# Статистический анализ данных. Спецкурс.

## Лекция 6. Методы многомерной статистики

Ботанический сад-институт ДВО РАН

Кислов Д.Е.

12 декабря 2016 г.

# Структура лекции

- Принцип наименьших квадратов;

# Структура лекции

- Принцип наименьших квадратов;
- Метод главных компонент;

# Структура лекции

- Принцип наименьших квадратов;
- Метод главных компонент;
- Линейный дискриминантный анализ;

# Структура лекции

- Принцип наименьших квадратов;
- Метод главных компонент;
- Линейный дискриминантный анализ;
- Классификация по прецедентам;

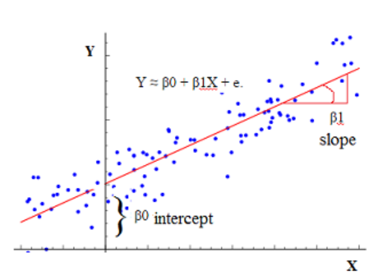
# Структура лекции

- Принцип наименьших квадратов;
- Метод главных компонент;
- Линейный дискриминантный анализ;
- Классификация по прецедентам;
- Оценка качества классификации, отбор признаков.

Предложен К.Ф. Гауссом (1795) для решения уравнений:

$$a_1x + b_1y = c_1$$
$$a_2x + b_2y = c_2$$
$$a_3x + b_3y = c_3$$

Задача регрессии

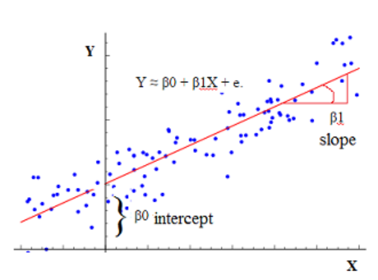


# Принцип наименьших квадратов

Предложен К.Ф. Гауссом (1795) для решения уравнений:

$$a_1 x + b_1 y = c_1 + \varepsilon_1$$
$$a_2 x + b_2 y = c_2 + \varepsilon_2$$
$$a_3 x + b_3 y = c_3 + \varepsilon_3$$
$$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 \to \min$$

Задача регрессии



# Принцип наименьших квадратов

## Применение

- Решение переопределенных/недоопределенных систем уравнений;

# Принцип наименьших квадратов

## Применение

- Решение переопределенных/недоопределенных систем уравнений;
- Статистическая оценка параметров;

# Принцип наименьших квадратов

## Применение

- Решение переопределенных/недоопределенных систем уравнений;
- Статистическая оценка параметров;
- Решение задач снижения размерности;

# Принцип наименьших квадратов

## Применение

- Решение переопределенных/недоопределенных систем уравнений;
- Статистическая оценка параметров;
- Решение задач снижения размерности;
- Построение регрессионных моделей;

# Принцип наименьших квадратов

## Применение

- Решение переопределенных/недоопределенных систем уравнений;
- Статистическая оценка параметров;
- Решение задач снижения размерности;
- Построение регрессионных моделей;
- ... в общем случае — любые другие задачи, связанные с минимизацией ошибок/погрешностей.

### Задача регрессии

Имеется набор измерений $y_j$ ($j = \overline{1, \ldots, N}$, предположительно зависящий от параметров $x_{ij}$ ($i = \overline{1, \ldots, M}$). Необходимо построить какую-либо модель этой зависимости, исходя из набора эмпирических данных.

# Принцип наименьших квадратов

**Задача регрессии**

Имеется набор измерений $y_j$ $(j = \overline{1, \ldots, N}$, предположительно зависящий от параметров $x_{ij}$ $(i = \overline{1, \ldots, M})$. Необходимо построить какую-либо модель этой зависимости, исходя из набора эмпирических данных.

**Положим,что зависимость между $y_i$ и $x_{ij}$ линейная**

# Принцип наименьших квадратов

## Задача регрессии

Имеется набор измерений $y_j$ ($j = \overline{1, \ldots, N}$, предположительно зависящий от параметров $x_{ij}$ ($i = \overline{1, \ldots, M}$). Необходимо построить какую-либо модель этой зависимости, исходя из набора эмпирических данных.

## Положим,что зависимость между $y_i$ и $x_{ij}$ линейная

$$
\begin{aligned}
y_1 &= a_0 + a_1 \cdot x_{11} + a_2 \cdot x_{21} + \ldots + a_M \cdot x_{M1} \\
y_2 &= a_0 + a_1 \cdot x_{12} + a_2 \cdot x_{22} + \ldots + a_M \cdot x_{M2} \\
&\vdots \\
y_N &= a_0 + a_1 \cdot x_{1N} + a_2 \cdot x_{2N} + \ldots + a_M \cdot x_{MN}
\end{aligned}
$$

# Принцип наименьших квадратов

## Задача регрессии

Имеется набор измерений $y_j$ $(j = \overline{1, \ldots, N}$, предположительно зависящий от параметров $x_{ij}$ $(i = \overline{1, \ldots, M})$. Необходимо построить какую-либо модель этой зависимости, исходя из набора эмпирических данных.

## Положим,что зависимость между $y_i$ и $x_{ij}$ линейная

$$
\begin{aligned}
&y_1 = a_0 + a_1 \cdot x_{11} + a_2 \cdot x_{21} + \ldots + a_M \cdot x_{M1} + \varepsilon_1 \\
&y_2 = a_0 + a_1 \cdot x_{12} + a_2 \cdot x_{22} + \ldots + a_M \cdot x_{M2} + \varepsilon_2 \\
&\vdots \\
&y_N = a_0 + a_1 \cdot x_{1N} + a_2 \cdot x_{2N} + \ldots + a_M \cdot x_{MN} + \varepsilon_N
\end{aligned}
$$

$$\sum_i \varepsilon_i^2 \to \min$$

# Принцип наименьших квадратов

## Решение

$$Y = X \cdot a + \varepsilon, a = (a_0, a_1, \ldots, a_M),$$
$$a = (X^T X)^{-1} X^T Y, \text{ или } a = X^+ Y$$

## Нелинейный МНК

$$y_j = F(a_i, x_{ij}) + \varepsilon_j$$

# Принцип наименьших квадратов

## Решение

$$Y = X \cdot a + \varepsilon, a = (a_0, a_1, \dots, a_M),$$
$$a = (X^T X)^{-1} X^T Y, \text{или } a = X^+ Y$$

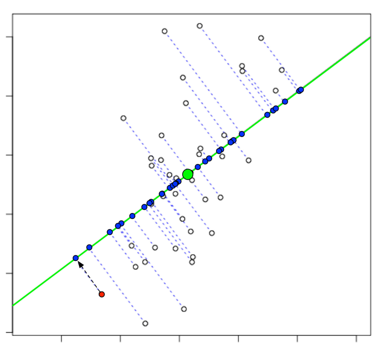
## Нелинейный МНК

$$y_j = F(a_i, x_{ij}) + \varepsilon_j$$

## Решение

Как правило – численные методы: нелинейные методы оптимизации, проблемно-ориентированные подходы.
А также ... Существуют частные случаи – легко приводимые к линейной задаче.

## Формулировка задачи



Отыскать такую ось – линейную комбинацию исходных координат, сумма квадратов расстояний от данных до которой минимальна; (метод предложен К. Пирсоном);

# Метод главных компонент

## Интерпретации

- Теория вероятностей – диагонализация ковариационной матрицы (преобразование Хотеллинга);

## Интерпретации

- Теория вероятностей – диагонализация ковариационной матрицы (преобразование Хотеллинга);
- Статистика – максимизация вариации (дисперсии) проекций данных на прямую;

# Метод главных компонент

## Интерпретации

- Теория вероятностей – диагонализация ковариационной матрицы (преобразование Хотеллинга);
- Статистика – максимизация вариации (дисперсии) проекций данных на прямую;
- Механика – отыскание главных осей инерции;

## Вычислительные аспекты

Пусть $X$ — исходная матрица данных (имеющих нулевое среднее); $S = \frac{1}{N-1} X^T X$ – выборочная ковариационная матрица. Тогда решение задачи отыскания главных компонент определяется ее спектральным разложением

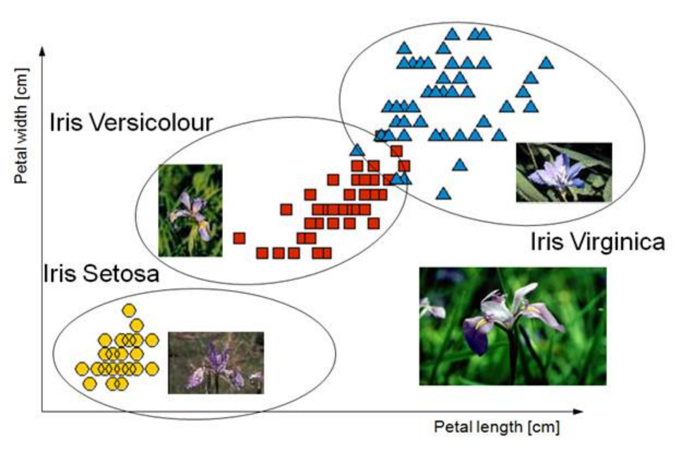
$$S = U^T \cdot \Sigma U,$$

где $U$ – ортогональная матрица; $\Sigma = (\lambda_1, \ldots, \lambda_2)$.

$$tr(\Sigma) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_m^2$$

$$\mu_k = \frac{\lambda_1 + \ldots + \lambda_k}{\sum\limits_i \lambda_i}$$

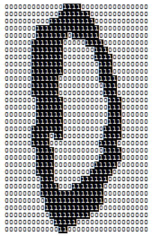
## Формулировка задачи на примере классификации ирисов



# Классификация по прецедентам

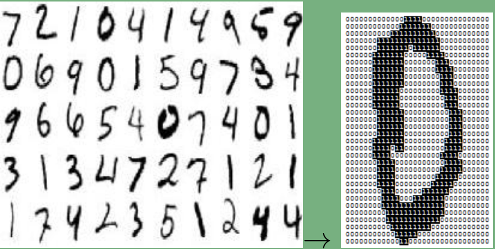
## Формулировка задачи на примере классификации рукописных цифр

$\rightarrow$



# Классификация по прецедентам

## Формулировка задачи на примере классификации рукописных цифр



## !

Реальные задачи распознавания образов могут иметь очень большие размерности.

# Основные этапы решения задач классификации по прецедентам

- Подготовка данных: приведение к единому масштабу (при необходимости), центрирование, удаление выбросов и т.п.

# Основные этапы решения задач классификации по прецедентам

- Подготовка данных: приведение к единому масштабу (при необходимости), центрирование, удаление выбросов и т.п.
- Отбор и формирование переменных (features engeneering, features extraction, features selection );

# Основные этапы решения задач классификации по прецедентам

- Подготовка данных: приведение к единому масштабу (при необходимости), центрирование, удаление выбросов и т.п.
- Отбор и формирование переменных (features engeneering, features extraction, features selection );
- Выбор классификатора и подстройка его параметров;

# Основные этапы решения задач классификации по прецедентам

- Подготовка данных: приведение к единому масштабу (при необходимости), центрирование, удаление выбросов и т.п.
- Отбор и формирование переменных (features engeneering, features extraction, features selection );
- Выбор классификатора и подстройка его параметров;
- Тестирование классификатора (при необходимости – повторение операций с п. 2).

# Методы решения задач классификации

- Наивный Байесовский классификатор;

# Методы решения задач классификации

- Наивный Байесовский классификатор;
- Метод k-ближайших соседей;

# Методы решения задач классификации

- Наивный Байесовский классификатор;
- Метод k-ближайших соседей;
- Линейная и квадратичная дискриминация;

# Методы решения задач классификации

- Наивный Байесовский классификатор;
- Метод k-ближайших соседей;
- Линейная и квадратичная дискриминация;
- Деревья решений;

# Методы решения задач классификации

- Наивный Байесовский классификатор;
- Метод k-ближайших соседей;
- Линейная и квадратичная дискриминация;
- Деревья решений;
- Машины опорных векторов;

# Методы решения задач классификации

- Наивный Байесовский классификатор;
- Метод k-ближайших соседей;
- Линейная и квадратичная дискриминация;
- Деревья решений;
- Машины опорных векторов;
- Нейросетевые классификаторы;

# Методы решения задач классификации

- Наивный Байесовский классификатор;
- Метод k-ближайших соседей;
- Линейная и квадратичная дискриминация;
- Деревья решений;
- Машины опорных векторов;
- Нейросетевые классификаторы;
- Комбинирование классификаторов;