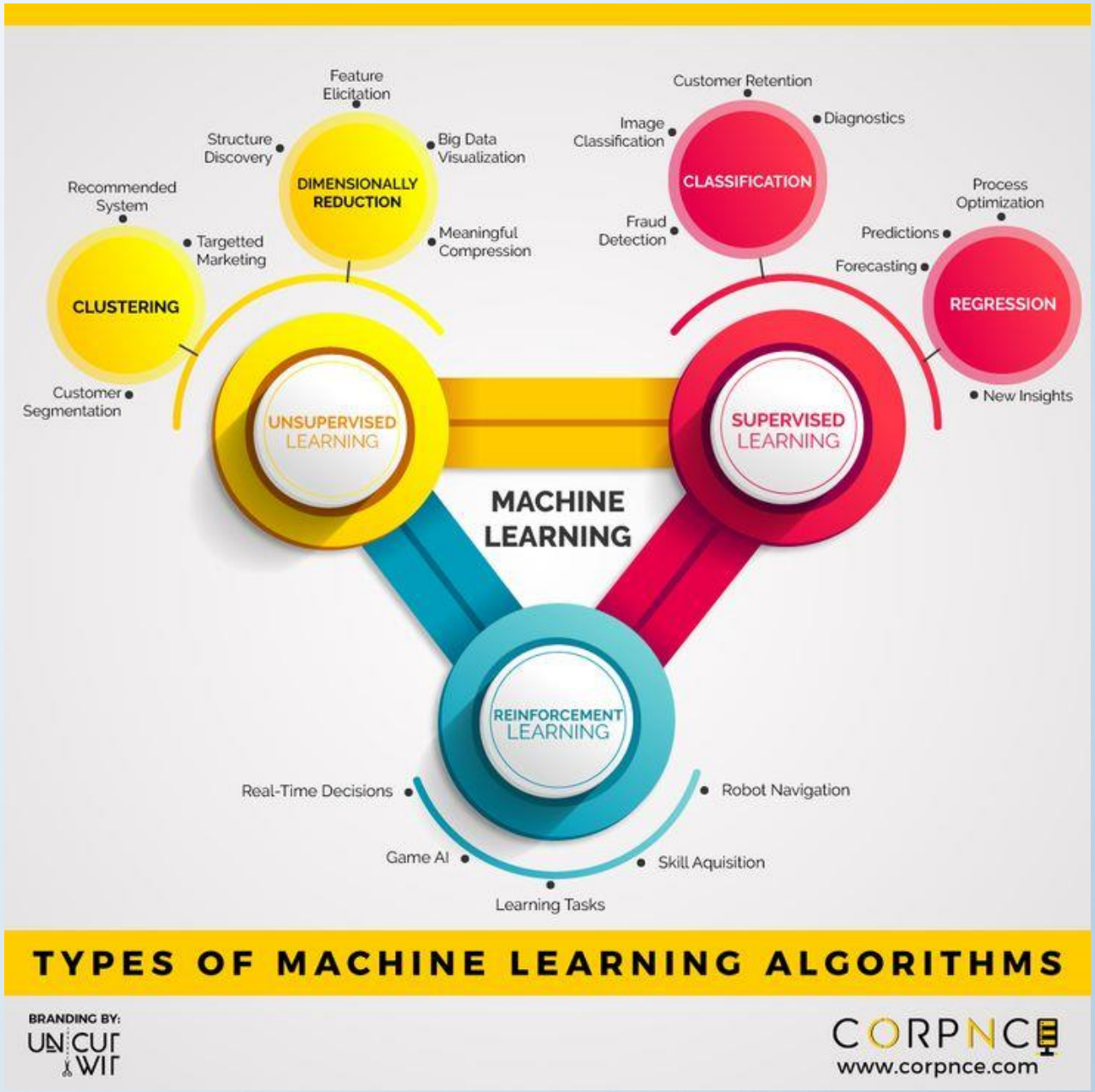
Методы машинного обучения

ИУ-5, магистратура, 2 семестр,
весна 2022 года

Методы работы с признаками

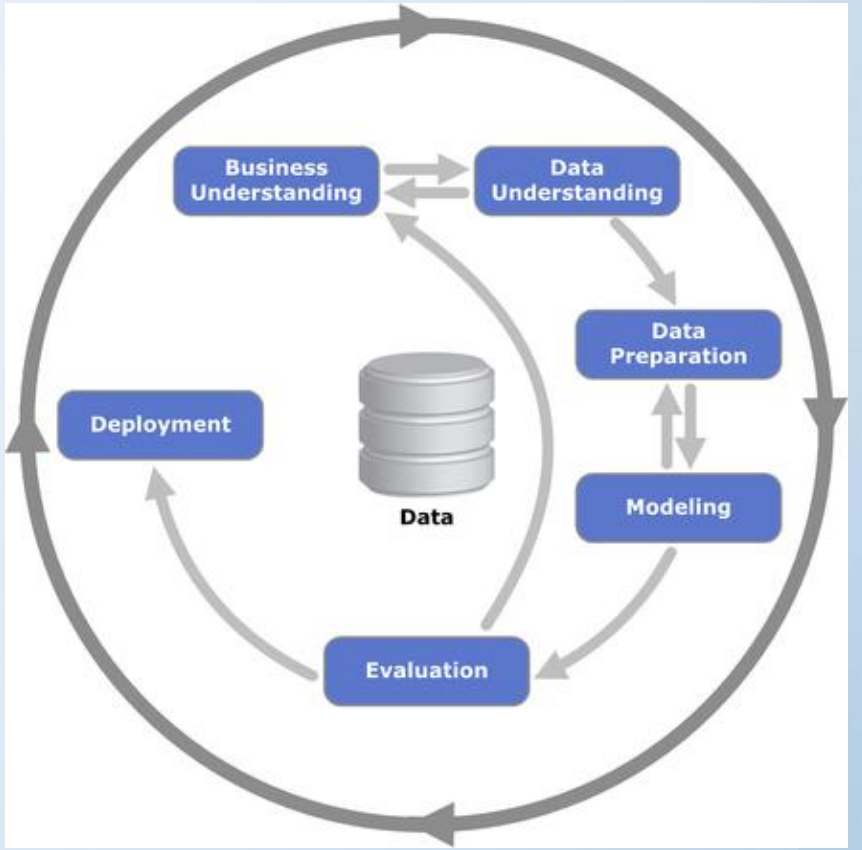
Типы («Классификация») задач ML

- Обучение с учителем (supervised learning)
  - Классификация
  - Регрессия
  - Прогнозирование временных рядов
- Обучение без учителя (unsupervised learning)
  - Кластеризация
  - Методы понижения размерности
- Обучение с подкреплением (reinforcement learning)



CRISP-DM

- CRISP-DM (Cross-Industry Standard Process for Data Mining – межотраслевой стандартный процесс для исследования данных) – проверенная в промышленности и наиболее распространённая методология по исследованию данных.
- Первые версии предложены в конце 1990-х годов.
- Модель жизненного цикла исследования данных состоит из шести фаз, а стрелки обозначают наиболее важные и частые зависимости между фазами. Последовательность этих фаз строго не определена. Как правило в большинстве проектов приходится возвращаться к предыдущим этапам, а затем снова двигаться вперед. Описание фаз:
  1. Понимание бизнес-целей (Business Understanding)
  2. Начальное изучение данных (Data Understanding)
  3. **Подготовка данных (Data Preparation)**
  4. Моделирование (Modeling)
  5. Оценка *качества модели* (Evaluation)
  6. Внедрение (Deployment)



Методы работы с признаками

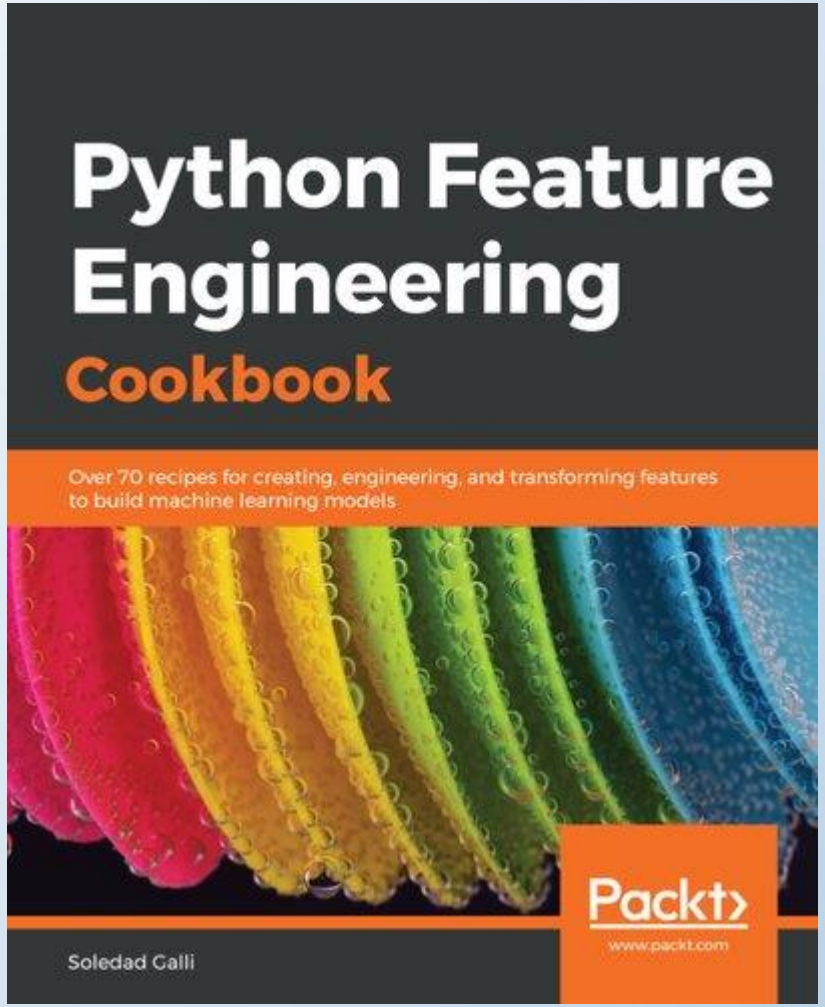
- Выделение признаков (Feature extraction)
- Отбор признаков (Feature selection)
  - «Техники отбора признаков следует отличать от выделения признаков. Выделение признаков создаёт новые признаки как функции от оригинальных признаков, в то время как отбор признаков возвращает подмножество признаков».
- Конструирование признаков (Feature engineering)
- Очистка данных (Data cleansing)

Какие задачи входят в «Feature Engineering for Machine Learning» (по версии Soledad Gally)

- Заполнение пропусков в данных
- Кодирование категориальных признаков
- Масштабирование и трансформация признаков
- Обработка выбросов в данных
- Особенности обработки различных видов признаков (дата и время)

Источники

- Курсы на Kaggle
  - Data Cleaning
  - Feature Engineering
- Курсы на Udemy (Soledad Gally)
  - Feature Engineering for Machine Learning
  - Feature Selection for Machine Learning
  - Machine Learning with Imbalanced Data
  - Deployment of Machine Learning Models
- Лекция в курсе ODS



Порядок обработки данных при выполнении «Feature Engineering» и «Feature Selection»

I. Анализ набора данных, типов признаков и характеристик признаков
II. Генерация дополнительных признаков
III. Устранение пропусков в данных
IV. Кодирование категориальных признаков
V. Нормализация числовых признаков
VI. Обработка выбросов в данных
VII. Масштабирование признаков
VIII. Отбор признаков (Feature selection)

Анализ набора данных, типов признаков и характеристик признаков

- Формально отличается от шага «Начальное изучение данных (Data Understanding)» или «Описательный (разведочный) анализ данных (Exploratory Data Analysis)». Но на практике данный шаг предполагает выполнение EDA.
- Какие характеристики исходных признаков анализируются:
  - Тип признака (числовой, категориальный, другого типа). Тексты и изображения обрабатываются специальными способами.
  - Распределение числовых и категориальных признаков.
  - Наличие выбросов.
  - Наличие пропусков.
  - Исходный масштаб признаков (требуется ли масштабирование).

Генерация дополнительных признаков

- Особенно важна для данных нестандартных типов.
- Позволяет обогатить набор признаков дополнительными сведениями из предметной области, что может улучшить качество модели.
  1. Генерация новых признаков на основе признаков исходного датасета.
  2. Привлечение признаков из дополнительных источников. При этом необходимо решать задачу корректного маппинга новых признаков на объекты существующего датасета.
- Для сгенерированных признаков необходимо также выполнить шаг I).