

# Метод K-средних (K-means algorithm)

Количественная аналитика — осень 2015

# Основная идея

Основная идея состоит в группировки немаркированных наблюдений в заданное количество кластеров (классов) путём минимизации расстояний до их центров

# Общая схема алгоритма

Задать начальные значения центроидов кластеров

Повторять {

    присвоить наблюдениям номер кластера с ближайшим к  
    ним центром

    передвинуть центроиды кластеров к среднему значению  
    координат их членов

}

# Функция потерь

$K$  — количество классов,  $c^{(i)}$  — класс  $i$ -го наблюдения,  $i \in \{1; \dots; m\}$

$\vec{\mu}_k = [1 \times n]$  — центроид  $k$ -го класса,  $k \in \{1; \dots; K\}$

$$J(c^{(1)}, \dots, c^{(m)}, \vec{\mu}_1, \dots, \vec{\mu}_K) = \frac{1}{m} \sum_{i=1}^m \|\vec{x}^{(i)} - \vec{\mu}_{c^{(i)}}\|^2$$

Более формальный алгоритм:


Повторять {

для  $i = 1$  до  $m$   $c^{(i)} :=$  индекс ближнего центроида


для  $k = 1$  до  $K$   $\vec{\mu}_k := \text{mean}(\vec{x}^{(i)} \in \text{кластер } k)$

}

$\min_{c^{(1)}, \dots, c^{(m)}} J$



$\min_{\vec{\mu}_1, \dots, \vec{\mu}_K} J$



# Метод К-средних в R

Пусть  $X$  — матрица наблюдений

```
km <- kmeans(X, centers = K, nstart = 10, iter.max = 20)
```

```
K-means clustering with 2 clusters of sizes 50, 50
```

```
Cluster means:
```

```
      [,1]      [,2]  
1  0.98398589  1.03541527  
2 -0.03894685 -0.02637371
```

```
Clustering vector:
```

```
[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
[38] 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
[75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
Within cluster sum of squares by cluster:
```

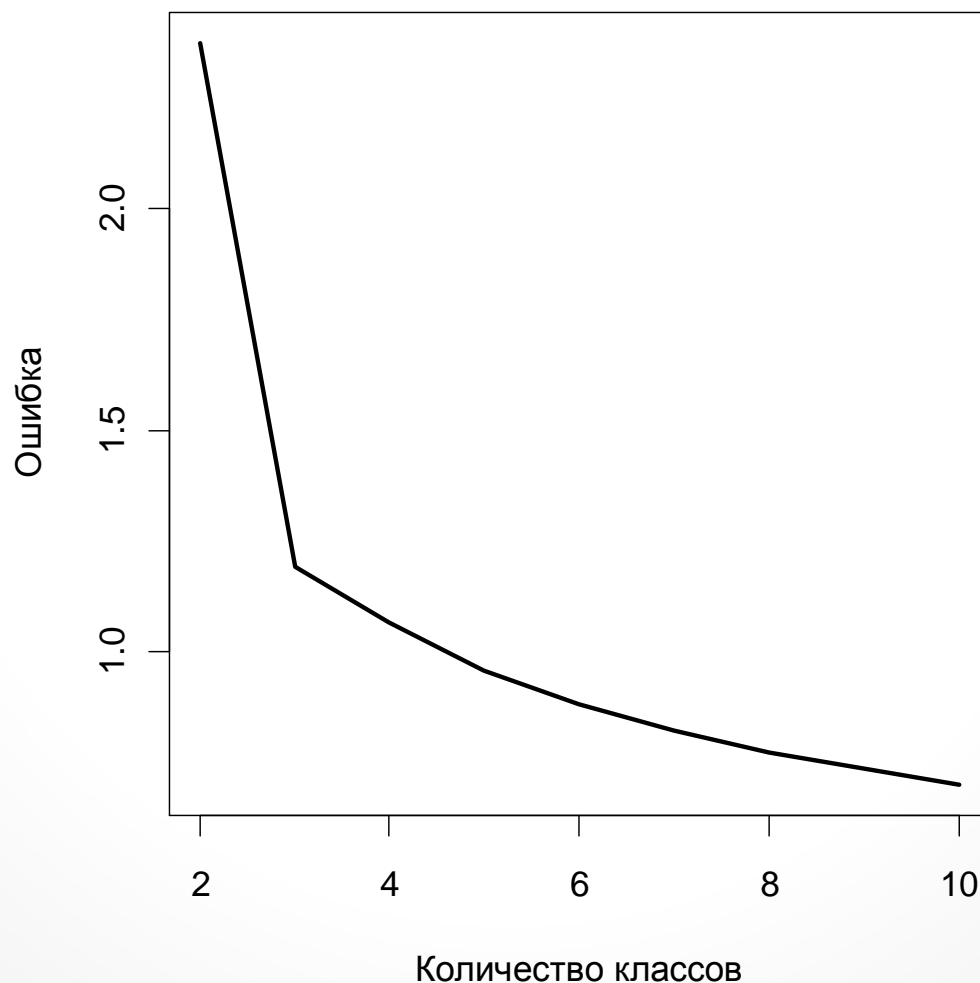
```
[1] 8.535306 10.700694  
(between_SS / total_SS = 73.9 %)
```

```
Available components:
```

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"  
[6] "betweenss"    "size"         "iter"         "ifault"
```

# Выбор количества классов

Количество классов рекомендуется увеличивать до тех пор, пока сохраняется быстрое снижение внутригрупповой ошибки



# Домашнее задание

В файле «[grades.csv](#)» содержатся оценки 304-х студентов по 9-ти предметам

Вашей задачей является разделение этих студентов на академические группы, которое должно осуществляться, исходя из их успеваемости