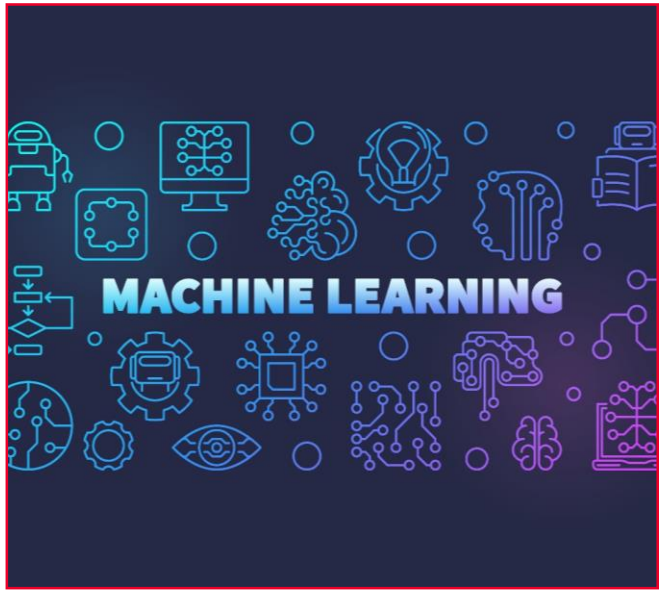
PROАГРО ЛЕКТОРИЙ

Анализ больших данных методами машинного обучения

Демичев Вадим Владимирович,
кандидат экономических наук,
доцент кафедры статистики и кибернетики РГАУ-МСХА имени К. А. Тимирязева

Вопросы лекции

- Введение в машинное обучение
- Преимущества использования машинного обучения
- Важнейшие библиотеки Python для машинного обучения
- Процесс построения модели.



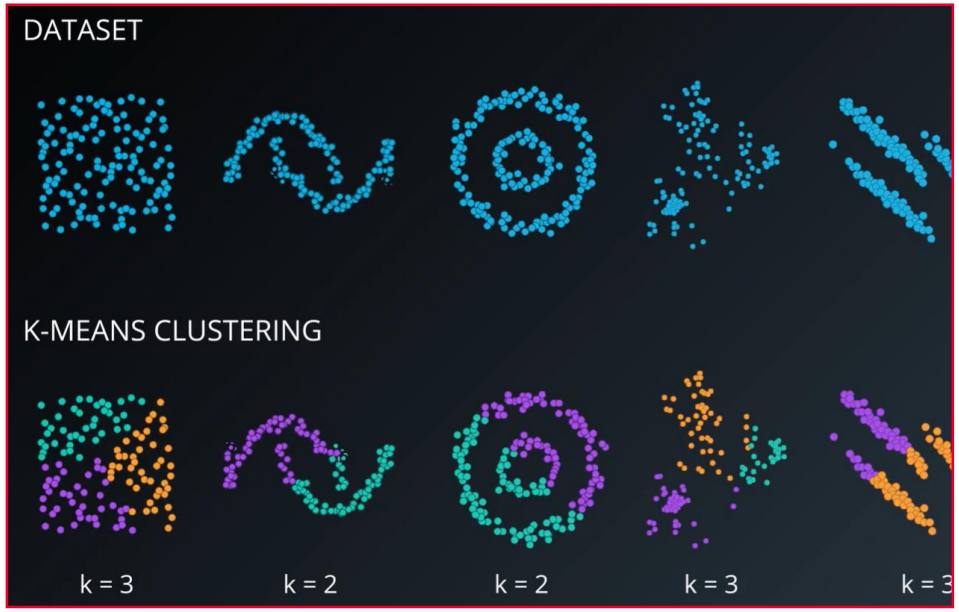
Введение в машинное обучение

- **Машинное обучение** - область исследований, которая наделяет компьютеры возможностью проявлять поведение, которое не было заложено в них явно (Артур Сэмюэл).

- **Машинное обучение** - процесс, в котором точность работы компьютера повышается по мере сбора данных и извлечения из них информации (Майк Робертс).

- **Машинное обучение** - обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Введение в машинное обучение



В машинном обучении применяется три ключевых метода:

- Регрессия
- Классификация
- Кластеризация

Введение в машинное обучение

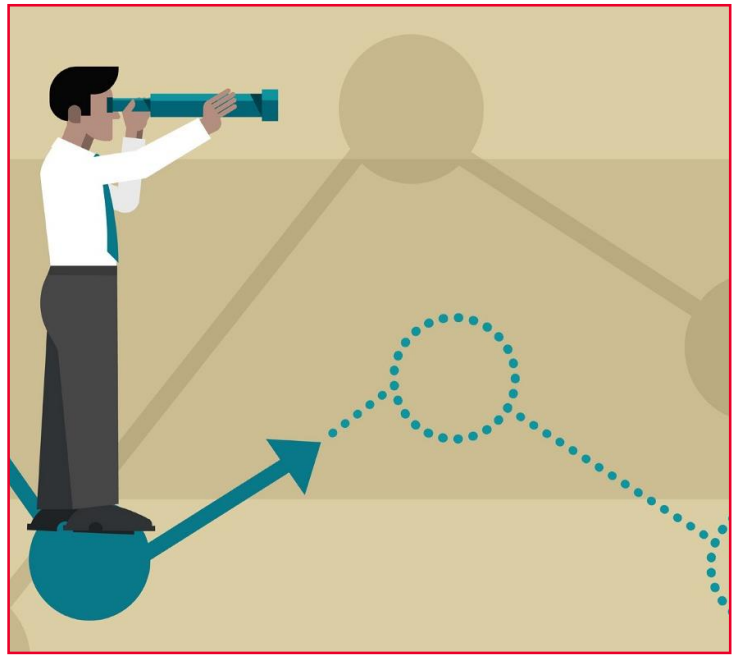
Ключевые понятия машинного обучения:

- Метка (label) – это то, что мы предсказываем, переменная Y в простой линейной регрессии (так же называют целевой переменной).
- Признак (feature) - это входная переменная - переменная X в простой линейной регрессии (так же называют независимыми переменными).
- В машинном обучении выделяют также понятие размеченный набор данных и неразмеченный набор данных. Размеченный набор данных используется для тренировки моделей машинного обучения.

Введение в машинное обучение

Выделяют два этапа жизни модели:

- Обучение - означает создание или изучение модели. То есть вы показываете модели помеченные данные и даете возможность постепенно изучать взаимосвязи между признаками и меткой.
- Прогнозирование - означает применение обученной модели к немаркированным данным. То есть вы используете обученную модель для составления полезных прогнозов (Y). Например, во время прогнозирования вы можете предсказать значение стоимости квартиры для новых немаркированных примеров.



Преимущества использования машинного обучения

Практические примеры использования моделей машинного обучения:

- Анализ состояния растительности (идентификация сорняков, классификация растительности, выявление болезни растений, прогнозирование урожайности).
- Анализ почвы (оценка состояния почвы, классификация почвы, прогнозирование плодородности почвы).
- Поиск нефтяных месторождений, золотых рудников и мест археологических раскопок на основании информации о существующих местах (классификация и регрессия).
- Поиск имен людей или названий географических мест в тексте (классификация).
- Идентификация людей по фотографиям или по записи голоса (классификация).
- Идентификация выгодных клиентов (регрессия и классификация).
- Активное выявление частей автомобиля, в которых с большей вероятностью произойдет поломка (регрессия).
- Прогнозирование суммы, которую человек потратит на продукт X (регрессия).
- Прогнозирование годового дохода компании (регрессия).

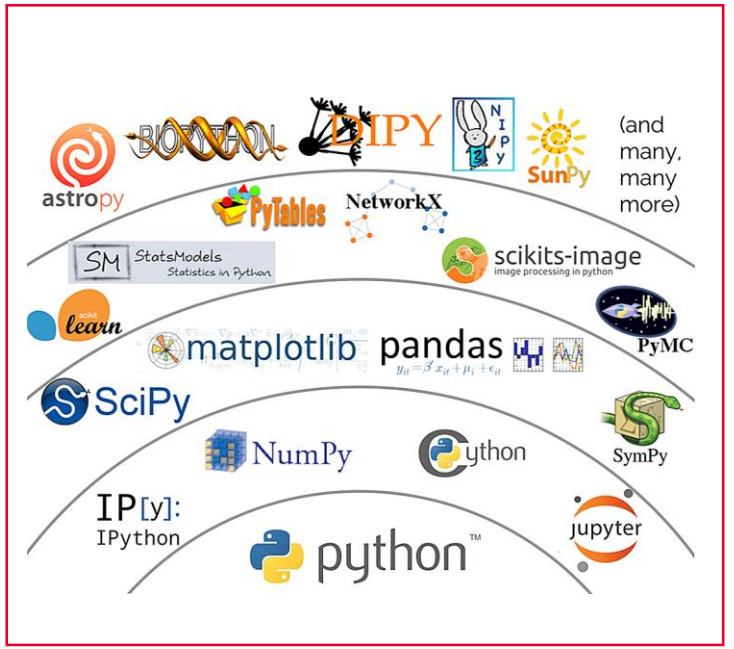
Важнейшие библиотеки Python для машинного обучения

Перечень библиотек Python для машинного обучения:

- Библиотеки для помещения данных в память: NumPy, Matplotlib, Pandas, StatModels, Scikit-learn, NLTK и другие

- Библиотеки оптимизации операций: Numba, PyCUBA, Blaze, Cython и другие

- Библиотеки для подключения инструментов машинного обучения к хранилищам данных: PyDoop, PySpark, Hadoopy и другие.



Процесс построения модели

Моделирование в машинном обучении состоит из 4-х шагов:

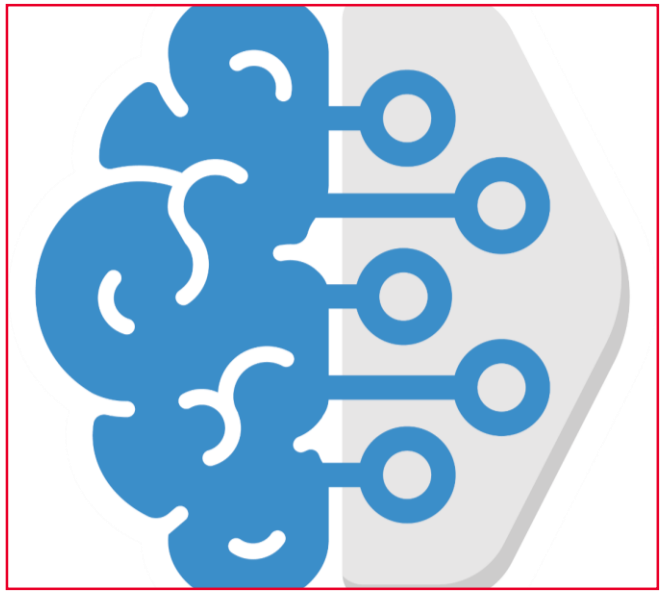
- Планирование показателей и выбор модели;
- Тренировка модели;
- Проверка адекватности и выбор модели;
- Применение тренированной модели к незнакомым данным.

Процесс построения модели

Планирование показателей - определение и идентификация возможных переменных модели. Данная задача является одной из важных, поскольку именно эти переменные используются для построения прогнозов.

Процесс построения модели

Тренировка модели. При наличии правильных свободных переменных и запланированного метода моделирования приступают к тренировке модели. В этой фазе модели передают данные, на которых она могла бы учиться.



Процесс построения модели

Проверка адекватности модели. Хорошая модель машинного обучения обладает двумя свойствами:

- она должна хорошо прогнозировать значения
- эффективно работать на ранее не известных для нее данных.

Для оценки этих свойств модели определяется метрика погрешности и стратегия проверки адекватности.

Процесс построения модели

Метрики погрешности в машинном обучении:

- частота ошибок классификации (для задач классификации);
- среднеквадратичная погрешность (для регрессионных задач).

Процесс построения модели

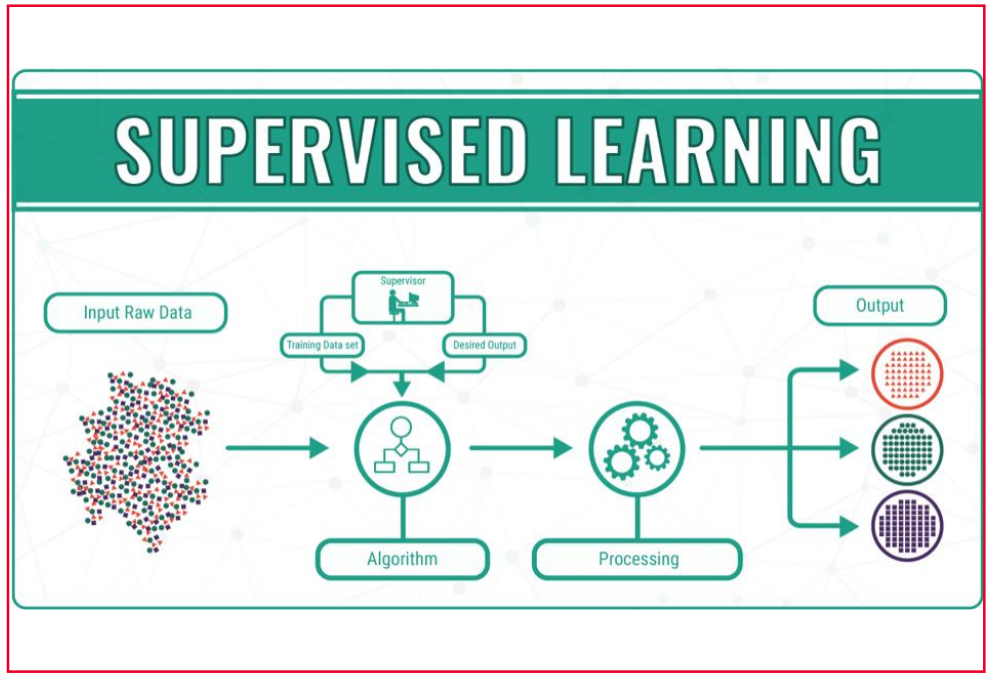
Стратегии проверки адекватности:

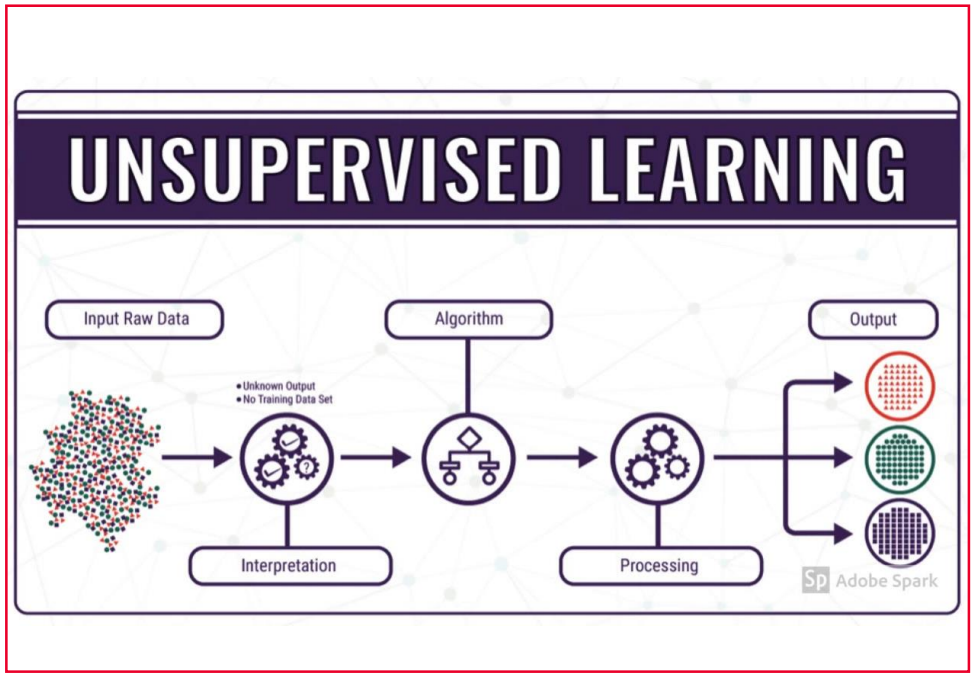
- разбиение данных на тренировочный набор с X% наблюдений и контрольную выборку с остальными данными.
- к – кратная перекрестная проверка – набор данных делится на k частей. Каждая часть однократно используется как тестовый набор данных, в то время как остальные части формируют тренировочный набор данных.
- исключение единицы – этот метод идентичен к-кратной проверке, но с к = 1. Одно наблюдение всегда исключается, а для тренировки используются остальные данные.
- регуляризация - подразумевает внесение штрафа за каждую дополнительную переменную, используемую для построения модели.
- L1 – регуляризация - строится модель с минимально возможным количеством независимых переменных.
- L2 – регуляризация направлена на минимизацию расхождений между коэффициентами независимых переменных.

Процесс построения модели

Прогнозирование новых наблюдений – завершающий этап процесса моделирования, подразумевающий применение модели на незнакомых данных.

Процесс построения модели





По способу использования помеченных данных различают методы контролируемого обучения, неконтролируемого обучения и методы с частичным контролем.

PROАГРО
ЛЕКТОРИЙ

Спасибо за внимание!