Машинное обучение

Лекция 1. Постановка задачи машинного обучения

Что такое машинное обучение?

Что такое машинное обучение?

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Что такое машинное обучение?

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Зачем нужно машинное обучение?

Что такое машинное обучение?

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Зачем нужно машинное обучение?

- Автоматизация

Что такое машинное обучение?

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Зачем нужно машинное обучение?

- Автоматизация
- Поиск закономерностей, которые человек не может найти

Постановка задачи

Задача машинного обучения:

- Данные (что такое объект, какие признаки);
- Что предсказывать;
- Оценка качества (критерий качества + способ валидации).

Постановка задачи: Данные

Матрица объекты-признаки

	Признак 1	**Признак 2**	...	**Признак K**
Объект 1	0.77	1742	···	red
Объект 2	0.79	436	···	blue
Объект 3	0.82	910	···	green
...	...	...	...	...
Объект K	0.83	1054	···	blue

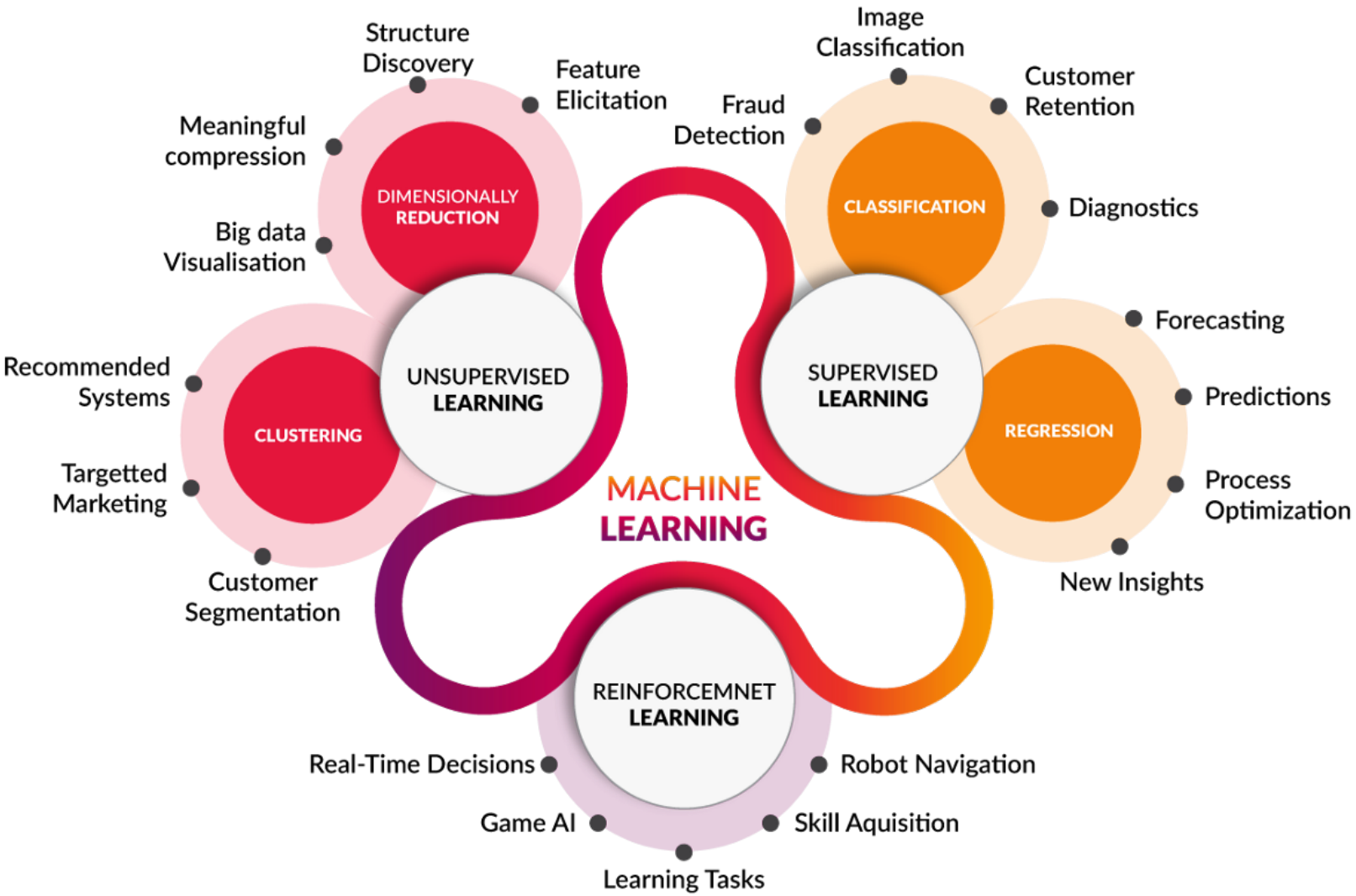
Постановка задачи: Данные

Объект — вектор в [конечномерном] пространстве признаков.
Виды признаков:

1. вещественные
2. бинарные
3. категориальные
4. порядковые (упорядоченные категориальные)
5. подмножество
6. строковые

Категориальные признаки можно кодировать через one-hot кодирование.

Постановка задачи: Что предсказываем?



Классификация

- X — множество объектов.

Классификация

- X — множество объектов.
- Y — множество номеров классов.

Классификация

- X — множество объектов.
- Y — множество номеров классов.
- Конечная обучающая выборка $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Классификация

- X — множество объектов.
- Y — множество номеров классов.
- Конечная обучающая выборка $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Существует *неизвестная целевая зависимость — отображение* $y^* : X \to Y$, значения которой известны только на объектах конечной обучающей выборки X^m. Требуется построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$.

Кластеризация

- X — множество объектов, $\rho(x, x')$ метрика на X.

Кластеризация

- X — множество объектов, $\rho(x, x')$ метрика на X.
- Y — множество номеров кластеров.

Кластеризация

- X — множество объектов, $\rho(x, x')$ метрика на X.
- Y — множество номеров кластеров.
- Конечная обучающая выборка $X^m = \{x_1, \ldots, x_m\} \subset X$.

Кластеризация

- X — множество объектов, $\rho(x, x')$ метрика на X.
- Y — множество номеров кластеров.
- Конечная обучающая выборка $X^m = \{x_1, \ldots, x_m\} \subset X$.

Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.

Кластеризация

- X — множество объектов, $\rho(x, x')$ метрика на X.
- Y — множество номеров кластеров.
- Конечная обучающая выборка $X^m = \{x_1, \ldots, x_m\} \subset X$.

Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.

В чем отличие от классификации?

Кластеризация

- X — множество объектов, $\rho(x, x')$ метрика на X.
- Y — множество номеров кластеров.
- Конечная обучающая выборка $X^m = \{x_1, \ldots, x_m\} \subset X$.

Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.

В чем отличие от классификации? Кластеризация отличается от классификации тем, что номера исходных объектов y_i изначально не заданы, и даже может быть неизвестно само множество Y.

Регрессия

- Задана выборка — множество $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N | \boldsymbol{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной.

Регрессия

- Задана выборка — множество $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N | \boldsymbol{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной.
- Задана регрессионная модель — параметрическое семейство функций $f(\boldsymbol{w}, \boldsymbol{x})$ зависящая от параметров $\boldsymbol{w} \in \mathbb{R}^W$ и свободных переменных $\boldsymbol{x}$.

Регрессия

- Задана выборка — множество $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N | \boldsymbol{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной.
- Задана регрессионная модель — параметрическое семейство функций $f(\boldsymbol{w}, \boldsymbol{x})$ зависящая от параметров $\boldsymbol{w} \in \mathbb{R}^W$ и свободных переменных $\boldsymbol{x}$.

Регрессия

- Задана выборка — множество $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N | \boldsymbol{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной.
- Задана регрессионная модель — параметрическое семейство функций $f(\boldsymbol{w}, \boldsymbol{x})$ зависящая от параметров $\boldsymbol{w} \in \mathbb{R}^W$ и свободных переменных $\boldsymbol{x}$.

Выборку обозначают как $D = \{(\boldsymbol{x}, y)_i\}$. Требуется найти наиболее вероятные параметры $\bar{\boldsymbol{w}}$:

$$\bar{\boldsymbol{w}} = \arg \max_{\boldsymbol{w} \in \mathbb{R}^W} p(D | \boldsymbol{w}, f).$$

Регрессия

- Задана выборка — множество $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N | \boldsymbol{x} \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1, ..., y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной.
- Задана регрессионная модель — параметрическое семейство функций $f(\boldsymbol{w}, \boldsymbol{x})$ зависящая от параметров $\boldsymbol{w} \in \mathbb{R}^W$ и свободных переменных $\boldsymbol{x}$.

Выборку обозначают как $D = \{(\boldsymbol{x}, y)_i\}$. Требуется найти наиболее вероятные параметры $\bar{\boldsymbol{w}}$:

$$\bar{\boldsymbol{w}} = \arg \max_{\boldsymbol{w} \in \mathbb{R}^W} p(D | \boldsymbol{w}, f).$$

В каком смысле правдоподобные?

Постановка задачи: Критерии качества

$$Q(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i))$$

- ℓ — размер тестовой выборки.
- y_i — правильный ответ на i-ом объекте.
- $a(x_i)$ — предсказанный ответ на i-ом объекте.
- L — функция потерь ("loss function")

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Для регрессии:

- $L(y_i, a(x_i)) = y_i - a(x_i)$

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Для регрессии:

- $L(y_i, a(x_i)) = y_i - a(x_i)$ — *плохо, почему?*

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Для регрессии:

- $L(y_i, a(x_i)) = y_i - a(x_i)$ — *плохо, почему?*
- $L(y_i, a(x_i)) = |y_i - a(x_i)|$ — mean absolute error

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Для регрессии:

- $L(y_i, a(x_i)) = y_i - a(x_i)$ — *плохо, почему?*
- $L(y_i, a(x_i)) = |y_i - a(x_i)|$ — mean absolute error
- $L(y_i, a(x_i)) = (y_i - a(x_i))^2$ — mean square error

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Для регрессии:

- $L(y_i, a(x_i)) = y_i - a(x_i)$ — *плохо, почему?*
- $L(y_i, a(x_i)) = |y_i - a(x_i)|$ — mean absolute error
- $L(y_i, a(x_i)) = (y_i - a(x_i))^2$ — mean square error

Для классификации:

Постановка задачи: Функция потерь

Какая может быть функция потерь?

Для регрессии:

- $L(y_i, a(x_i)) = y_i - a(x_i)$ — *плохо, почему?*
- $L(y_i, a(x_i)) = |y_i - a(x_i)|$ — mean absolute error
- $L(y_i, a(x_i)) = (y_i - a(x_i))^2$ — mean square error

Для классификации:

$$L(y_i, a(x_i)) = \begin{cases} 1 & y_i = a(x_i) \\ 0 & y_i \neq a(x_i) \end{cases}$$

— accuracy