Занятие 1. Введение в машинное обучение

Машинное обучение и большие данные, 4 семестр

Иван Евгеньевич Бугаенко,
ассистент каф. ПМиФИ

Формат работы

Курс состоит из:

- практических занятий (Моисеева Н.А., 4.5 балла)
- 8 лабораторных работ (46.5 баллов)
- расчетно-графической работы (9 баллов)
- экзамена (40 баллов)

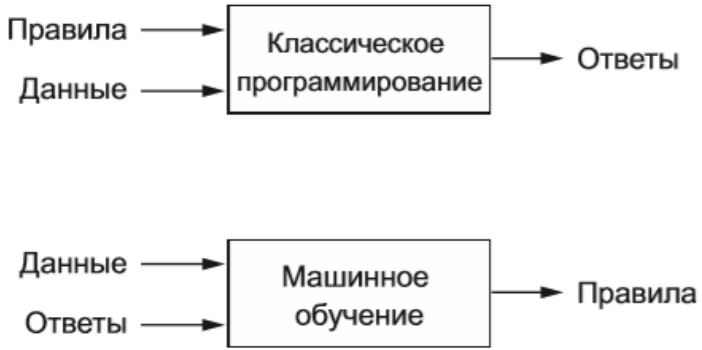
Итого: 100 баллов

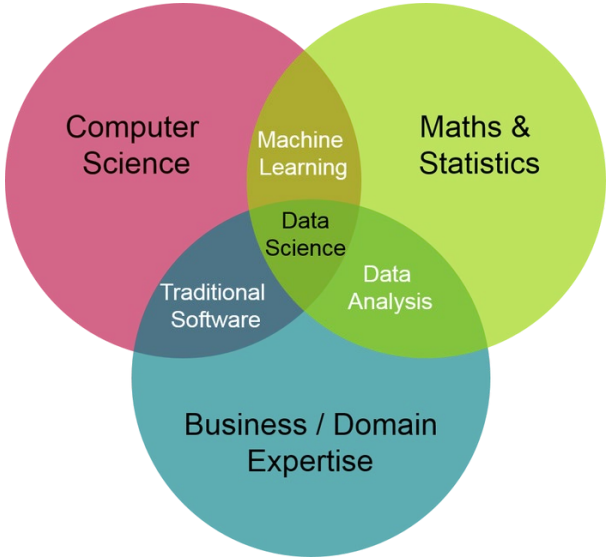
Все материалы по ЛР выкладываются на wiki.pmifi.ru

Рейтинг ведется в таблице

Что такое машинное обучение?

Машинное обучение – это алгоритмы, которые находят в данных скрытые закономерности





Виды задач в машинном обучении

Без чего невозможно машинное обучение?

Виды данных для машинного обучения

Машинное обучение превращает данные в знания

Основные понятия в ML

x – объект
$\mathbb{X}$ – множество объектов ($x \in \mathbb{X} \subseteq \mathbb{R}^d$)

$y = y(x)$ – ответ (метка), истинное правило, как объекту соответствует метка
$\mathbb{Y}$ – множество ответов ($x \in \mathbb{Y}$, $y: \mathbb{X} \to \mathbb{Y}$)

$x = (x_1, x_2, \dots, x_d)$ – признаковое описание объекта (разложение в признаковом пространстве), x_j – число ($j = \overline{1, d}$), d – число признаков (предикторов)

$x^i = \left(x_1^i, x_2^i, \dots, x_d^i\right)$ – признаковое описание x^i объекта (i-го объекта)

Основные понятия в ML

$x^i = \left(x_1^i, x_2^i, \dots, x_d^i\right)$ – признаковое описание x^i объекта (i-го объекта)

Если мы собрали M объектов ($i = \overline{1, M}$), то:

$$X = \begin{pmatrix} x^1 \\ x^2 \\ x^3 \\ \vdots \\ x^M \end{pmatrix} = \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_d^1 \\ x_1^2 & x_2^2 & \dots & x_d^2 \\ x_1^3 & x_2^3 & \dots & x_d^3 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \dots & x_d^M \end{pmatrix}$$ – матрица «объект-признак»

Основные понятия в ML

Если мы собрали M объектов ($i = \overline{1, M}$), то:

$$X = \begin{pmatrix} x^1 \\ x^2 \\ x^3 \\ \vdots \\ x^M \end{pmatrix} = \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_d^1 \\ x_1^2 & x_2^2 & \dots & x_d^2 \\ x_1^3 & x_2^3 & \dots & x_d^3 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \dots & x_d^M \end{pmatrix}$$ – матрица «объект-признак»

$$\chi = (x^i \quad y_i)_{i=1}^M = \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_d^1 & y_1 \\ x_1^2 & x_2^2 & \dots & x_d^2 & y_2 \\ x_1^3 & x_2^3 & \dots & x_d^3 & y_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^M & x_2^M & \dots & x_d^M & y_M \end{pmatrix}$$ – датасет (обучающая выборка)

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

Основные понятия в ML

# price	# area	# latitude	# longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

четвертый объект (x^4) – параметры конкретного дома

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.001640226154105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

четвертая метка (y_4) – стоимость конкретного дома

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

признаки (дескрипторы)

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

целевой признак (то, что хотим предсказать)

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

входные признаки (то, по чему хотим предсказать)

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

матрица объект-признак
(матрица входных данных)

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.001640226154105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

столбец целевых меток

Основные понятия в ML

# price	# area	▲ latitude	▲ longitude	# Bedrooms	# Bathrooms	# Balcony
22400000.000000004	629.0	19.0327996	72.8963568	2.0	2.0	
35000000.0	974.0	19.0327996	72.8963568	3.0	2.0	
31700000.0	968.0	19.0856	72.909277	3.0	3.0	
18700000.0	629.0	19.155756	72.846862	2.0	2.0	2.0
13500000.0	1090.0	19.177555	72.849887	2.0	2.0	
13000000.0	630.0	19.148058	72.937725	2.0	2.0	
20700000.0	1188.0	19.1549192106419	72.8435657622205	2.0	2.0	2.0
22900000.0	968.0	19.147269	72.848351	3.0	3.0	
17000000.0	820.0	19.0151285	72.8580644	2.0	2.0	
81000000.0	3260.0	19.0016402261541105	72.826532	3.0	3.0	3.0
46000000.0	2500.0	19.157602	72.945243	4.0	4.0	2.0
34000000.0	1085.0	19.1595755	72.9465747	3.0	3.0	1.0
65000000.0	1500.0	19.116129	72.821959	2.0	2.0	
20300000.0	891.0	18.9249211	72.8297224	3.0	3.0	1.0
25000000.0	800.0	19.07907	72.907976	2.0	2.0	1.0

датасет