# Лабораторная работа
# "Машинное обучение, или обучение по прецедентам"

**Цель работы:** ***Обучение по прецедентам***. Построение предсказательной модели.

## Теоретические сведения

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

**Тип обучения**: *Обучение по прецедентам*, или *индуктивное обучение*, основано на выявлении общих закономерностей по частным эмпирическим данным.

**Машинное обучение = Обучение по прецедентам**

Машинное обучение находится на стыке *математической статистики, методов оптимизации* и *классических математических дисциплин*, но имеет также и собственную специфику, связанную с проблемами *вычислительной эффективности* и *переобучения*. Многие методы тесно связаны с *извлечением информации* и *интеллектуальным анализом данных* (*Data Mining*).

*Машинное обучение* — не только математическая, но и практическая, инженерная дисциплина. Практически ни одно исследование в *машинном обучении* не обходится без *эксперимента* на *модельных* или *реальных* данных, подтверждающего практическую работоспособность метода.

## Общая постановка задачи обучения по прецедентам

### Постановка задачи.

Дано конечное множество ***прецедентов*** (объектов, ситуаций), по каждому из которых собраны (измерены) некоторые ***данные***. Данные о прецеденте называют также его ***описанием***. Совокупность всех имеющихся описаний прецедентов называется **обучающей выборкой**.

Требуется по этим *частным* данным выявить *общие зависимости, закономерности*, взаимосвязи, присущие не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые ещё не наблюдались. Т.е. ***требуется восстановить закон природы по экспериментальным наблюдениям***. ***Можно сказать, что это одна из основных задач естествознания***.

***Объекты – это люди, веб-страницы, документы, изделия, фирмы, по которым собирается какая-то информация, ….***

### Формальная постановка задачи

Пусть $X$ – множество описаний объектов, $Y$ – конечное множество возможных ответов.

Существует неизвестная *целевая зависимость* (*target function*) – отображение $y : X \to Y$, значения которой известны только на объектах конечной *обучающей выборки* $X^l = \{(x_i, y_i)\}$, $i=1\ldots l$.

Требуется построить *алгоритм*, *решающую функцию* (*decision function*) $a: X \to Y$, способный приблизить $y$ на всем $X$.

Вопросы к решению:

- Как задаются объекты, и какими могут быть ответы?
- Что означает «*a* приближает *y*»?
- Как строить функцию *a*?

## Признаковое описание.

Наиболее распространённым способом описания прецедентов является **признаковое описание.**

***Признак*** (***feature***) $f$ объекта $x$ – это результат измерения некоторой характеристики объекта.

Следует понимать, что **признаки** – это всего лишь функция над объектом.

**Определение. Признаком** называется отображение $f : X \to D_f$, где $D_f$ – множество допустимых значений признака.

В зависимости от природы множества $D_f$ признаки делятся на несколько типов:

- Если $D_f = \{0,1\}$, то $f$ – *бинарный* признак.
- Если $D_f$ – конечное множество, то $f$ – *номинальный* признак.
- Если $D_f$ – конечное упорядоченное множество, то $f$ – *порядковый* признак.
- Если $D_f = \mathbb{R}$, то $f$ – *количественный* признак.

Фиксируется совокупность $n$ показателей, измеряемых у всех объектов. Если все $n$ показателей числовые, то *признаковые описания* представляют собой числовые векторы размерности $n$.

В более сложных случаях прецеденты описываются *временными рядами* или *сигналами*, *изображениями*, *видеорядами*, *текстами* и т. д.

## Модель восстанавливаемой зависимости и функционал качества.

Для решения *задачи обучения по прецедентам* в первую очередь фиксируется *модель* восстанавливаемой зависимости.

**Определение 1**. **Моделью алгоритмов** называется параметрическое семейство отображений $A=\{ g(x,\theta) \mid \theta\in\Theta\}$, где $g: X \times \Theta \to Y$ – некоторая фиксированная функция, $\Theta$ - множество допустимых значений параметра $\theta$, называемое *пространством параметров* или *пространством поиска* (*search space*).

Процесс подбора оптимального параметра модели $\theta$ по обучающей выборке $X^l$ называют **настройкой** (*fitting*) или **обучением** (*training*, *learning*) алгоритма $a \in A$.

**Определение 2**. **Метод обучения** (*learning algorithm*) – это отображение $\mu : (X \times Y)^l \to A$, которое произвольной конечной выборке $X^l = (xi, yi)$, $i=1\ldots l$, ставит в соответствие некоторый алгоритм $a \in A$. Говорят, что метод $\mu$ *строит* алгоритм $a$ по выборке $X^l$.

Метод обучения должен допускать эффективную программную реализацию.

Затем вводится *функционал качества*, значение которого показывает, насколько хорошо модель описывает наблюдаемые данные. *Алгоритм обучения* (*learning algorithm*) ищет такой набор параметров модели, при котором *функционал качества* на заданной обучающей выборке принимает оптимальное значение. Процесс *настройки* (*fitting*) модели по выборке данных в большинстве случаев сводится к применению численных *методов оптимизации*.

## Примеры

**Пример 1.** Задача о подборе параметров *a*, *b* некоторого физического закона

$$y = a\,e^{-t} + b\,t$$

по результатам измерения величины *y* в моменты времени *t*, приведенным в таблице:

| $t_i$ | 0 | .1 | .2 | .3. | .4 | .5 |
|---|---|---|---|---|---|---|
| $y_i$ | 4.25 | 3.95 | 3.64 | 3.41 | 3.21 | 3.04 |

Для нахождения параметров потребуем соответствия измерений физическому закону, т.е. выполнения шести равенств вида:

$$y_i = a\,e^{-t_i} + b\,t_i \qquad (1)$$

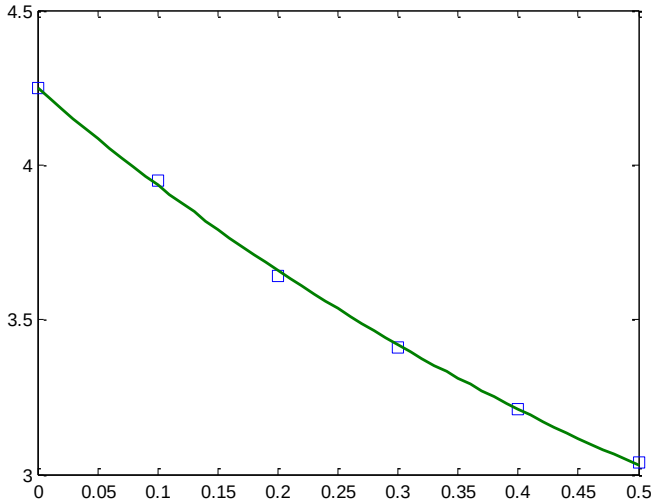
Эти равенства являются *переопределенной системой* из 6 линейных алгебраических уравнений с двумя неизвестными *a* и *b*. Систему (1) можно переписать в векторно-матричном виде

$$A \cdot x = Y. \qquad (2)$$

Матрица *A* и вектор правой части *Y* системы (2) имеют вид

$$A = \begin{bmatrix} e^{-t_1} & t_1 \\ e^{-t_2} & t_3 \\ e^{-t_3} & t_3 \\ e^{-t_4} & t_4 \\ e^{-t_5} & t_5 \\ e^{-t_6} & t_6 \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix}, \quad x = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Решая систему (2), находим требуемые значения параметров *a* и *b*.



## Практические задания

**Задание 1.** Подобрать параметры *a*, *b* и *c* некоторого физического закона

$$y = a\frac{1}{t} + b\sqrt{t} + c\,e^{t}$$

по результатам измерения величины *y* в моменты времени *t*, приведенным в таблице:

| ti | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | 1.2 | 1.4 | 1.6 | 1.8 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| yi | 0.64 | 0.36 | 0.16 | 0.04 | 0.00 | 0.04 | 0.16 | 0.36 | 0.64 | 1.00 |

a) Напишите программу, определяющую параметры **a**, **b**, **c** и строящую маркерами график исходных данных и линией график функции с получившимися параметрами.

b) Исследуемую функцию оформить в виде m-функции от 4 аргументов – переменной **t** и параметров **a**, **b**, **c**.

c) Текст программы оформить в виде m-сценария. В командное окно вывести найденные значения параметров.

d) Нанести на график всю необходимую информацию: заголовок – уравнение с найденными параметрами, подписи к осям, координатную сетку.

**Задание 2**. Известно, что уравнение

**y(t)=a sin t+b cos t+c sin 2t+d cos 2t,**

приближает функцию одной переменной, заданную таблицей значений

| $t_i$ | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| $y_i$ | -6 | -7 | -3 | -2 | 7 | 11 |

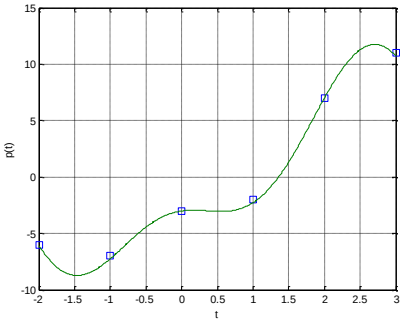
e) Напишите программу, определяющую параметры **a**, **b**, **c**, **d** и строящую маркерами график исходных данных и линией график функции с получившимися параметрами.

f) Исследуемую функцию оформить в виде m-функции от пяти аргументов – переменной **t** и параметров **a**, **b**, **c**, **d**.

g) Текст программы оформить в виде m-сценария. В командное окно вывести найденные значения параметров.

h) Исходные данные должны считываться из текстового файла.

i) Нанести на график всю необходимую информацию: заголовок – уравнение с найденными параметрами, подписи к осям, координатную сетку.



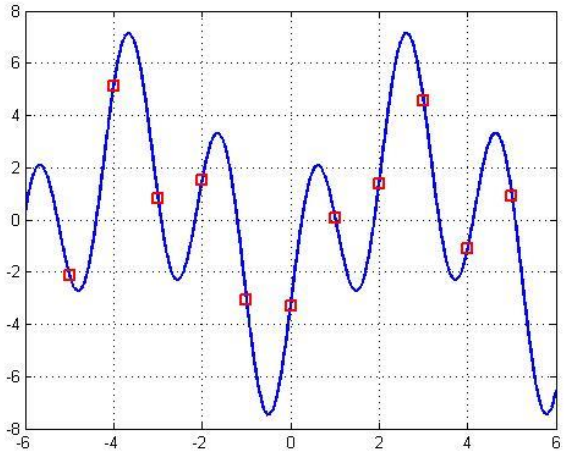
**Задание 3**. Известно, что уравнение

$$y(t)=a \sin t+b \cos t+c \sin 3t+d \cos 2t,$$

приближает функцию одной переменной, заданную двумя обучающими выборками:

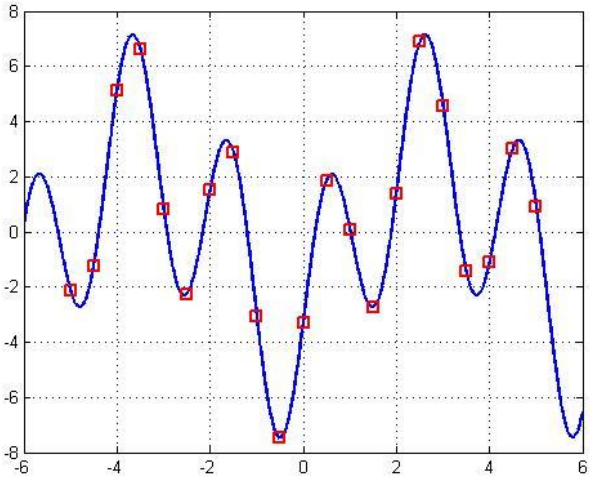
в таблице 1

| $t_i$ | -5 | -4 | -3 | -2- | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | -2 | 5 | 0.8 | 1.5 | -3 | -3.3 | 0 | 1.4 | 4.5 | -1 | 1 |



И в таблице 2:

| $t_i$ | -5 | -4.5 | -4 | -3.5 | -3 | -2.5 | -2 |
|---|---|---|---|---|---|---|---|
| $y_i$ | -2.1 | -1.2 | 5.1 | 6.6 | 0.8 | -2.2 | 1.4 |
| $t_i$ | -1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5 |
| $y_i$ | 2.8 | -3.0 | -7.4 | -3.3 | 1.8 | 0.09 | -2.7 |
| $t_i$ | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| $y_i$ | 1.3 | 6.8 | 4.5 | -1.4 | -1.1 | 3.0 | 0.9 |



a) Напишите программу, определяющую параметры **a**, **b**, **c**, **d** и строящую маркерами график исходных данных и линией график функции с получившимися параметрами.

b) Исследуемую функцию оформить в виде m-функции от пяти аргументов – переменной **t** и параметров **a**, **b**, **c**, **d**.

c) Текст программы оформить в виде m-сценария. В командное окно вывести найденные значения параметров.

d) Исходные данные должны считываться из текстового файла.

e) Нанести на график всю необходимую информацию: заголовок – уравнение с найденными параметрами, подписи к осям, координатную сетку.

f) Проанализируйте, как влияют представленные выборки на точность решения.

**Задание 4**. Задача о подборе параметров *a, b* физического закона

$$v = a\, m^b \qquad (3)$$

описывающего зависимость объема мозга ($v$, $cm^3$) от массы тела ($m$, kg) данного вида млекопитающих. Параметры *a, b* – положительные константы. Результаты измерения объема мозга и веса тела взрослых шимпанзе приведены в таблице:

| $m_i$ | 31 | 36 | 38 | 41 | 42 | 45 | 47 | 48 | 50 | 53 | 55 | 57 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $v_i$ | 365 | 380 | 382 | 395 | 397 | 410 | 410 | 415 | 420 | 427 | 437 | 440 |

a) Напишите программу, определяющую параметры *a, b* и строящую маркерами график исходных данных и линией график функции с получившимися параметрами.

b) Исследуемую функцию оформить в виде m-функции от трех аргументов – переменной *m* и параметров *a, b*.

c) Текст программы оформить в виде m-сценария. В командное окно вывести найденные значения параметров.

d) В одном графическом окне в разных системах координат построить графики полученной функции в логарифмическом масштабе (слева) и декартовой с\к (справа.). Нанести на график всю необходимую информацию: заголовок – уравнение с найденными параметрами, подписи к осям, координатную сетку.

**Подсказка**: Если данные удовлетворяют уравнению степенной зависимости (*power law equation*), то значения параметров *a* и *b* могут быть определены решением системы ЛУ и построением соответствующего графика в логарифмическом масштабе. Для этого прологарифмируем левую и правую части уравнения (3):

$$\ln(v) = b\,\ln(m) + \ln(a).$$

Для нахождения параметров потребуем соответствия измерений физическому закону, т.е. выполнения 12-ти равенств вида:

$$\ln(v_i) = b\,\ln(m_i) + \ln(a)$$

или, соответственно, обозначив $y=\ln(v)$, $x=\ln(m)$, $k=\ln(a)$, получим

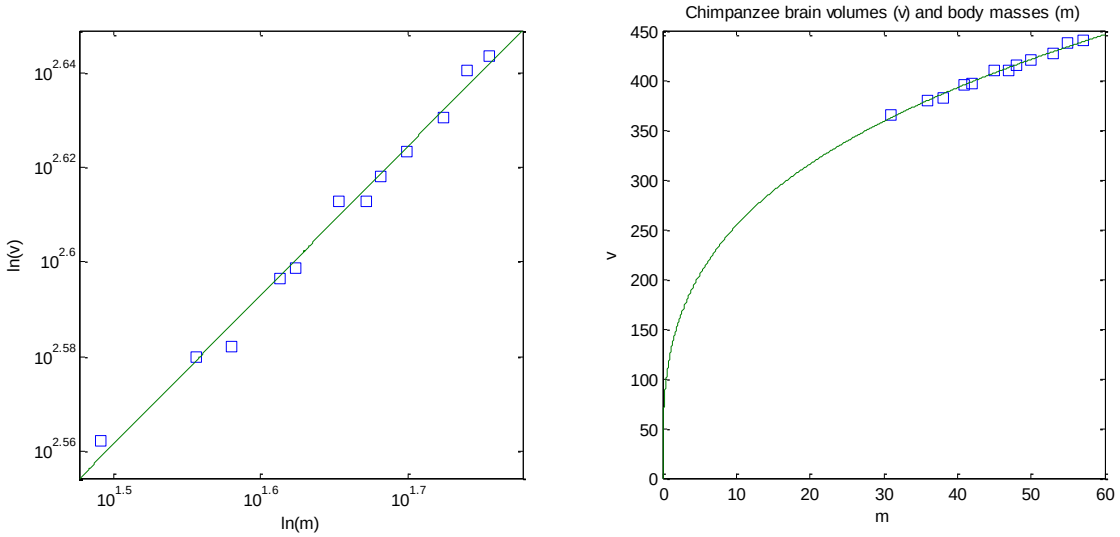
$$y_i = b\,x_i + k. \qquad (4)$$

Эти равенства являются переопределенной системой из 12 линейных алгебраических уравнений с двумя неизвестными *b* и *k*. Матрица *A* и вектор правой части *Y* системы (4) имеют вид

$$A=\begin{bmatrix} \ln(m_1) & 1 \\ \ln(m_2) & 1 \\ \dots & \dots \\ \ln(m_{12}) & 1 \end{bmatrix}, \quad Y=\begin{bmatrix} \ln(v_1) \\ \ln(v_2) \\ \dots \\ \ln(v_{12}) \end{bmatrix}, \quad x=\begin{bmatrix} b \\ k=\ln(a) \end{bmatrix}.$$

Тогда система (4) имеет вид **`A·x = Y`**. Решая ее, находим коэффициенты *b* и *k*.. После этого находим коэффициент $a=e^k$.

В log-log-plot мы должны получить прямую линию, что отображаем в левой системе координат. В правой системе координат строим степенную функцию.




**`Задание 5.`** Задача о подборе параметров *a*, *b* физического закона

$$v = a\, m^b$$

описывающего зависимость объема древесины (*v*, hundreds of board feet, board foot=2359.7 $cm^3$) хвойного дерева от его возраста (*m*, years) данного вида млекопитающих. Параметры *a*, *b* – положительные константы. Результаты измерения объема древесины и возраста хвойных деревьев приведены в таблице:

| $m_i$ | 50 | 75 | 100 | 125 | 150 | 175 | 200 |
|---|---|---|---|---|---|---|---|
| $v_i$ | 11 | 28 | 56 | 98 | 158 | 225 | 330 |

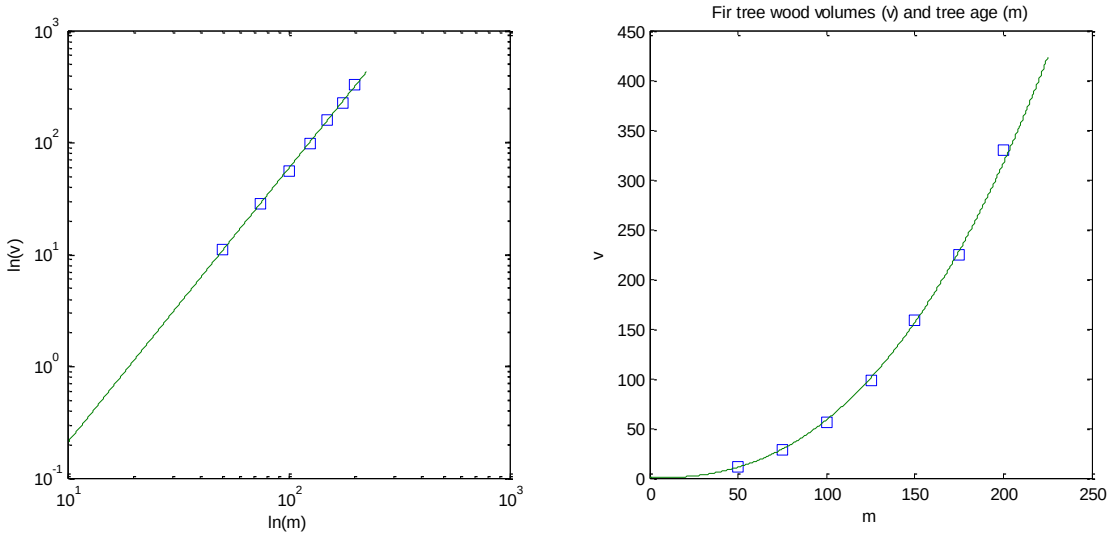
a) Напишите программу, определяющую параметры *a*, *b* и строящую маркерами график исходных данных и линией график функции с получившимися параметрами.

b) Исследуемую функцию оформить в виде m-функции от трех аргументов – переменной *m* и параметров *a*, *b*.

c) Текст программы оформить в виде m-сценария. В командное окно вывести найденные значения параметров.

d) В одном графическом окне в разных системах координат построить графики полученной функции в логарифмическом масштабе (слева) и декартовой с\к (справа.). Нанести на гра-

фик всю необходимую информацию: заголовок – уравнение с найденными параметрами, подписи к осям, координатную сетку.




Закон: $v = 7.26 \times 10^{-4}\ m^{2.45}$.

## Литература

1. Дьяконов А. Чему не учат в анализе данных и машинном обучении. http://alexanderdyakonov.narod.ru/lpot4emu.pdf

2. Enns R.H., McGuire G.C. An Introductory Guide to the Mathematical Models of Science.

3. Воронцов К.В. Курс лекций «Машинное обучение» Вводная лекция http://shad.yandex.ru/lectures/machine_learning.xml

4. http://www.machinelearning.ru/wiki

5. Репозиторий реальных данных UCI (ун-т Ирвина, Калифорния): http://archive.ics.uci.edu/ml

6. Полигон алгоритмов классификации: http://poligon.MachineLearning.ru

7. Конкурсы по решению задач анализа данных: http://www.kaggle.com, http://tunedit.org

8. Сайты, на которых можно тестировать различные алгоритмы на различных данных: http://poligon.MachineLearning.ru, http://mlcomp.org