Методы многомерной классификации

Кучерявский С.В.
svk@asu.ru

План

- **Теоретические основы**
  - Что это такое
  - Виды и этапы классификации
  - Как оценить результаты
  - Геометрическая интерпретация
- **Методы многомерной классификации**
  - МГК
  - SIMCA
- **Примеры, обсуждения и выводы**

Часть I. Теоретические основы

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

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

Классификация новых образцов
отождествление неизвестных образцов с одним из известных классов

С чем работаем?

- **Объект** — все, что угодно: пациент, вещество, предмет и т.д.
- **Вектор признаков** — набор переменных и их значений, 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

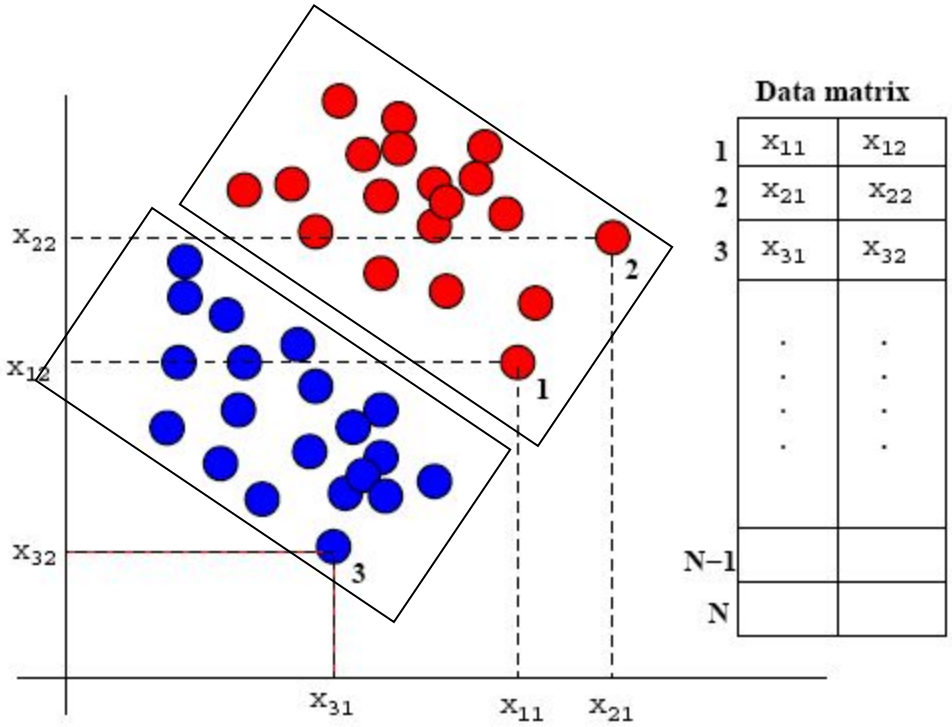
возможные группы — по полу, по материку, по стране и т.п.

Геометрическая интерпретация

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

Объекты – точки в пространстве признаков

Группы или классы – ограниченные подпространства в пространстве признаков: гиперкуб, гиперсфера и т.п.



Алгоритмы классификации

Без обучения (**Unsupervised**)	С обучением (**Supervised**)
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)	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

Возможные ситуации

1. **В начале ни одного класса не определено**

   первым шагом в этом случае является предварительный анализ данных на предмет обнаружения потенциальных групп. В зависимости от результата возможны варианты:

   - **Имеется одна ярко выраженная группа**
   - **Имеется несколько ярко выраженных групп**

   Эти же варианты могут быть известны априори

Возможные ситуации

2. **Имеется одна ярко выраженная группа**

   В этом случае основная задача классификации найти и выделить типичную зависимость в данных для объектов, принадлежащих к одной группе и использовать ее для классификации новых объектов

3. **Имеется несколько ярко выраженных групп**

   Необходимо использовать методы распознавания образов для выяснения принадлежности новых объектов к тому или иному классу. Задачу можно свести к предыдущей ситуации.

Как определить класс?

- Есть данные и некоторая информация о них, как на ее основе определить класс?
- Что такое схожесть объектов, принадлежащих одному классу?

Все зависит от уровня начальных знаний:

A. Известно некоторое характерное свойство

B. Имеется репрезентативный набор данных

C. Известны релевантные переменные

D. Известна зависимость между ними

Фундаментальные знания о классе

Как определить класс? Уровень А

Известно некоторое характерной свойство, если объект обладает этим свойством, он принадлежит классу, в противном случае – нет

Примеры: пол человека или животного, спин частицы, способность лекарства снимать боль и т.п.

Возможные проблемы: очень часто одно свойство не определяет класс, в котором объекты распределены неравномерно, особенно если данное свойство может быть результатом действия разных механизмов

Как определить класс? Уровень B

Аналитик имеет в своем распоряжении набор данных среди которых находятся объекты заведомо принадлежащие данному классу – репрезентативную выборку

Пример: данные химического или спектрального анализа качественных лекарств и подделок, но какие образцы являются подделками, а какие качественными препаратами – неизвестно

Возможные проблемы: необходимо, чтобы выборка как можно полнее покрывала различные вариации, характерные для объектов класса

Как определить класс? Уровень C

В дополнение к уровню B известно так же какие именно переменные из исходного набора определяют принадлежность к классу, т.е. являются релевантными

Пример: данные химического или спектрального анализа качественных лекарств и подделок, причем известно какие образцы относятся к подделкам, а какие – к качественным лекарствам

Возможные проблемы: обычно выявление релевантных переменных происходит методом проб и ошибок и требует времени.

Как определить класс? Уровень D

На данном уровне класс определяется совокупностью релевантных переменных и зависимостью между их значениями. Этот уровень знаний позволяет классифицировать новые, неизвестные объекты

Пример: модель, связывающая данные анализа с принадлежностью образцов к тому или иному классу

Что дальше? Одна из самых простых возможностей углубить уровень знаний после уровня D – интерпретировать известную зависимость и использовать результаты интерпретации

Распознавание образов

Итак класс или классы определены, каким образом классифицировать новые значения? Будем использовать **геометрическую интерпретацию**!

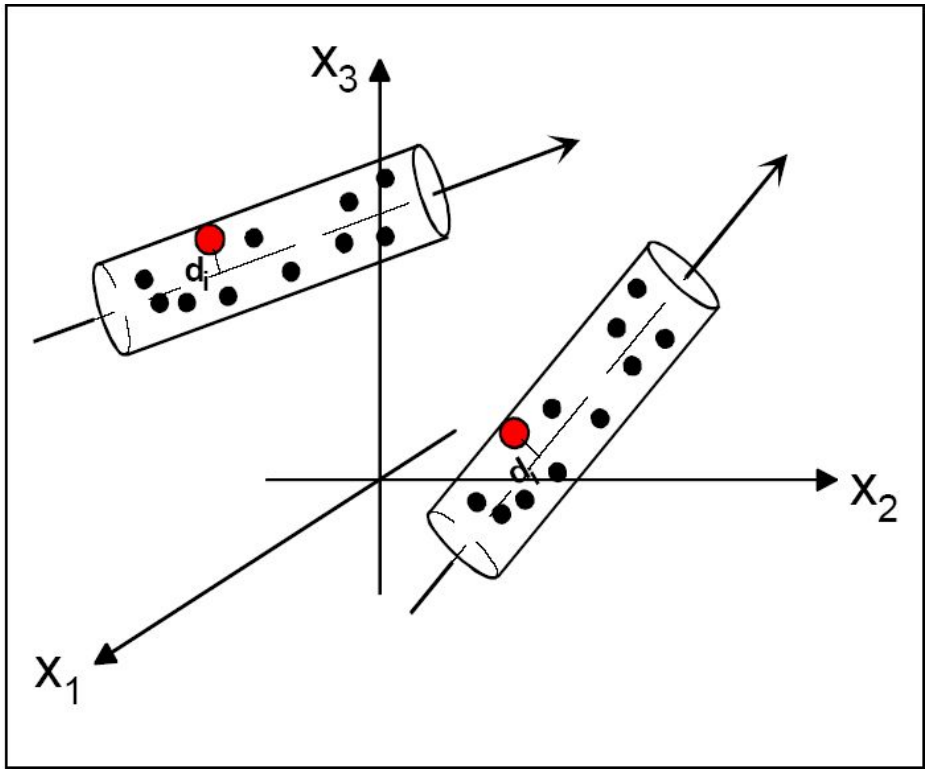
В начале рассмотрим два уровня распознавания образов:

1. Классификация как однозначное определение принадлежности к одному из классов
2. Классификация с определением класса и учетом возможности выбросов

Распознавание образов. Уровень 1

Предполагаем,
что все объекты (как исходного так и тестового набора) принадлежать одному из заранее определенных классов – лежать в соответствующем гиперобъеме

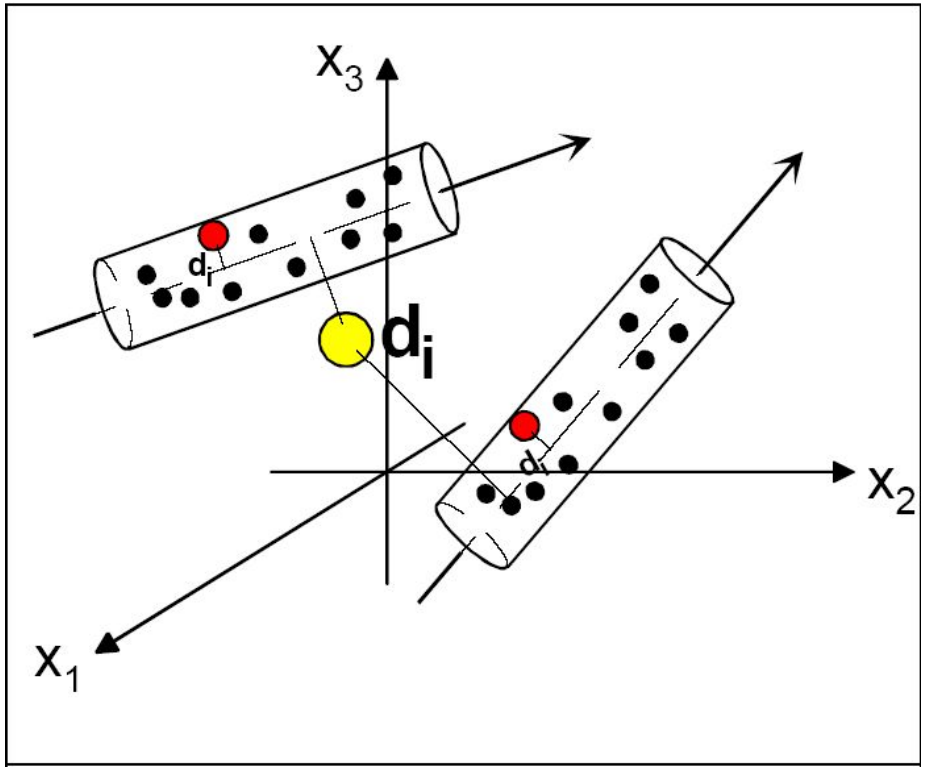
Проблема: в реальных ситуациях такое встречается очень редко



Распознавание образов. Уровень 2

Предполагаем, что помимо объектов, принадлежащих тому или иному классу, возможны выбросы – объекты, не соответствующие ни одному классу, т.е. не попадающие ни в один гиперобъем

Проблемы: один из классов может не иметь определенной геометрической структуры

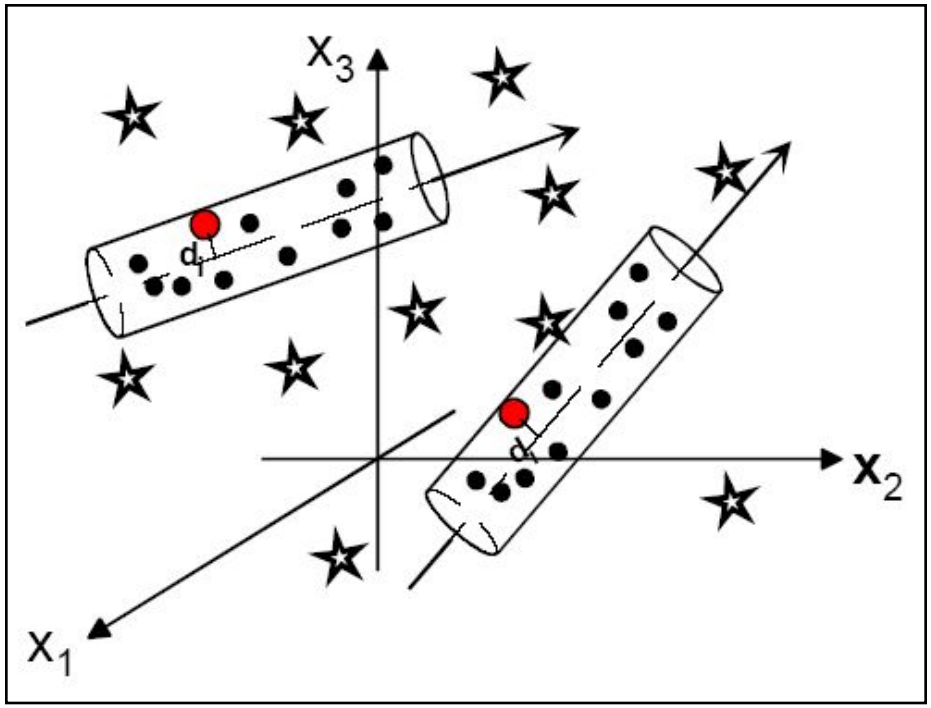


Распознавание образов. Уровень 2а

Асимметричный

случай – один из классов не имеет характерной структуры

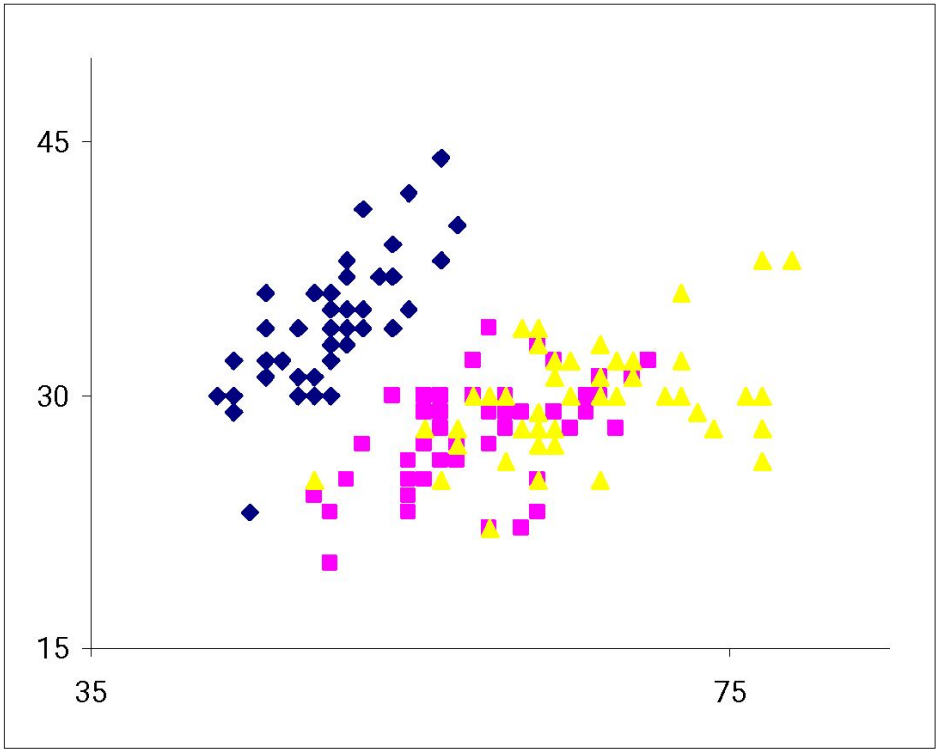
Пример: контролируемый процесс (параметры в жестких рамках) или неконтролируемый (может происходить что угодно)



Какие еще могут быть проблемы?

- Некоторые гиперобъемы могут перекрываться

- Не всегда можно определить правильный геометрический эквивалент группы или класса



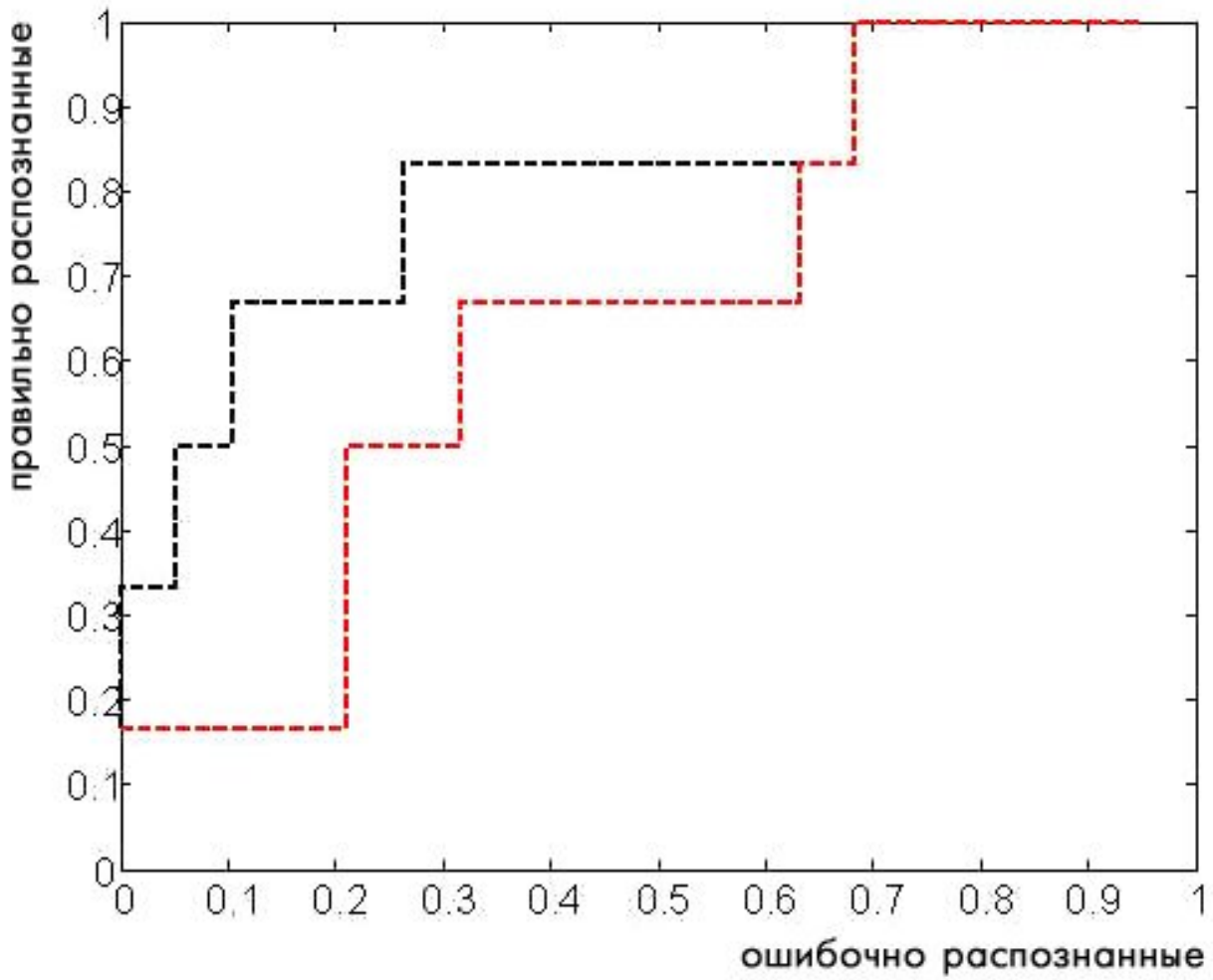
Как оценить эффективность?

Рассмотрим работу одноклассового классификатора:

- **Ошибки первого рода** — образцы, являющиеся членами класса, но ошибочно отклоненные классификатором

- **Ошибки второго рода** — образцы, ошибочно определенные классификатором как члены класса

Кривая мощности критерия



Какие ошибки уменьшать?

Все зависит от конкретного случая:

- Уменьшение **ошибок первого рода**: риск упустить важную информацию выше, чем последствия ее переоценки. Примеры — определения ядовитых веществ, медицинский диагноз

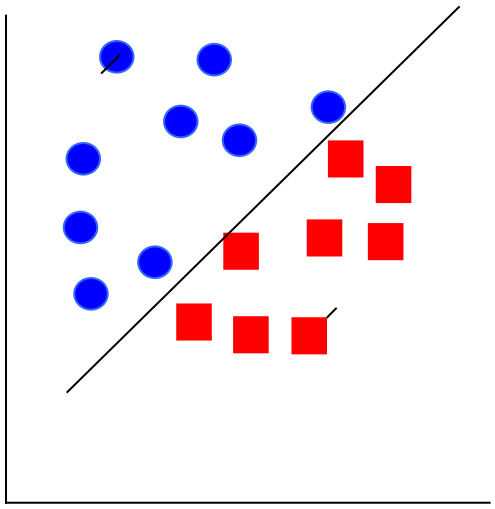
- Уменьшение **ошибок второго рода**: с точностью до наоборот. Примеры — судопроизводство

Как это все реализовать?

- Как определить гиперобъем?
- Как определить схожесть объектов?
- Как вычислить попадает ли объект к данному классу, если объемы перекрываются?
- Как классифицировать выбросы?
- Как уменьшить ошибки первого или второго рода?

Как это все реализовать?

Будем и дальше использовать геометрическую интерпретацию

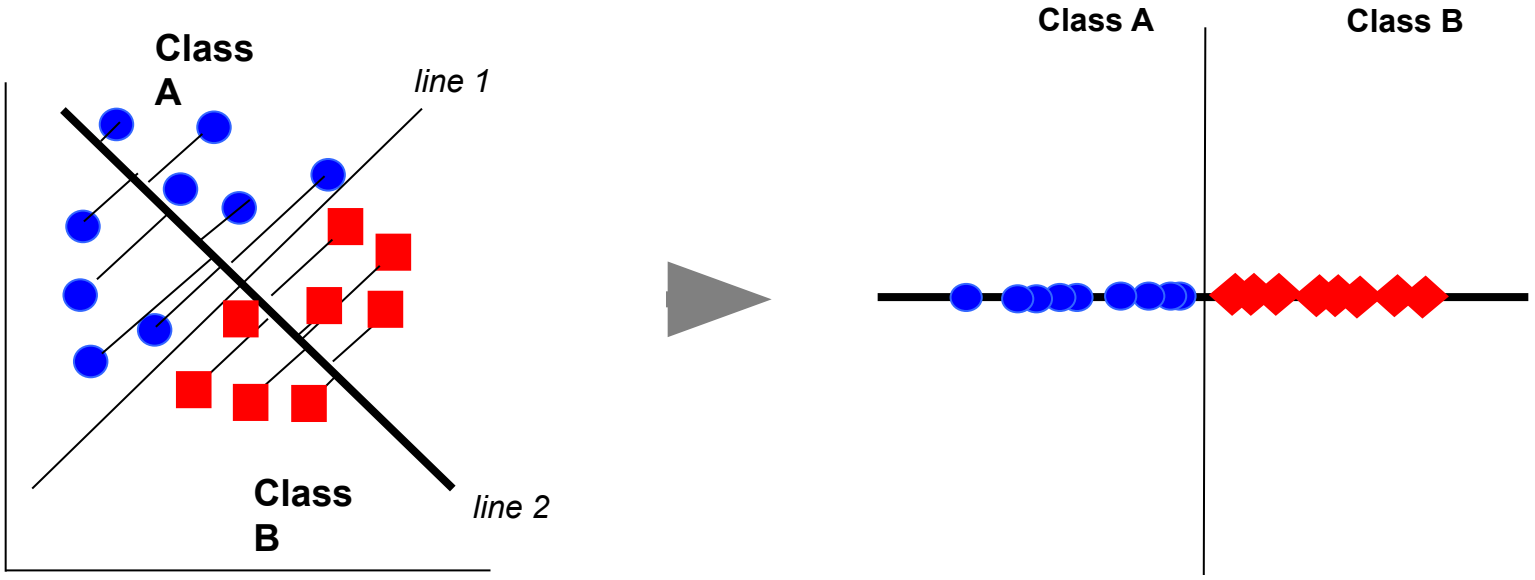


Как определить класс?

Используем линейную границу: все, что выше — первый класс, все, что ниже — второй

Как это все реализовать?

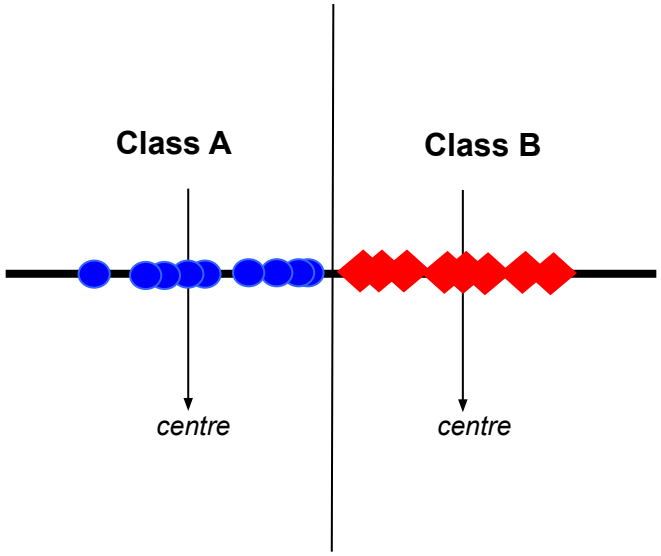
Используем проекционный подход:
объекты, с меньшей координатой — класс A, с большей — класс B



Как это все реализовать?

Находим **центроиды** — центры моделей:

Объекты, расстояние от которых до первого центра меньше, чем до второго, принадлежат классу A и наоборот



Как это все реализовать?

Нет четкого разделения между классами:

- Устанавливаем ограничение на максимальное расстояние от центра и все остальное считаем выбросами – **уменьшение ошибок второго рода**
- Устанавливаем приоритетный класс и максимальное расстояние до его центра ставим больше, чем до центра второго класса — **уменьшение ошибок первого рода**

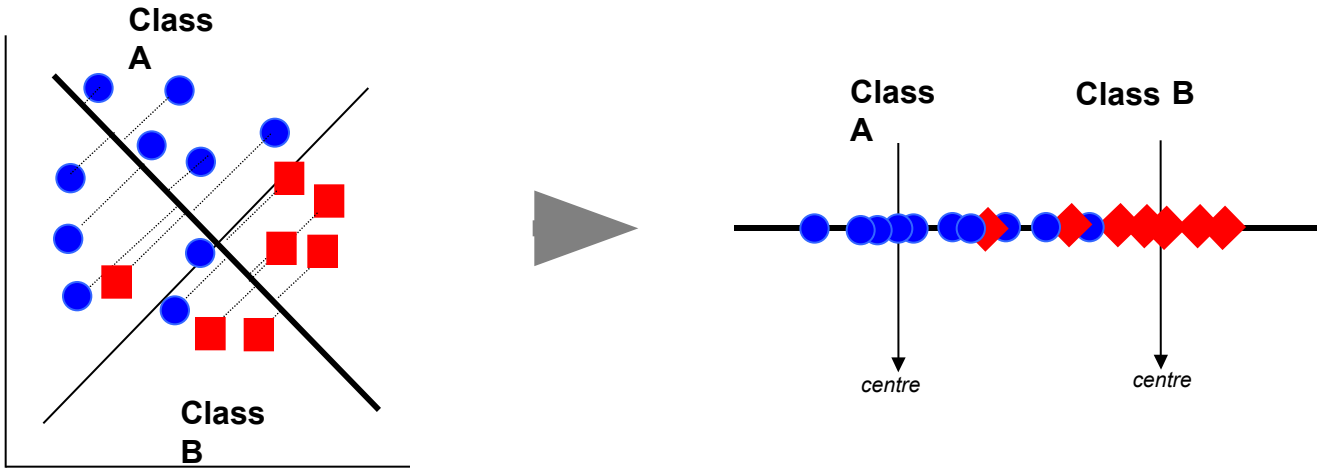


График расстояний: используем проекцию

Оси — расстояния от объекта до центров каждого из классов

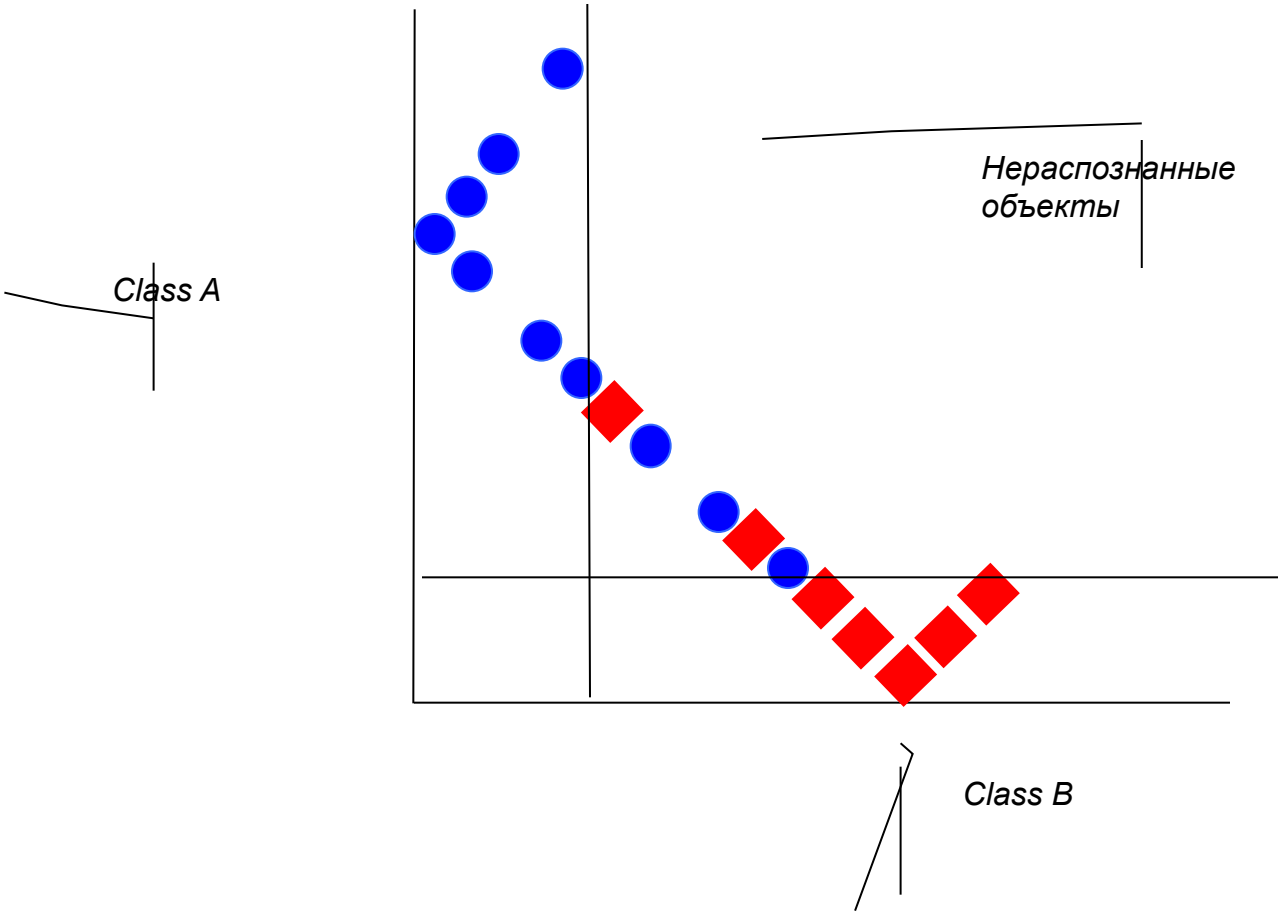
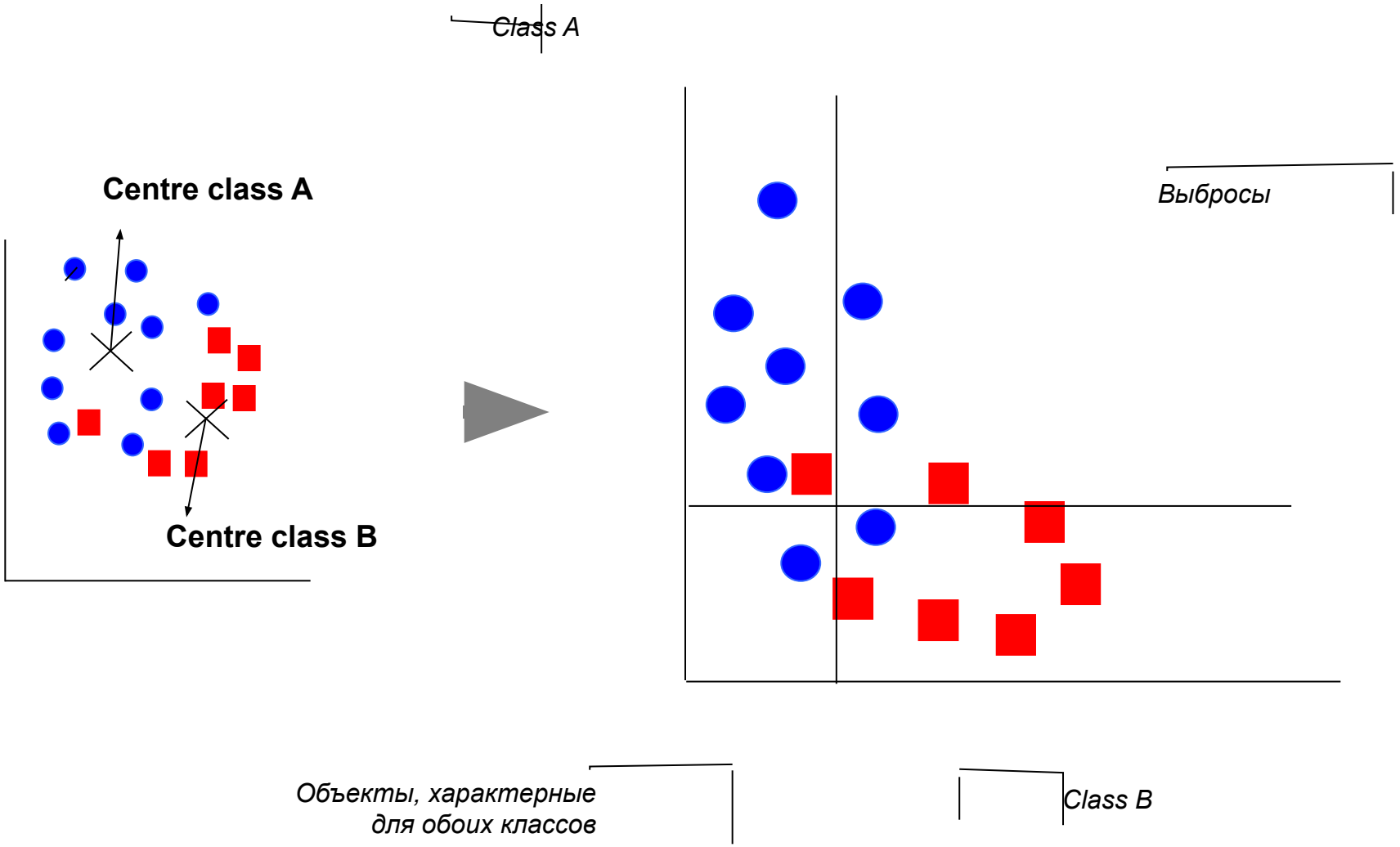


График расстояний: работаем в пространстве



Как вычислить расстояние?

Евклидово расстояние:

$$d_{kl} = (\mathbf{x_k} - \mathbf{x_l})(\mathbf{x_k} - \mathbf{x_l})'$$

Здесь k и l — номера объектов, $\mathbf{x}_k$, $\mathbf{x}_l$ — их векторы признаков

Основные характеристики

- Каждая переменная вектора признаков дает одинаковый вклад наряду с остальными — считается что они ортогональны
- Если между переменными имеется корреляция то они будут иметь непропорциональное влияние на результаты анализа

Как вычислить расстояние?

Расстояние Махаланобиса

$$d_{kl} = (\mathbf{x_k} - \mathbf{x_l})C^{-1}(\mathbf{x_k} - \mathbf{x_l})'$$

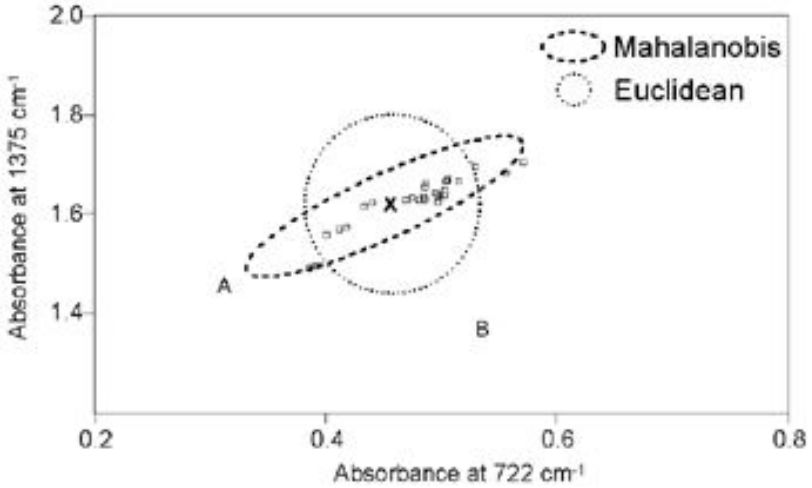
Здесь k и l — номера объектов, $\mathbf{x}_k$, $\mathbf{x}_l$ — их векторы признаков, C — ковариационная матрица признаков

Основные характеристики

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

Как вычислить расстояние?

Расстояние Махаланобиса



Альтернатива – метод ближайших соседей

- Подсчитывается число ближайших k соседей рассматриваемого объекта
- Тот класс, к которому принадлежит большинство соседей и соотносится с объектом
  - Метод ближайшего соседа

Использование исходных данных

- Вектор признаков зачастую состоит из десятков, сотен переменных, что делает невозможным визуальный анализ данных
- Матрица исходных данных содержит лишь несколько релевантных переменных и большое число нерелевантных
- Данные могут содержать шум
- Данные могут быть линейно зависимы

Выход: использовать проекционные методы!

Часть II. Методы многомерной классификации

Методы многомерной классификации

- **Unsupervised**
  - МГК
- **Supervised**
  - SIMCA
  - PLS DA
  - SVM
  - Neural networks
  - …

Набор данных: Elements

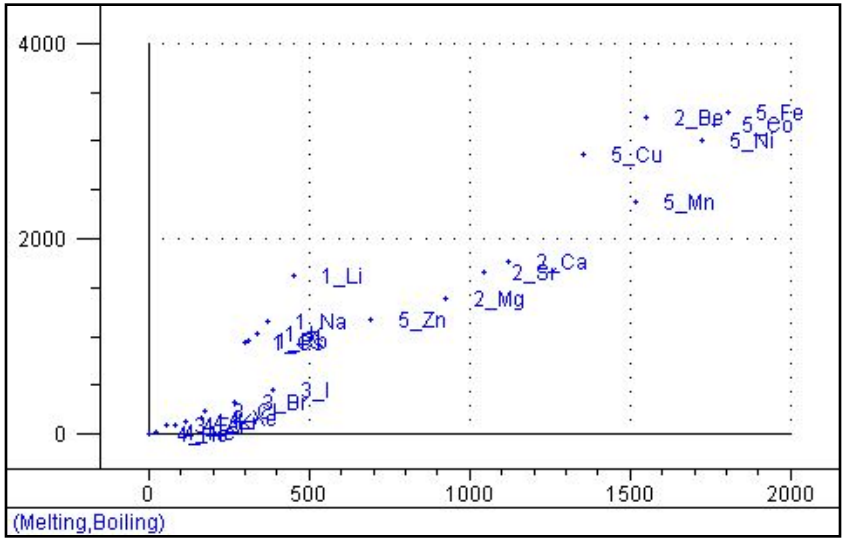
Свойства некоторых элементов таблицы Менделеева:

- **25 образцов x 5 переменных**
- **5 групп**

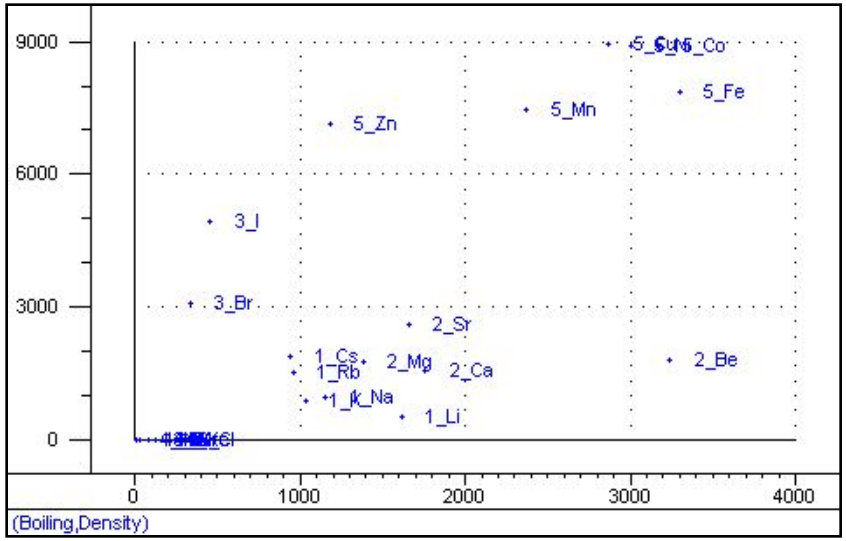
Element	G	Melting	Boiling	Density	Oxidation	Electronegativity
Li	1	453.69	1615	534	1	0.98
Na	1	371	1156	970	1	0.93
Be	2	1550	3243	1800	2	1.57
Mg	2	924	1380	1741	2	1.31
F	3	53.5	85	1.7	-1	3.98
Cl	3	172.1	238.5	3.2	-1	3.16
He	4	0.9	4.2	0.2	0	0
Ne	4	24.5	27.2	0.8	0	0
Zn	5	692.6	1180	7140	2	1.6
Co	5	1765	3170	8900	3	1.8

Предварительный анализ (2D)

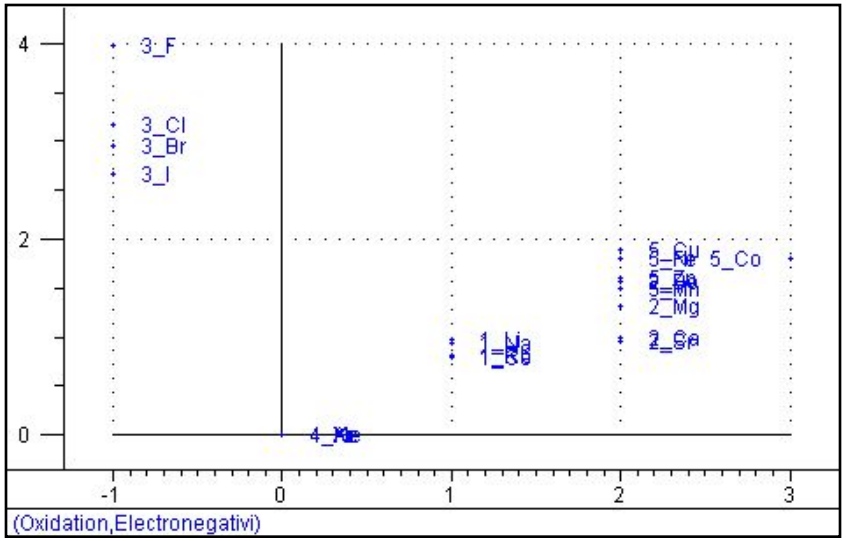
A



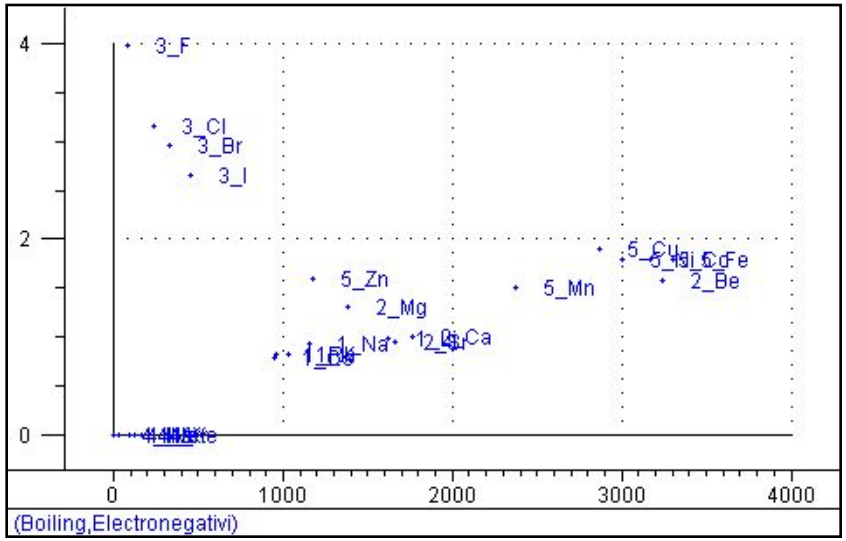
C



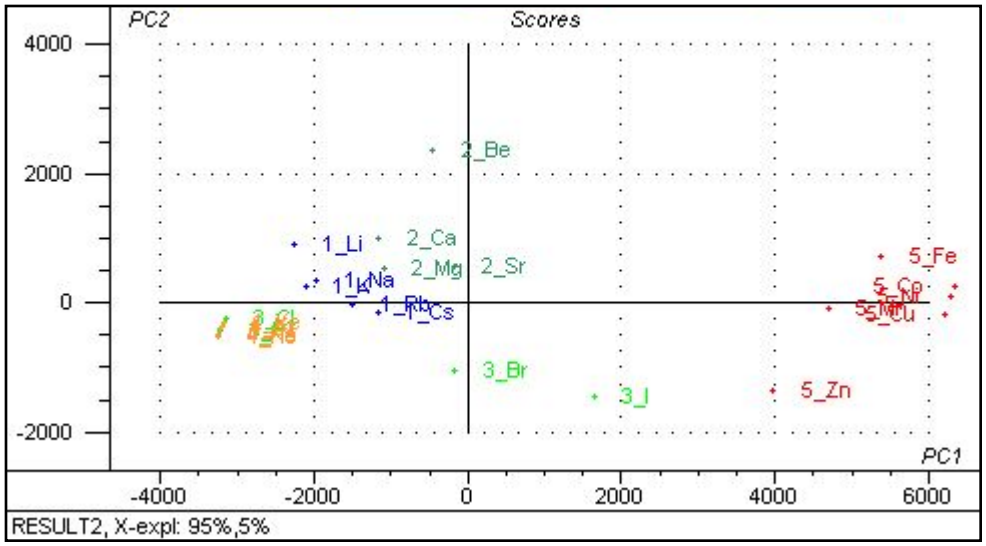
B



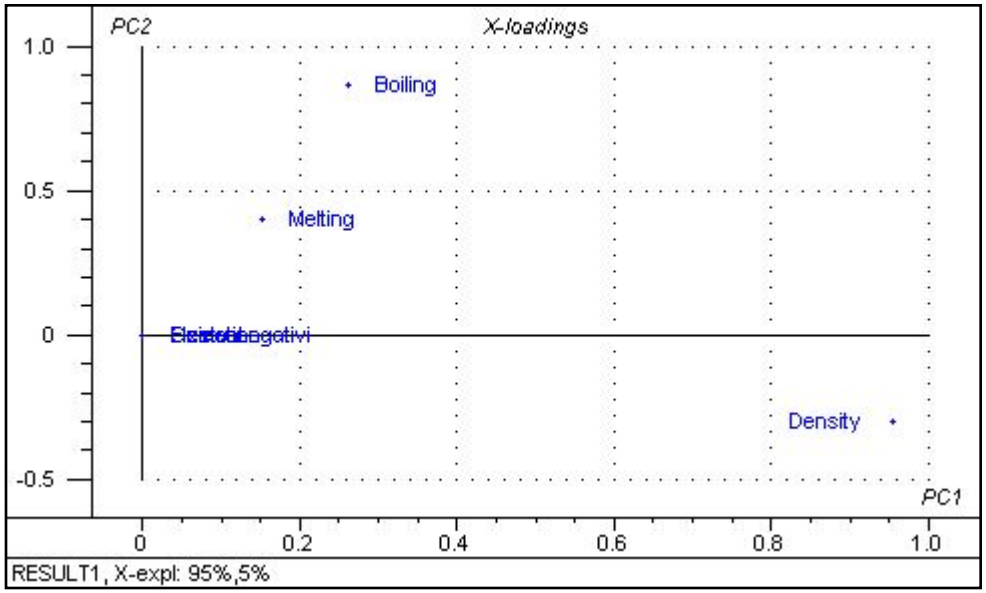
D



МГК-анализ



Счета



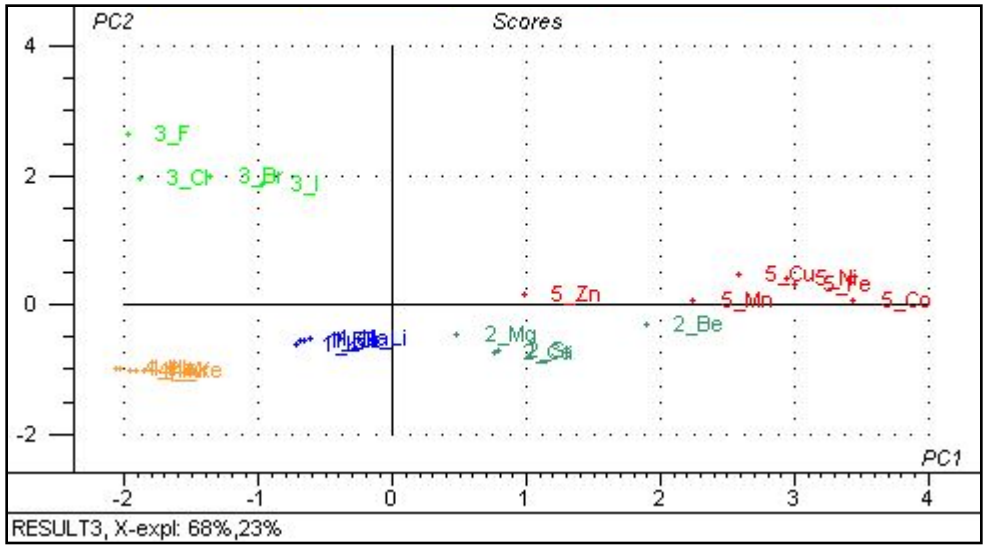
Нагрузки

МГК-анализ

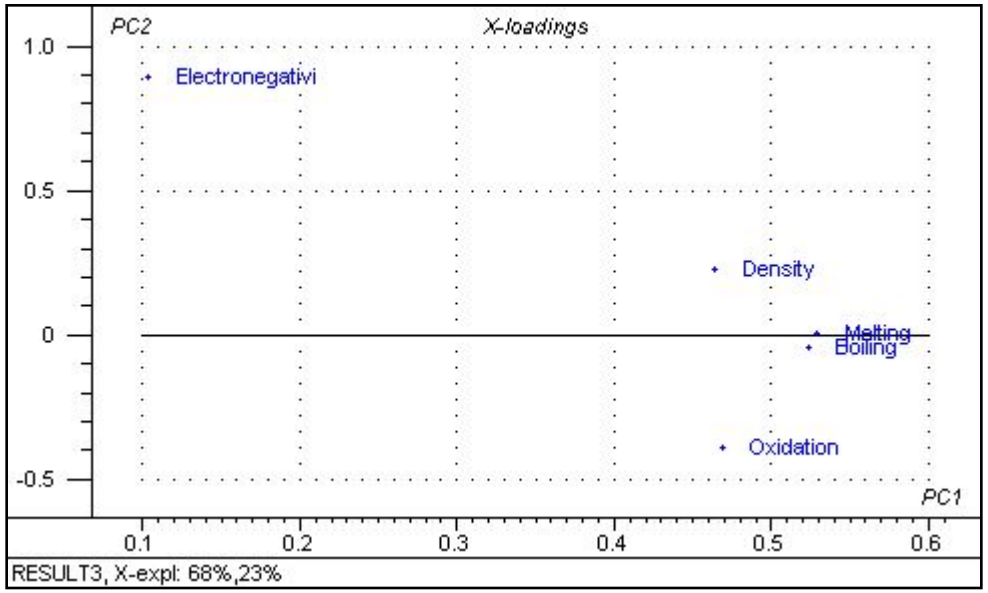
Element	G	Melting	Boiling	Density	Oxidation	Electronegativity
Li	1	453.69	1615	534	1	0.98
Na	1	371	1156	970	1	0.93
Be	2	1550	3243	1800	2	1.57
Mg	2	924				1.31
F	3	53.5				
Cl	3	172.1				
He	4					
Ne	4					
Zn	5					
Co	5					

Автошкалирование!

МГК-анализ

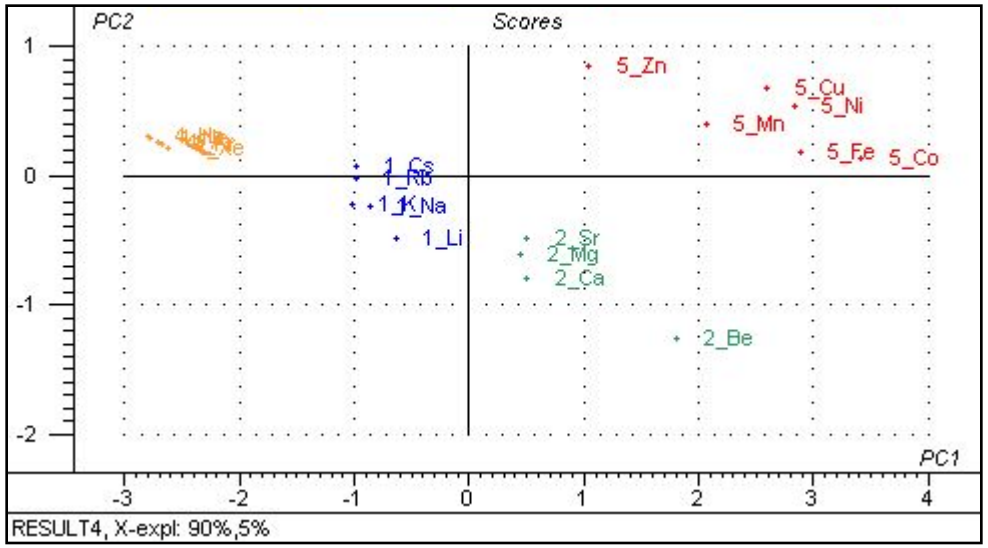


Счета

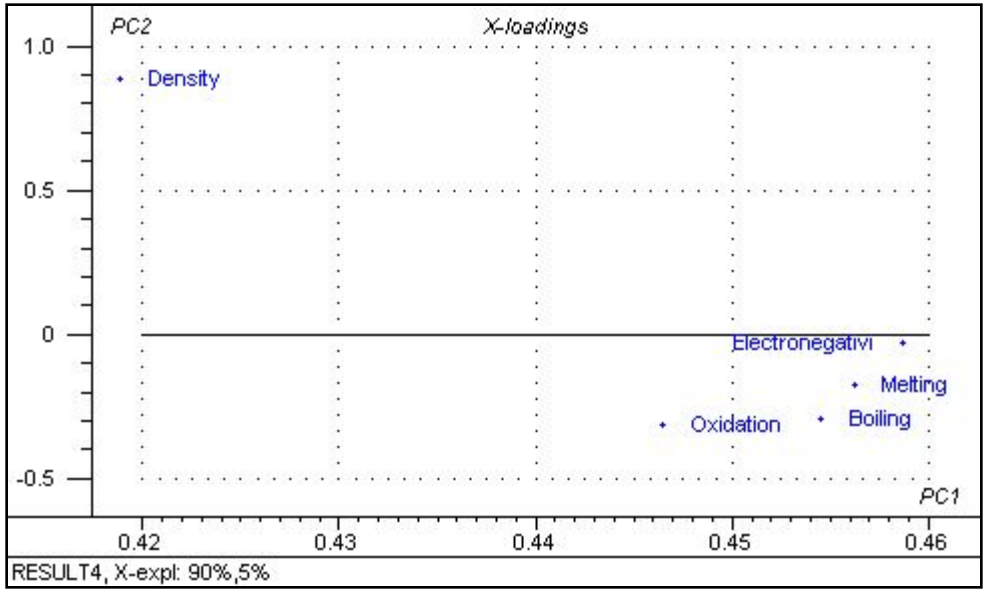


Нагрузки

МГК-анализ



Счета

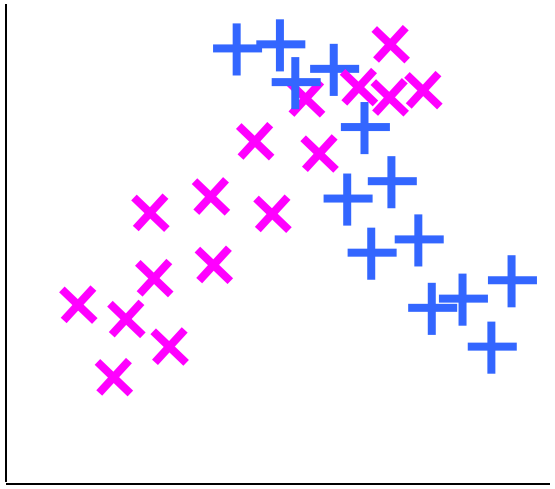


Нагрузки

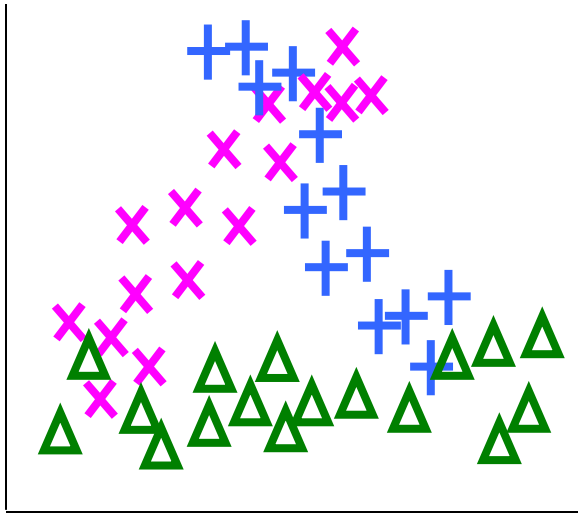
Soft **I**ndependent **M**odeling of **C**lass **A**nalogy

Предложен Svante Wold, 1970-е годы

Объект может относится одновременной к нескольким классам, что очень часто может встречаться в реальной жизни



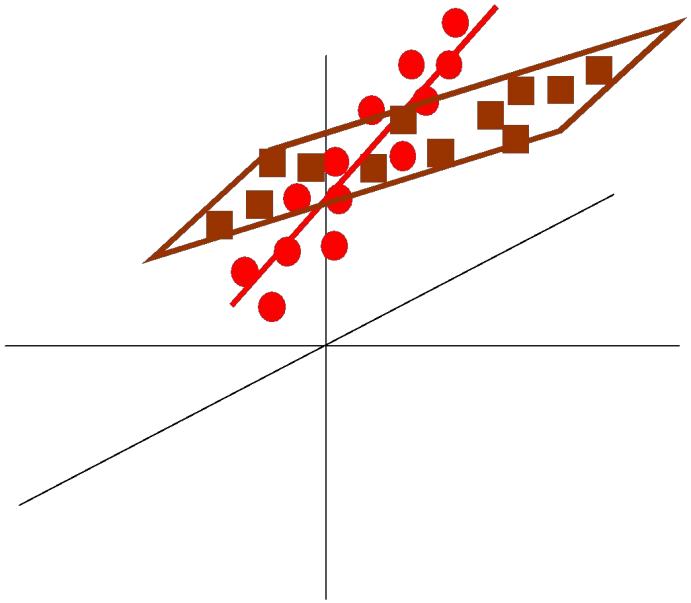
Основная идея:
моделировать каждый класс независимо от других и для каждого объекта определять принадлежит он данному класс или нет



SIMCA: основные этапы

1. Каждый класс моделируется методом главных компонент

Для каждого класса может использоваться разное число компонент, которое определяется в соответствии с методами, изложенными в курсе по МГК



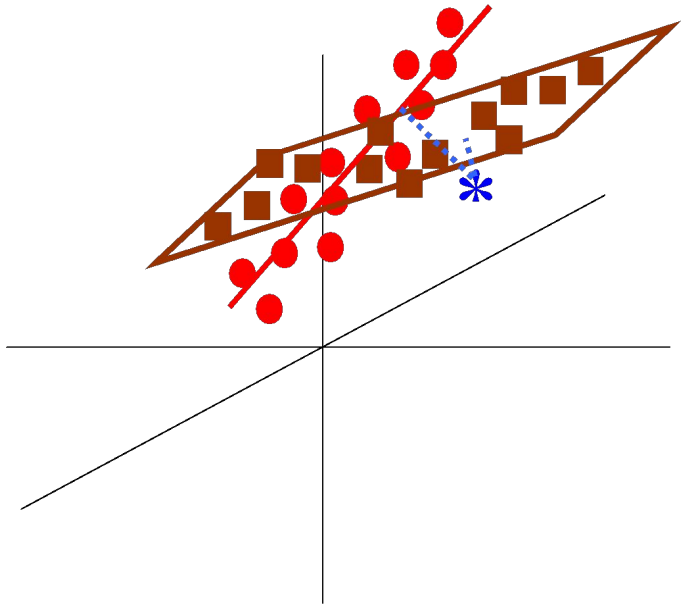
При построении обязательно необходимо проверить модель на предмет наличия выбросов и/или необходимости предобработки данных

SIMCA: основные этапы

2. Вычисляется расстояние от объекта до каждого класса

В данном случае расстояние от нового образца (звездочка) до плоскости ближе, чем до прямой

Может использоваться так же вероятностный подход



SIMCA: исходные данные – вино

3 класса, 178 образцов x 13 переменных

Тренировочный набор: 148 образцов

Проверочный набор: 30 образцов

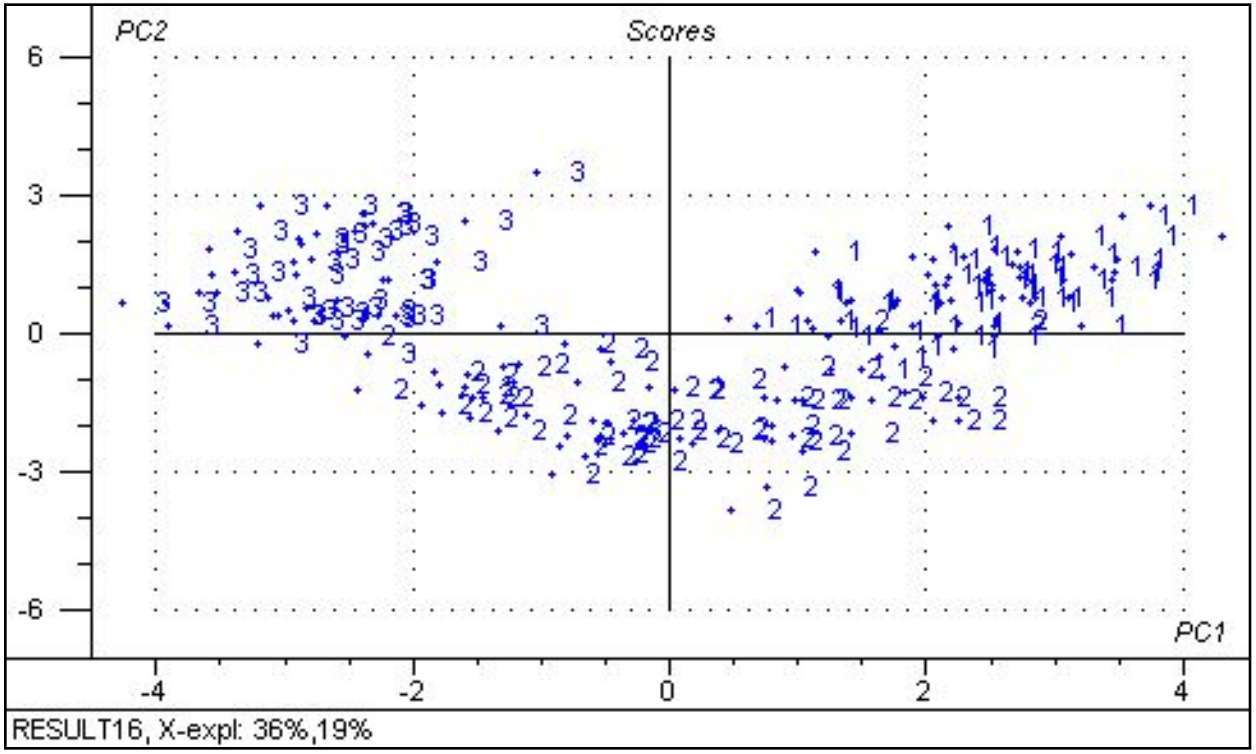
Type	Alcohol	Malic acid	Ash	Alcalinity	Magnesium	Total phenols	Flavanoids	Nonflavanoid ph	Proanthocyanins	Color intensity	Hue	OD280/OD315	Proline
1	14,23	1,71	2,43	15,60	127,00	2,80	3,06	0,28	2,29	5,64	1,04	3,92	1065,00
1	13,20	1,78	2,14	11,20	100,00	2,65	2,76	0,26	1,28	4,38	1,05	3,40	1050,00
1	13,16	2,36	2,67	18,60	101,00	2,80	3,24	0,30	2,81	5,68	1,03	3,17	1185,00
2	12,37	0,94	1,36	10,60	88,00	1,98	0,57	0,28	0,42	1,95	1,05	1,82	520,00
2	12,33	1,10	2,28	16,00	101,00	2,05	1,09	0,63	0,41	3,27	1,25	1,67	680,00
2	12,64	1,36	2,02	16,80	100,00	2,02	1,41	0,53	0,62	5,75	0,98	1,59	450,00
3	12,81	2,31	2,40	24,00	98,00	1,15	1,09	0,27	0,83	5,70	0,66	1,36	560,00
3	12,70	3,55	2,36	21,50	106,00	1,70	1,20	0,17	0,84	5,00	0,78	1,29	600,00
3	12,51	1,24	2,25	17,50	85,00	2,00	0,58	0,60	1,25	5,45	0,75	1,51	650,00

SIMCA: основные результаты

- Графики моделей
- Классификационная таблица
- Расстояние между моделями
- Модельная мощность переменных
- Дискриминационная мощность переменных
- Расстояние от образца до моделей (классов)
- Размах образца
- График Кумана

SIMCA: основные результаты

Общая МГК-модель



SIMCA: основные результаты

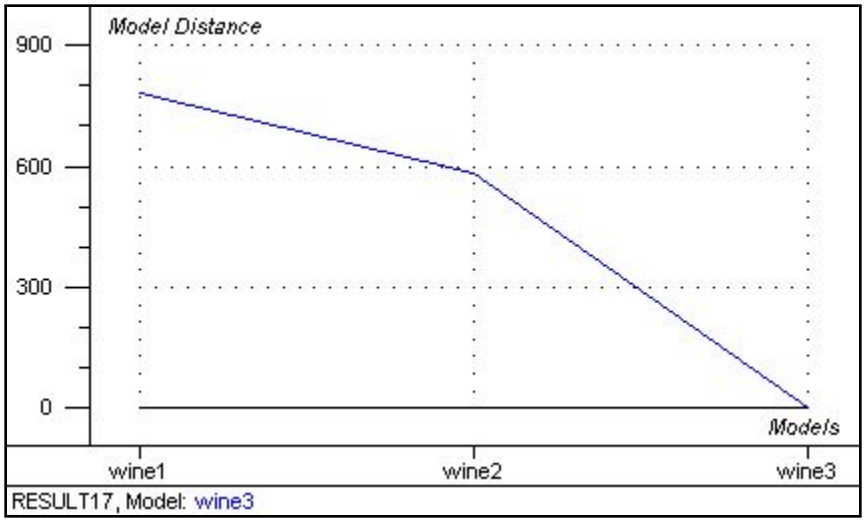
Таблица классификации

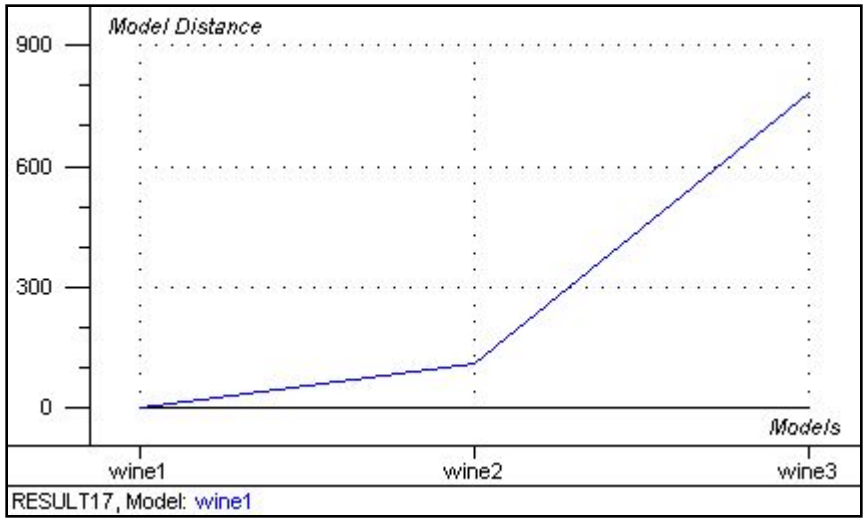
sample	wine1	wine2	wine3
1	*		
1	*		
1	*		
1	*	*	
1	*		
1	*		
1	*		
1	*		
1	*		
1	*	*	

sample	wine1	wine2	wine3	Sample	wine1	wine2	wine3
2		*		3			*
2				3			*
2		*		3			*
2		*		3			*
2		*		3			*
2	*	*		3			*
2				3			*
2		*		3			*
2				3			*
2		*		3			*

Расстояние между моделями

Все объекты одной модели соотносятся с ней же, затем соотносятся с другой моделью, затем результат сравнивается с единицей. Чем больше данный параметр тем более хорошо различаются модели





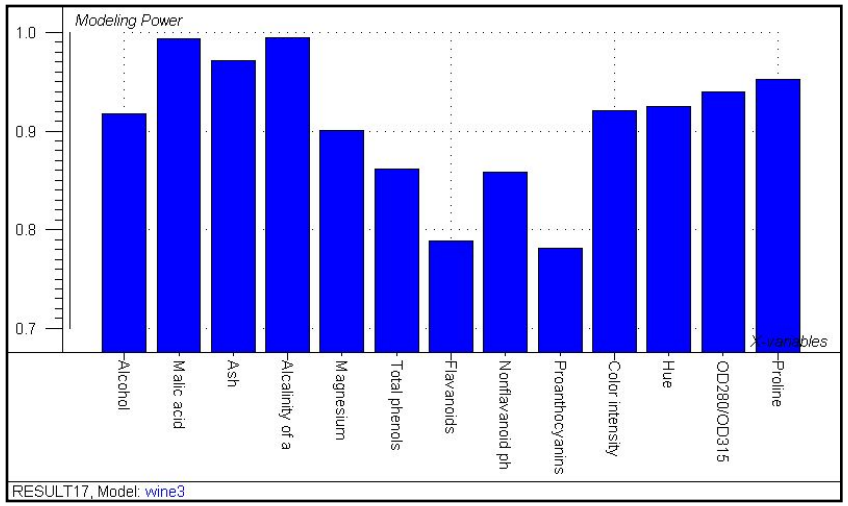
SIMCA: основные результаты

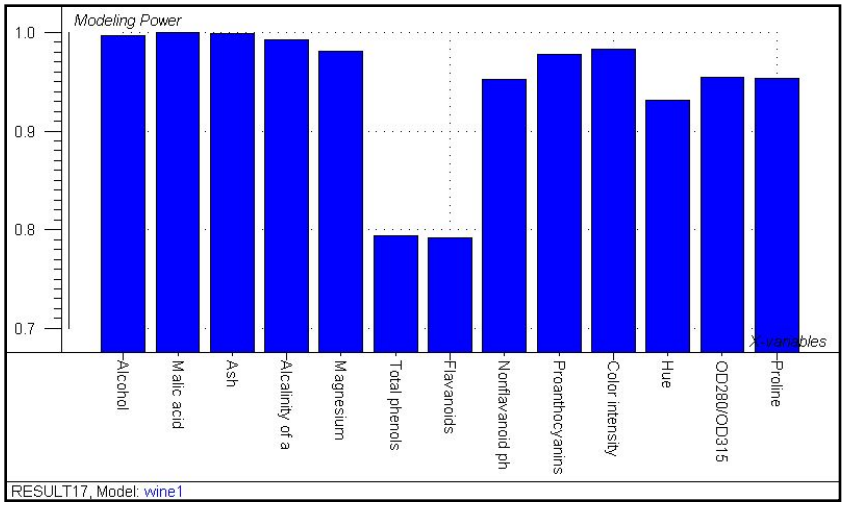
Модельная мощность переменной

Данный параметр показывает насколько сильное влияние оказывает данная переменная на моделирование данного класса

Рассчитывается по формуле $M_j = 1 - s_{jresid} / s_{jraw}$

Разброс значений: 1 – сильное влияние … 0 – влияния нет





SIMCA: основные результаты

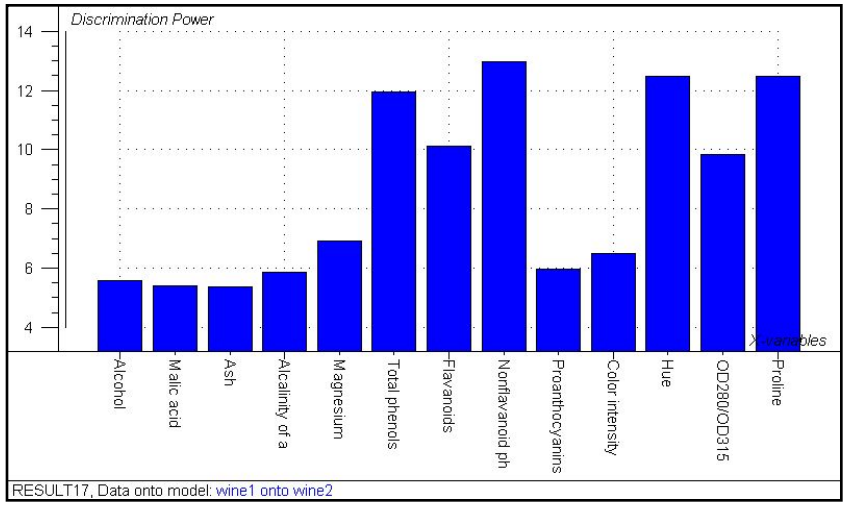
Дискриминационная мощность переменной

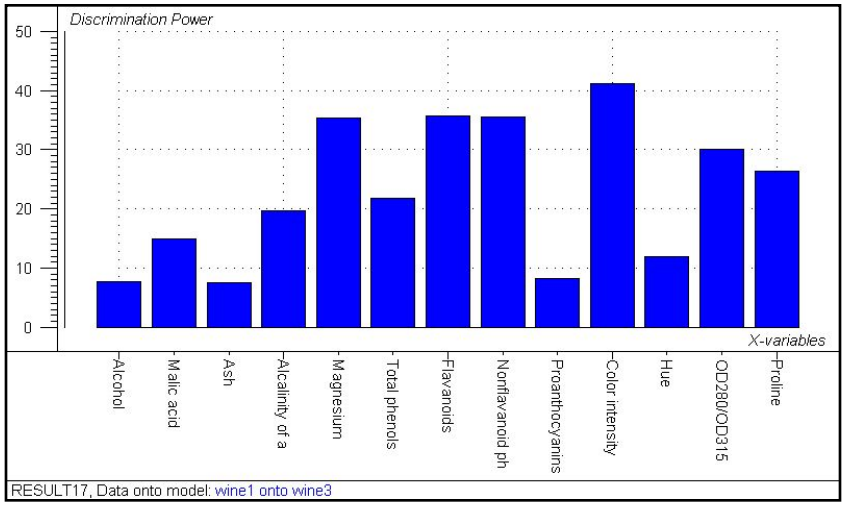
Данный параметр показывает способность переменной разделять два класса (способность переменной моделировать класс не влечет за собой автоматом способность разделять)

Рассчитывается по формуле

$$D_j = \sqrt{\frac{\text{class}A\ \text{model}B\ s^2_{jresid} + \text{class}B\ \text{model}A\ s^2_{jresid}}{\text{class}A\ \text{model}A\ s^2_{jresid} + \text{class}B\ \text{model}B\ s^2_{jresid}}}$$

Разброс значений: чем больше значение, тем больше способность к дискриминации





SIMCA: основные результаты

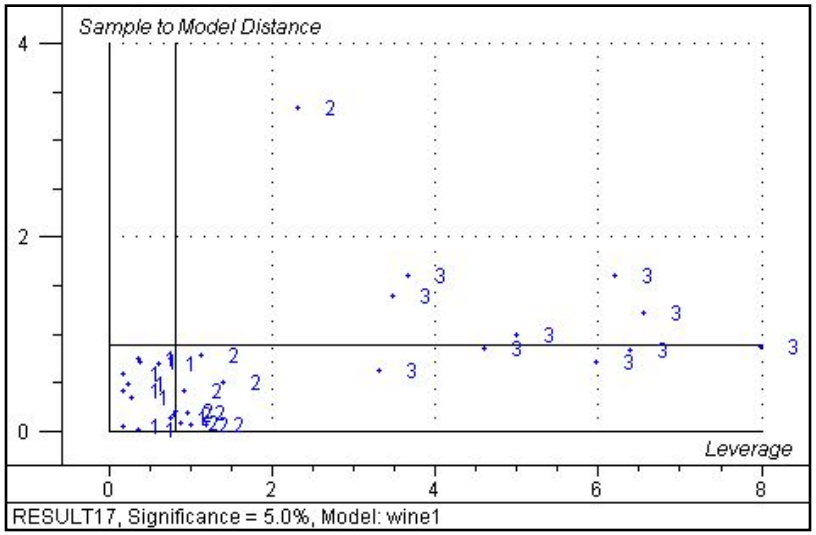
Расстояние от образца до модели

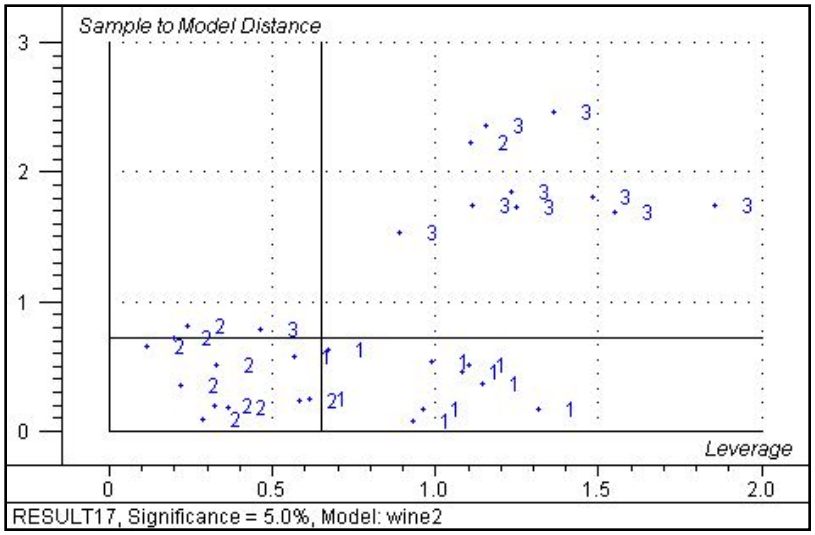
Рассматривается насколько далеко образец находится от модели данного класса (используется отношение дистанции до центроида и вариация)

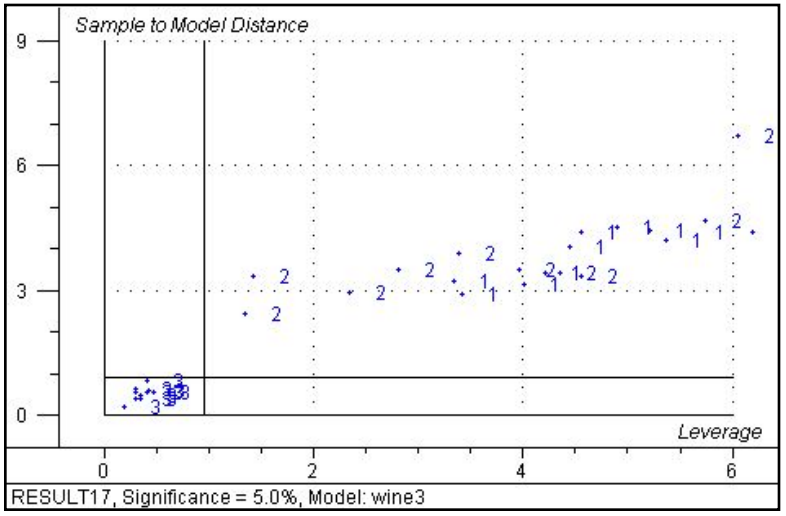
Размах образца

Рассматривается насколько проекция образца на данную модель далека от ее центроида (т.е. насколько он отличается от других образцов данной модели)

Зависимость расстояния от размаха

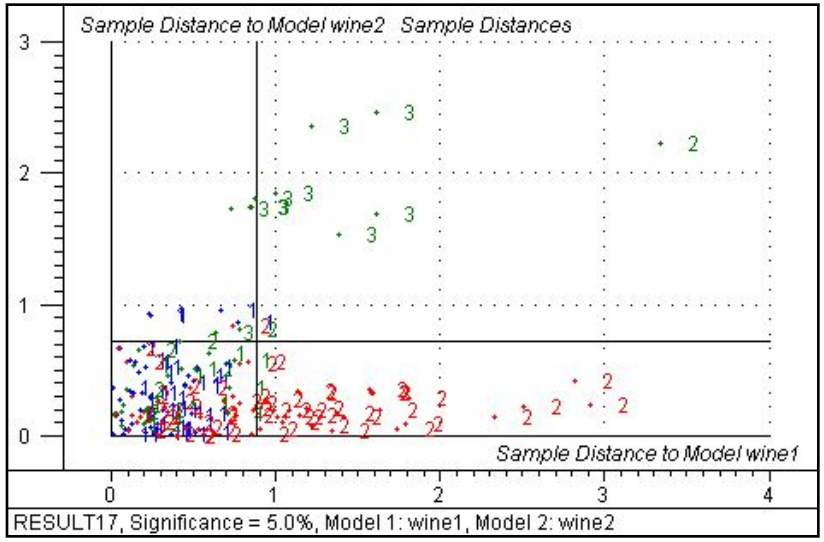


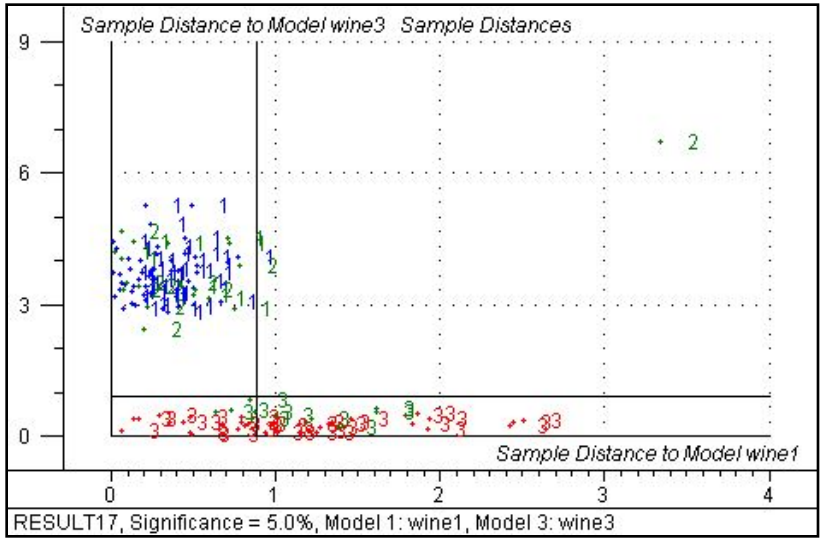


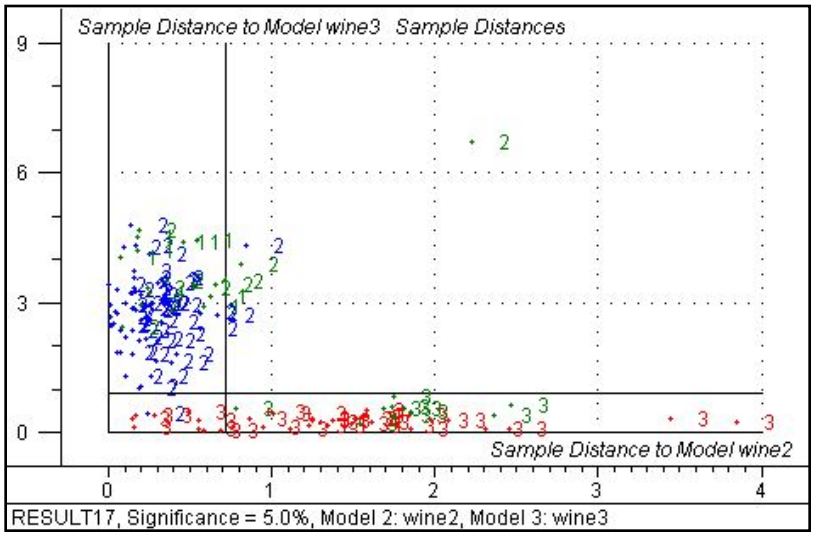


SIMCA: основные результаты

График Кумана







Резюме

Классификация шаг за шагом

- **Предварительная обработка данных**

  Большинство проекционных методов весьма чувствительны к предварительной обработке данных. Поэтому, если нет априорной информации, какие переменные имеют более сильное влияние, а какие – нет, необходимо центрировать данные и шкалировать

- **Предварительный анализ данных**

  Второй этап представляет собой построение МГК и/или ПЛС модели исходных данных и предварительный обзор результатов на предмет наличия групп, выбросов и прочих аномалий

Классификация шаг за шагом

- **Раздельное построение моделей для классов**

  Для классов, которые были выявлены на втором этапе, строятся раздельно модели для лучшей кластеризации и анализа поведения объектов внутри класса. Очень важно на этом этапе использовать кросс-валидацию

- **Интерпретация моделей**

  На данном этапе полученные модели анализируются и интерпретируются на предмет выявления наиболее значимых для них переменных

Классификация шаг за шагом

- **Классификация объектов**

  На данном этапе объекты или результаты наблюдений проецируются на полученные классы. Для определения, насколько хорошо они соответствуют модели, для каждого случая вычисляется расстояние от объекта до нее. Здесь нужно учесть, что могут быть как объекты, описывающие несколькими моделями, так и те, которые не удовлетворяют ни одной из них

- **Классификация новых образцов**

  Для достоверной оценки способности классификации необходимо использовать независимый, тестовый набор данных, если есть такая возможность