т. е. данных, которые не были использованы для обучения модели.

Качество модели — это соответствие модели определённым требованиям с точки зрения того, для чего она предназначена.

Классификация — это один из разделов машинного обучения, посвящённый решению следующей задачи. Имеется множество объектов (ситуаций), разделённых, некоторым образом, на классы. Задано конечное множество объектов, о которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Кластеризация — это задача разбиения заданной выборки данных (объектов) так, чтобы каждый кластер (объединение нескольких однородных элементов) состоял из схожих объектов, а объекты разных кластеров значительно отличались друг от друга. Задача кластеризации: используя все имеющиеся данные, предсказать соответствие объектов выборки их классам, сформировав таким образом кластеры. Кластеризацию применяют для анализа и поиска признаков, по которым можно объединить объекты сжатия данных и поиска новизны, что не входит ни в один кластер.

Машинное обучение (machine learning, ML) — это методики анализа данных, которые позволяют аналитической системе обучаться в ходе решения множества сходных задач. Машинное обучение базируется на идее о том, что аналитические системы могут учиться выявлять закономерности и принимать решения с минимальным участием человека.

Нормализация данных — это метод предварительной обработки данных для изменения их масштаба в каждой строке данных. Полезно использовать в наборе разреженных данных, где содержится много нулей.

Объект — это сущность в цифровом пространстве, обладающая определённым состоянием и поведением, имеющая определённые свойства (атрибуты) и операции над ними (методы). Каждый объект характеризуется набором признаков. Например, сообщение электронной почты.

Переобучение — это явление, когда построенная модель хорошо объясняет примеры из обучающей выборки, но относительно плохо работает на примерах, не участвовавших в обучении (на примерах из тестовой выборки).

Признаки объекта — это индивидуальное измеримое свойство или характеристика наблюдаемого явления, объекта. Например, набор слов, длина сообщения, дата, отправитель, получатель, язык.

Регрессия — это метод, который был принят в машинном обучении, когда по заданному набору признаков необходимо спрогнозировать некую целевую переменную. Задача регрессии — предсказать место на числовой прямой. Например, загруженность дорог в зависимости от времени суток и время на путь из пункта А в пункт Б в зависимости от пробок.

Средняя квадратичная ошибка (MSE) — это распространённая функция потерь, которая используется в линейной регрессии в качестве показателя эффективности. Чтобы рассчитать MSE, надо взять разницу между предсказанными значениями и истинными, возвести её в квадрат и усреднить по всему набору данных.

Целевая переменная — это переменная, которая описывает результат (цель) процесса. Например, 0 — нет осложнений, 1 — есть осложнения.