Классификация: общие принципы

Жилин Сергей Иванович

Алтайский государственный университет

План

- Классификация: общие принципы
- Кластеризация методом К средних
- Классификация методом SIMCA

Классификация: постановка задачи

- Можно ли по спектру отличить кетон от эфира?
- Можно ли определить пол человека по его ответам на вопросы анкеты об автомобилях?
- Можно ли по хроматограмме узнать происхождение вина и если да, то какие именно особенности хроматограммы позволяют это сделать?
- **Как**, зная размеры лепестков, определить к какому виду относится изучаемый цветок?
- Как зная содержание элементов в почве определить из какого она района?

Этапы классификации

Кластеризация (классификация без обучения)

изучение исходных данных на предмет наличия в них групп, классов и определение признаков, которые за это отвечают



Построение модели (классификация с обучением)

нахождение зависимости между значениями признаков объектов и принадлежность их к определенной группе



Классификация новых образцов (распознавание образов)

отнесение неизвестных образцов к одному из известных классов

Алгоритмы классификации

Без обучения (Unsupervised)

Априори не известно существуют ли скрытые группы в данных и сколько их

Основной механизм – поиск аналогий в поведении значений параметров объектов

Основная цель – установить наличие групп (классов), а так же причину – переменные или их комбинации, которые на это влияют (являются схожими для объектов той или иной группы)

С обучением (Supervised)

Априори известно о том, какой группе принадлежит объекты из исходного набора данных

Основной механизм — построение модели, связывающей значения параметров объектов образующих ту или иную группу

Основная цель — использование полученной модели для классификации новых образцов

С чем работаем?

- Объект все, что угодно: пациент, вещество, предмет, *пиксел*, *изображение* и т.д.
- Вектор признаков набор значений переменных, характеризующих объект
- Группа или класс совокупность объектов обладающих схожими характеристиками, например (все или только некоторые) значения признаков которых лежат в определенных границах

Пример:

объект — человек

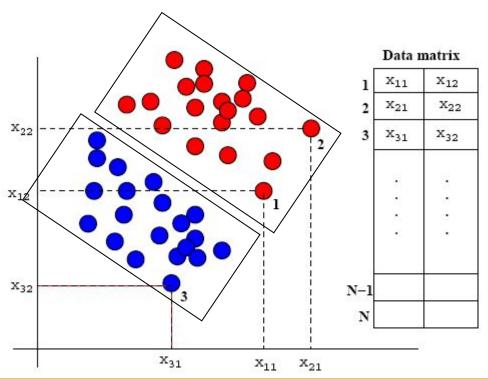
вектор признаков — рост, вес, длина волос, умение плавать, размер обуви, кулинарные предпочтения

возможные группы — по полу, по материку, по стране и т.п.

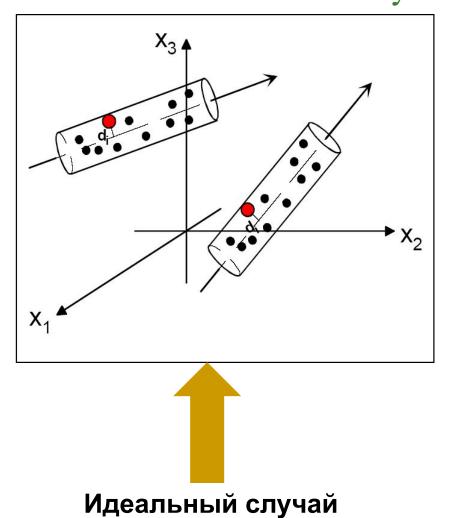
Вектор признаков – переменные (степени свободы) образующие *N*-мерную систему координат (*N* – число переменных в векторе признаков)

Объекты – точки в пространстве признаков

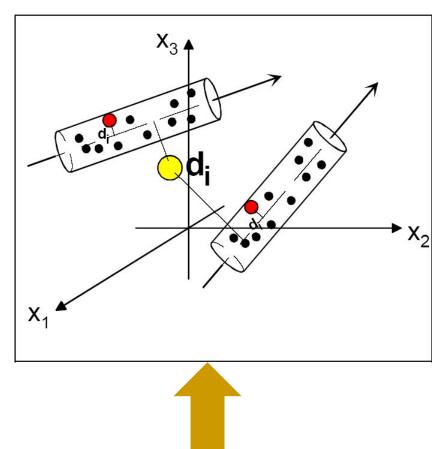
Группы или классы – части пространства признаков



Возможные ситуации

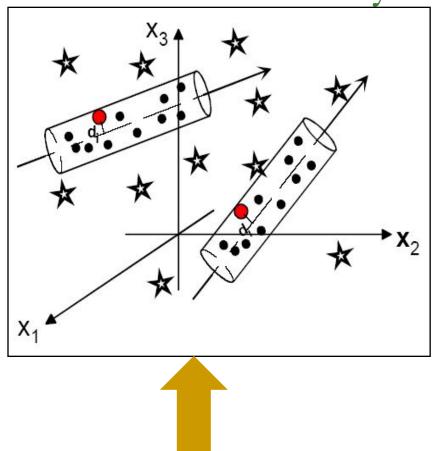


разделения

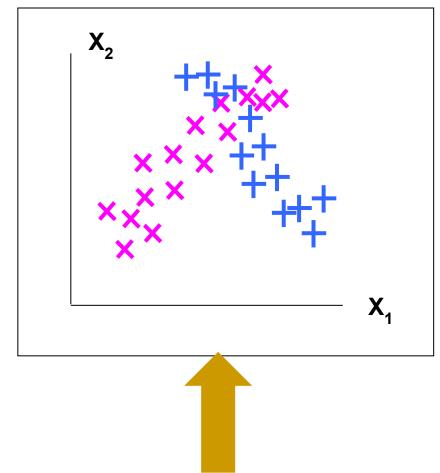


Имеются выбросы

Возможные ситуации

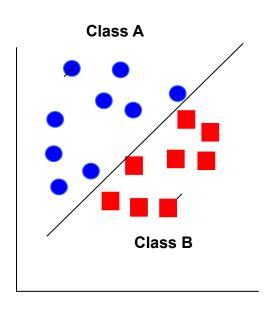


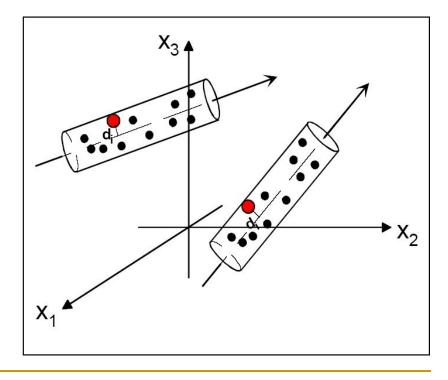




Классы перекрываются

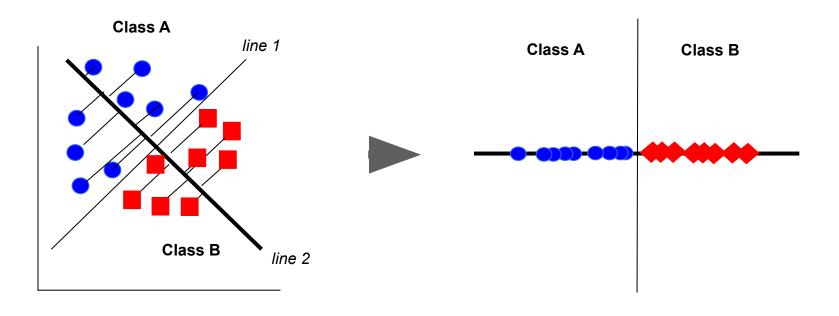
- Как задать классы?
 - 1. Явное задать границы части пространства, соответствующей классу (полупространство, гиперсфера, гиперпрямоугольник, и т.п.)



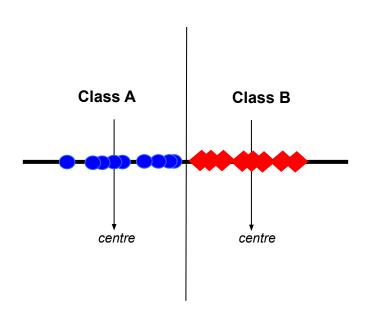


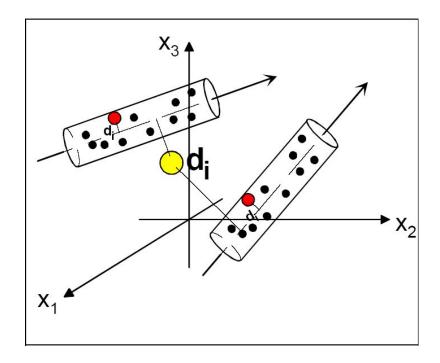
- Как задать классы?
 - 1. Явное задать границы части пространства, соответствующей классу (полупространство, гиперсфера, гиперпрямоугольник, и т.п.)

Часто удобнее оперировать проекциями объектов



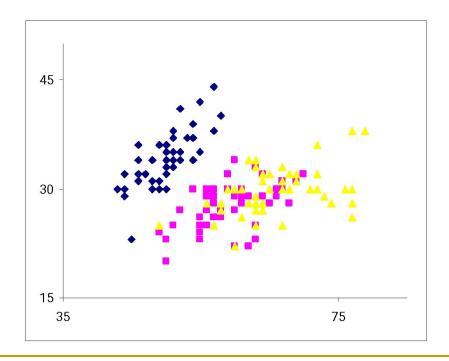
- Два подхода к заданию классов:
 - 2. Степень принадлежности классу определяется расстоянием до класса (до центра, до «каркаса», до границы)





- Два подхода к заданию классов:
 - 2. Степень принадлежности классу определяется расстоянием до класса (до центра, до «каркаса», до границы)

Особенно актуально для ситуаций, когда классы перекрываются



- Два подхода к заданию классов:
 - 2. Степень принадлежности классу определяется расстоянием до класса (до центра, до «каркаса», до границы)

Особенно актуально для ситуаций, когда классы перекрываются

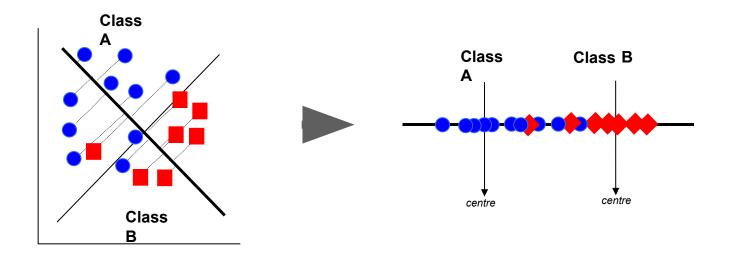
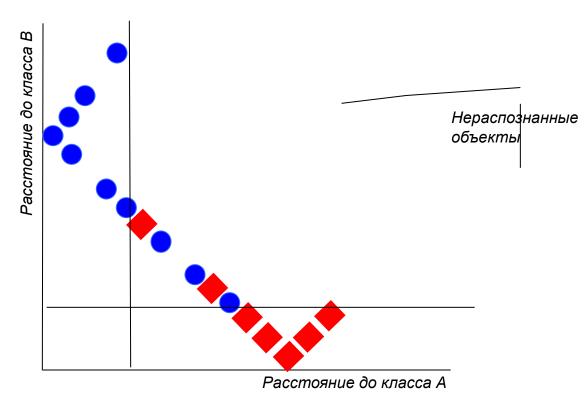


График расстояний

Для проекций объектов

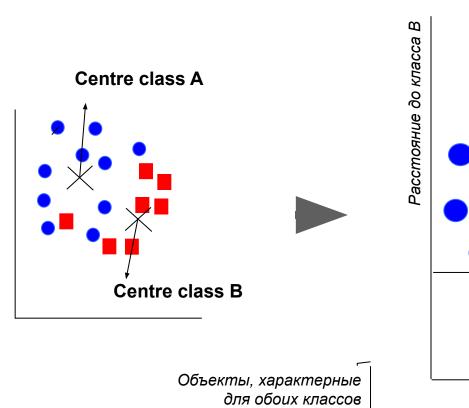


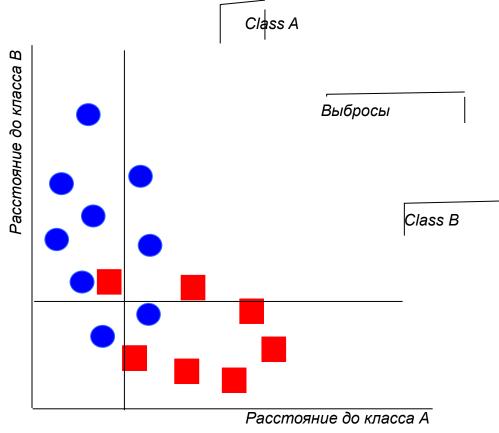
Class A

Class B

График расстояний

В исходном пространстве характеристик





Расстояние

в пространстве характеристик

- Расстояние может задаваться разными метриками!
- Евклидова метрика: $d_{kl} = (\mathbf{x_k} \mathbf{x_l})(\mathbf{x_k} \mathbf{x_l})'$
 - Каждая переменная вектора признаков дает одинаковый вклад наряду с остальными (признаки ортогональны)
 - Если между переменными имеется корреляция то они будут иметь непропорциональное влияние на результаты анализа
- Метрика Махалонобиса: $d_{kl} = (\mathbf{x_k} \mathbf{x_l})C^{-1}(\mathbf{x_k} \mathbf{x_l})'$

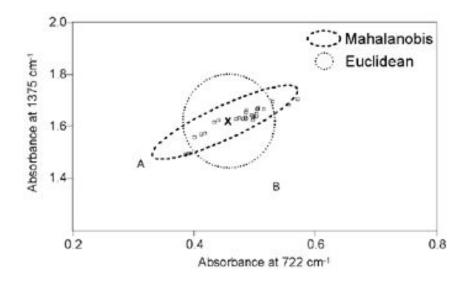
(C - ковариационная матрица)

- Учитывает возможную корреляцию между переменными
- Если корреляция между переменными отсутствует, то расстояние Махаланобиса равно расстоянию Евклида

Расстояние

в пространстве характеристик

- Расстояние может задаваться разными метриками!
 - Евклидова метрика и метрика Махалонобиса:



Расстояние

в пространстве характеристик

• Расстояние может задаваться разными метриками!

1.	Euclidean Distance (L2)	10.	Chord Distance
2.	City Block Distance	11.	Non-Correlation
3.	Canberra Metric	12.	Matusita Distance
4.	Histogram Intersection	13.	Soergel
5.	Jeffrey Divergence	14.	Wave-Hedges
6.	Bhattacharyya Distance	15.	WED Distance
7.	Chi-Square	16.	Kolmogorov-Smirnov
8.	Bray Curtis Distance		Statistic
9.	Angular Separation	17.	Kuiper
	Distance	18.	Mean Distance

Виды ошибок

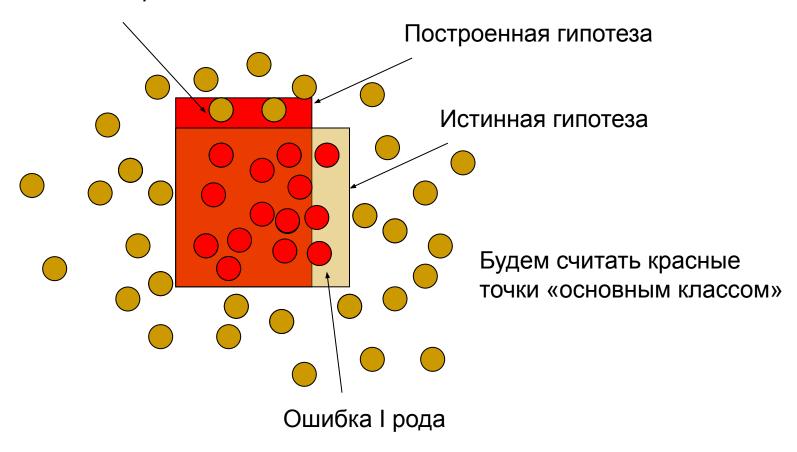
- Измерения ошибки как «вероятности выдать неверный ответ» может быть не всегда достаточно
 - 15% ошибки при постановке диагноза может означать как и то что, 15 % больных будут признаны здоровыми (и возможно умрут от отсутствия лечения), так и то, что 15% здоровых больными (и деньги на лечение будут потрачены зря)
- При неравнозначности ошибок для разных классов вводят понятие ошибки первого и второго рода и замеряют их по отдельности

Ошибки I и II рода

- Пусть, существует «основной класс»
 - Обычно, это класс, при обнаружении которого, предпринимается какое-либо действие;
 - Например, при постановке диагноза основным классом будет «болен», а вторичным классом «здоров».
- Ошибка первого рода принять основной класс за вторичный
 - Вероятность «промаха», когда искомый объект будет пропущен
- Ошибка второго рода равна вероятности принять вторичный класс за основной
 - Вероятность «ложной тревоги», когда за искомый объект будет принят «фон»

Ошибки I и II рода

Ошибка II рода



Ошибки I и II рода

- Что считать основным классом зависит полностью от прикладной специфики
- Особенно важно оценивать ошибки I и II рода раздельно при несбалансированности классов:
 - 🗅 Пусть

$$P(y = +1) = 0.01;$$
 $P(y = -1) = 0.99$

□ Тогда при нулевой ошибке II рода и ошибке I рода 0.5

$$P(a(x) = -1 | y = +1) = 0.5$$

Общая ошибка всего лишь

$$P(a(x) \neq y) = 0.005$$

Чувствительность vs Избирательность

 Чувствительность – вероятность дать правильный ответ на пример основного класса

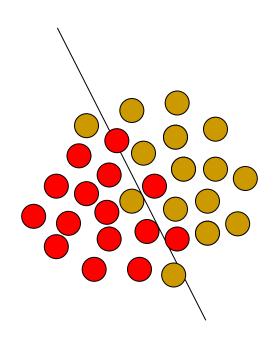
$$sensitivity = P(a(x) = y \mid y = +1)$$

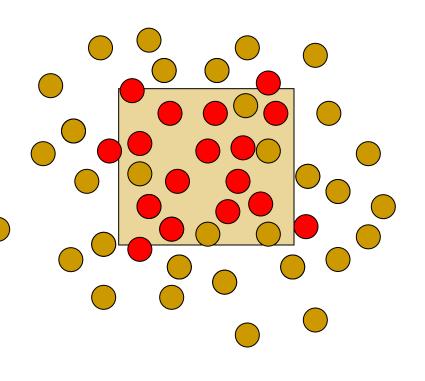
 Избирательность – вероятность дать правильный ответ на пример вторичного класса

$$specificity = P(a(x) = y \mid y = -1)$$

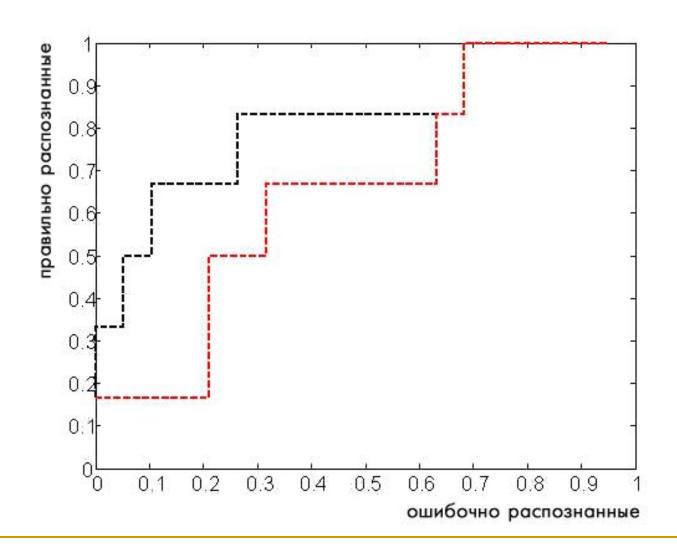
Регулировка баланса

Почти все алгоритмы
классификации допускают
регулировку соотношения ошибки І
и ІІ рода за счет варьирования
некоторого параметра



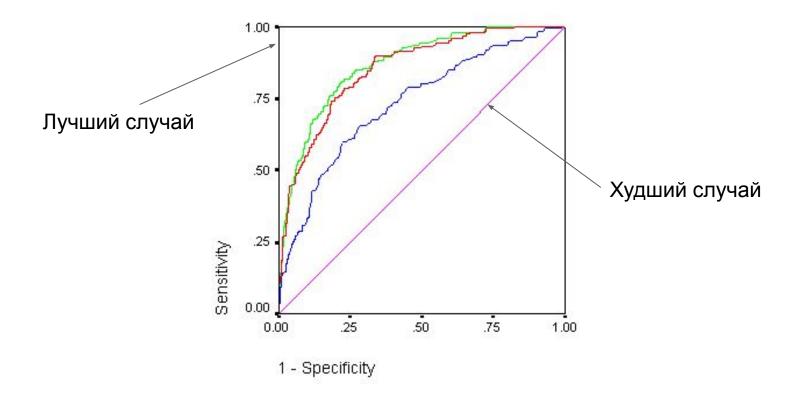


Кривая мощности критерия



ROC-кривая

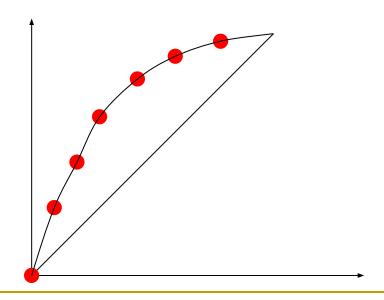
- ROC Receiver Operating Characteristic curve
 - Кривая, отражающая зависимость чувствительности и ошибки второго рода



Построение ROC-кривой

- Для различных значений параметра строится таблица ошибок
 - Сам параметр в таблице не участвует!
 - Классификатор строится и оценивается на разных выборках!
- По таблице строится набор точек в плоскости sensitivity/FP
 - Каждая строка таблицы точка
- По точкам строится кривая

Sensitivity	False Positive	
0.0	0.0	
0.25	0.5	
0.5	0.8	
1.0	1.0	



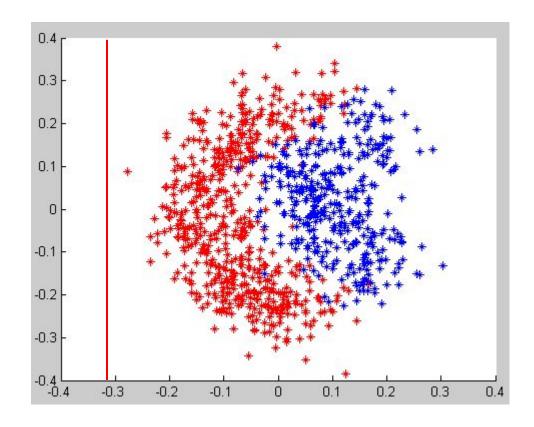
Анализ ROC кривой

- Площадь под графиком AUC
 - Дает некоторый объективный показатель качества классификатора
 - □ Позволяет сравнивать разные кривые
- Соблюдение требуемого значения ошибок I и II рода
 - Зачастую, для конкретной задачи существуют рамки на ошибку определенного рода. С помощью ROC можно анализировать возможность текущего решения соответствовать требованию

ROC: построение таблицы

Меняем порог и оцениваем ошибку

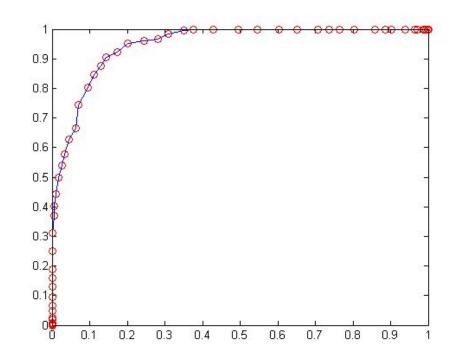
Sensitivity	False Positive
0.0	0.0
0.25	0.5
0.5	0.8
1.0	1.0



ROC: построение кривой

■ По таблице строим точки

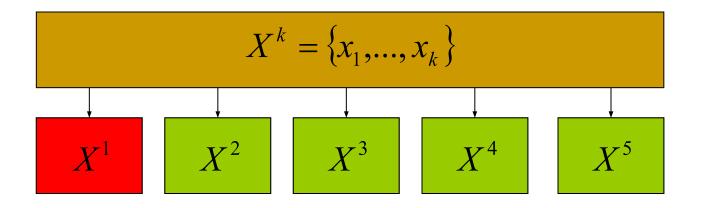
Точки интерполируем кривой



Этапы классификации

• МГК I. Выявление групп • Факторный анализ • Кластерный анализ II. Построение SIMCA модели PLS-DA III. Классификация новых образцов

Перекрестная проверка (cross-validation)





Обучение

Итоговая ошибка – средняя ошибка по всем итерациям

Другие методы классификации с обучением

- Статистические методы
- Нейронные сети
- Деревья классификации
- Бустинг (комитеты решающих правил)
- Метод опорных векторов

