Московский государственный университет имени М.В. Ломоносова
физический факультет
NeuroFuzzy Lab.

Признаковое описание объектов реального мира. Задачи распознавания образов

Зубюк Андрей Владимирович
zubjuk@physics.msu.ru
http://NeuroFuzzy.Phys.MSU.ru

Числовое описание объектов реального мира

Математика «умеет» работать с числами, но не с объектами реального мира.

Чтобы сделать возможным применение математических методов к объектам реального мира, их (объекты) необходимо описать числами.

В распознавании образов и машинном обучении числовые характеристики объектов принято называть **признаками**. Набор значений признаков объекта — это его **образ**.

Признаки человека: рост h, масса m, возраст a.

Вектор признаков

Обычно признаки объектов объединяются в таблицы (матрицы), например, в таблицы-строки или таблицы-столбцы:

$$\begin{pmatrix} h & m & a \end{pmatrix} \qquad \begin{pmatrix} h \\ m \\ a \end{pmatrix}$$

Такие матрицы будем называть **векторами**.

Такое название может быть объяснено следующим образом. В «обычной» геометрии точка имеет 3 координаты, которые можно записать в матрицу-строку или матрицу-столбец, например, $\begin{pmatrix} x & y & z \end{pmatrix}$ или $\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}$. Вектор — это направленный отрезок. Будем рассматривать только такие векторы, началом которых является точка с координатами $\begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$. Тогда координаты точки-конца вектора полностью определяют весь вектор.

Вообще, понятия «вектор» и «точка» в математике часто употребляются как синонимы.

Некоторые задачи распознавания образов

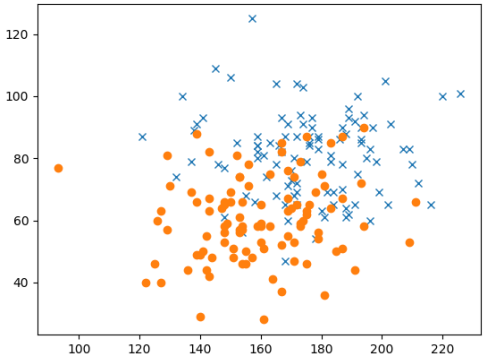
Классификация. Заданы классы объектов. Пример: класс мужчин, класс женщин. Необходимо по значениям признаков объекта определить, к какому классу он относится.

Кластеризация. Необходимо разделить имеющиеся объекты на кластеры (группы, «облака») так, что в пределах одного кластера объекты «схожи» друг с другом, а между кластерами — «заметно различаются». Понятия схожести и различия могут быть определены разными способами, поэтому задача кластеризации имеет не единственное верное решение.

Регрессия. Оценивание значений ненаблюдаемых признаков объектов по значениям наблюдаемых признаков (возможно, других объектов). Пример — прогнозирование погоды.

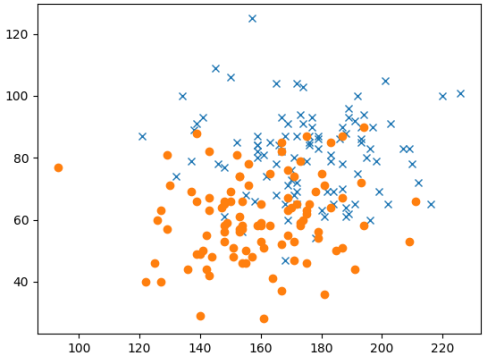
Пример задачи классификации

Даны векторы признаков ряда **известных** объектов из разных классов (рост и масса мужчин и женщин)

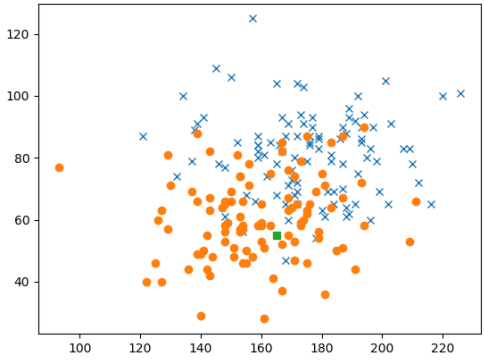


Пример задачи классификации

Даны векторы признаков ряда **известных** объектов из разных классов
(рост и масса мужчин и женщин)

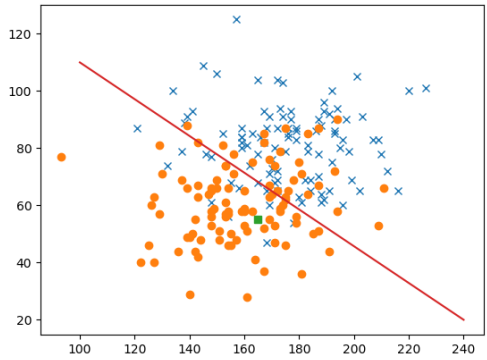


Необходимо по вектору признаков **неизвестного** объекта отнести его к одному из классов
(мужчина или женщина)

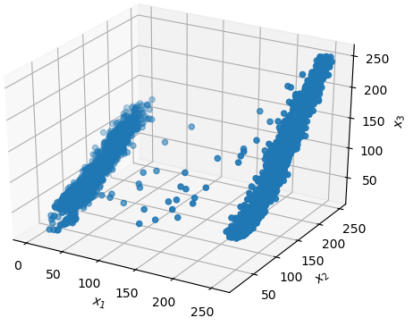


Пример задачи классификации

Способ разделения плоскости прямой

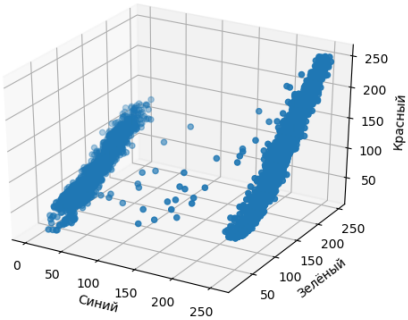


Пример задачи кластеризации



Явно выделяются 2 кластера (группы, «облака») точек. Задача их автоматического выделения — это и есть задача кластеризации.

Пример задачи кластеризации



Признаковое пространство

Понятия «близость», «прямая» и др. должны быть формализованы на языке математики. Для этого необходимо на **множестве всех (гипотетических, не обязательно реальных) векторов признаков** ввести ряд операций, которые превращают это множество в **признаковое пространство**.

Пусть $x = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}$ и $y = \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix}$ — векторы признаков 2-ух разных гипотетических объектов.

Линейные операции: умножение на число α и сложение:

$$\alpha x = \begin{pmatrix} \alpha x_1 & \alpha x_2 & \alpha x_3 \end{pmatrix}, \qquad \alpha y = \begin{pmatrix} \alpha y_1 & \alpha y_2 & \alpha y_3 \end{pmatrix},$$
$$x + y = \begin{pmatrix} x_1 + y_1 & x_2 + y_2 & x_3 + y_3 \end{pmatrix}.$$

Вычитание — это умножение одного вектора на число -1 с последующим сложением.

Признаковое пространство

Скалярное произведение:

$$(x, y) = x_1 y_1 + x_2 y_2 + x_3 y_3 = \sum_{i=1}^{3} x_i y_i.$$

Норма (длина, модуль) вектора признаков:

$$\|x\| = \sqrt{(x, x)} = \sqrt{\sum_{i=1}^{3} x_i^2}, \qquad \|y\| = \sqrt{(y, y)} = \sqrt{\sum_{i=1}^{3} y_i^2}.$$

Расстояние между векторами (точками-концами векторов) x и y:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{3} (x_i - y_i)^2}.$$

Операции в признаковом пространстве и задача классификации

Прямая на плоскости «$x1, x2$» задаётся уравнением

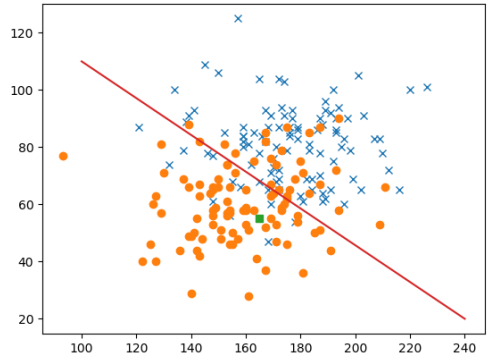
$$w_1 x_1 + w_2 x_2 + b = 0 \quad \Leftrightarrow \quad (w, x) + b = 0.$$

Все векторы $x = \begin{pmatrix} x_1 & x_2 \end{pmatrix}$, лежащие на плоскости с одной стороны от этой прямой, удовлетворяют условию

$$(w, x) + b \leqslant 0,$$

с другой стороны — условию

$$(w, x) + b > 0.$$



$w = \begin{pmatrix} 9 & 14 \end{pmatrix} \quad b = -2440$

Операции в признаковом пространстве и задача кластеризации

Схожесть объектов из одного кластера и различие объектов из разных кластеров могут быть определены с использованием расстояния (метрики) в признаковом пространстве

$$\rho(x, y) = \|x - y\|.$$

- Чем меньше $\rho(x, y)$, тем более схожи соответствующие объекты.
- Чем больше $\rho(x, y)$, тем сильнее они различаются.