# Понижение размерности и отбор признаков

Данил Лыков

ФПФЭ МФТИ

Москва, 2018

# План

- **Линейная алгебра**
  - Вектора
  - Базис
  - Преобразование, собственные вектора
- **Principal Component Ananlysis**
  - Постановка задачи
  - Реализация
- **Отбор признаков**
- Метрики, кросс-валидация

# Вектор

Абстрактный объект и операции для которых:

# Вектор

Абстрактный объект и операции для которых:

- $\alpha(a + b) = \alpha a + \alpha b$
- $(\alpha + \beta)a = \alpha a + \beta a$

# Вектор

Абстрактный объект и операции для которых:

- $\alpha(a + b) = \alpha a + \alpha b$
- $(\alpha + \beta)a = \alpha a + \beta a$
- $a + b = b + a$
- $\alpha(\beta a) = (\alpha\beta)a$
- $a + (b + c) = (b + a) + c$

# Вектор

Абстрактный объект и операции для которых:

- $\alpha(a+b) = \alpha a + \alpha b$
- $(\alpha+\beta)a = \alpha a + \beta a$
- $a + b = b + a$
- $\alpha(\beta a) = (\alpha\beta)a$
- $a + (b + c) = (b + a) + c$
- $o + a = a$
- $a + \overline{a} = o$
- $a = 1a$

# Например

$$\alpha(\beta a) \neq (\alpha\beta)a$$


# Например 2

$$\alpha(a+b) \neq \alpha a + \alpha b$$

не существует отрицательного элемента

```
>>> [1,2]+[12,42]
[1, 2, 12, 42]
>>> 2*[1,2]+2*[12,42]
[1, 2, 1, 2, 12, 42, 12, 42]
>>> 2*([1,2]+[12,42])
[1, 2, 12, 42, 1, 2, 12, 42]
```

# Базис

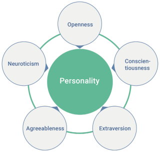
Множество $E = \{e_i\}$ для которого:
любой элемент представим в виде суммы элементов из множества $E$ с коэфициентами.
Дают ноль только с нулевыми коэффициентами

$$a = \sum_{n=1}^{d} \alpha_n e_n$$

Количество базисных векторов – размерность пространства

| Типы по Гиппократу-Галену | Черты темперамента | ICD-10 диагнозы | ICD-10 коды † |
|---|---|---|---|
| Холерик | Импульсивность, агрессия | Импульсивное расстройство личности | F60.30 |
| Флегматик | Социально отстраненный, погруженный в себя | Шизоидное расстройство личности | F60.1 |
| Меланхолик | Грустный, боязливый, подавленный, слабый | Тревожное расстройство личности | F60.6 |
| Сангвиник | Подвижный, социальный, уверенный в себе | Гипомания | F30.0 |

Cambridge Analytica




- MBTI, OCEAN

Хотим закодировать цвет - RGB (3-х мерный вектор)

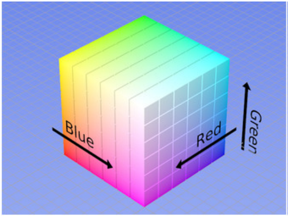
Физически - длина волны, один параметр.

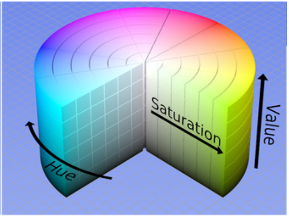
Нужно как-то преобразовать RGB

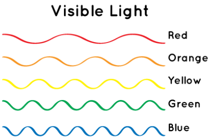
Хотим закодировать цвет - RGB (3-х мерный вектор)

Физически - длина волны, один параметр.

Нужно как-то преобразовать RGB



# Скалярное произведение

Любая функция $f : V^2 \mapsto \mathbb{R}$ $(f(a, b) = \alpha)$ для которой:

# Скалярное произведение

Любая функция $f : V^2 \mapsto \mathbb{R}$ $(f(a, b) = \alpha)$
для которой:
$f(\alpha a + \beta b, c) = \alpha f(a, c) + \beta(b, c)$
$f(a, b) = \overline{f(b, a)}$
$f(a, a) \equiv \langle a, a \rangle \geq 0$

$$a + b = \sum e\langle e, a\rangle + \sum e\langle e, b\rangle = \sum e\langle e, a + b\rangle$$

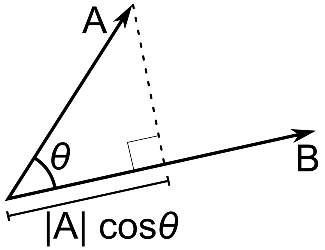
Например: $a_1b_1 + a_2b_2 + a_3b_3 + \ldots = \sum_{i=1}^{d} a_i b_i$

# Principal Component Ananlysis

Хотим выбарть другой базис, меньшей размерности, $dim(X) = D, dim(X') = d$

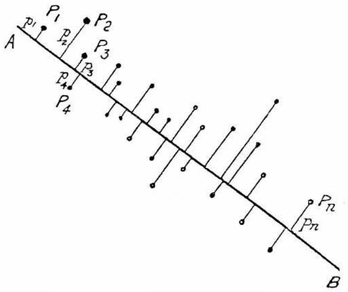
Можно выбрать $\{e_i\}_{i=1}^{d}$ произвольно,затем оптимизировать

Проекции: $p_i = e_i X$ ($x_j$ – столбцы, $p_i$ – строчки)



Цель – найти такие вектора, проекции $X$ на которые ближе всего исходным данным

Хорошо когда расстояние до вектора меньше – больше проекция.



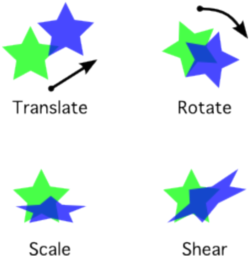
# PCA, Реализация

При условии что выборка центрирована,
дисперсия: $p_i p_i^T = \langle e_i X, e_i X \rangle = e_i X X^T e_i^T = \lambda_i$
Требование: $\langle e_i, e_i \rangle = 1, E^T E = I$
$e_i X X^T e_i^T = \lambda_i$
Перепишем в немного другом виде
$X X^T e_i^T = \lambda_i e_i^T$
$X X^T$ – матрица ковариаций

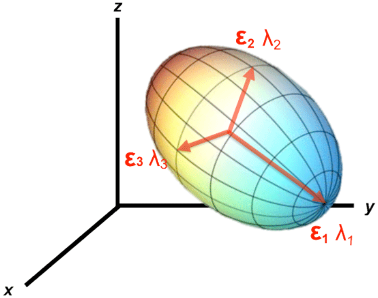
# PCA, Реализация

$$XX^Te_i^T = \lambda_i e_i^T$$

Каждой матрице соответствует какое-то преобразование пространства (*визуализация*)



# Собственные вектора $Wa = \lambda a$ (*визуализация*)



# PCA, Реализация

Как найти?
$X = UDV^T$ Singular Value Decomposition
$U, V$ – поворот или отражение
$D$ – все элементы кроме диагональных – нули

# PCA, Реализация

Как найти?

$X = UDV^T$ Singular Value Decomposition

$U, V$ – поворот или отражение

$D$ – все элементы кроме диагональных – нули

$U$ – собственные вектора $X^T X$

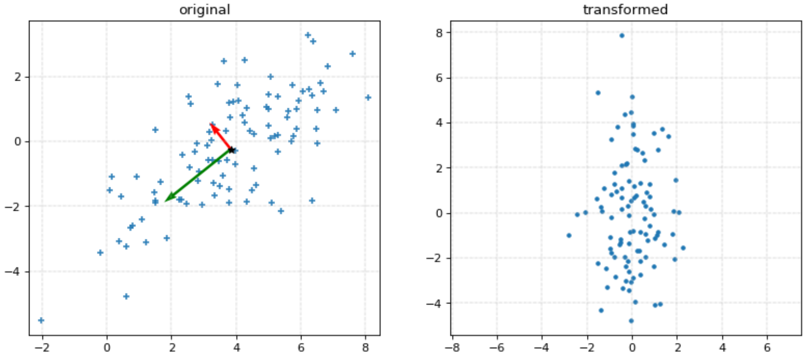
$V$ – собственные вектора $XX^T$

$D = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..)$ – диагональная матрица из корней собственных значений

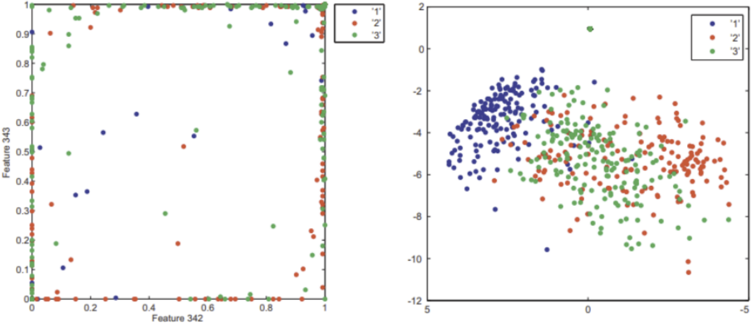
# PCA, код

```
def pca(X,num_components):
    X=np.array([l-np.mean(l) for l in X]) # Вычитаем среднее
    U,s,V=np.linalg.svd(X)                # Находим собственные вектора
    eps=np.sort(s)[-num_components]       # пороговое собственное значение
    E = np.array([vec for val,vec in zip(s,V)
                  if val>eps])            # берем только важные вектора
    X_=np.dot(E,X.T).T                    # Преобразуем данные
    return X_,E
```

# PCA, Визуализация



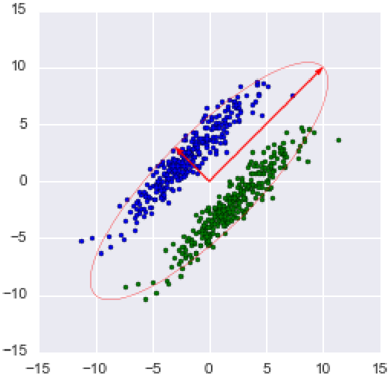
# PCA, Визуализация




(a) (b)

# PCA, Визуализация

Бывают случаи, когда проецировать нужно на малую компоненту



# Kernel PCA

$$\langle \phi(a), \phi(b) \rangle = K(a, b)$$

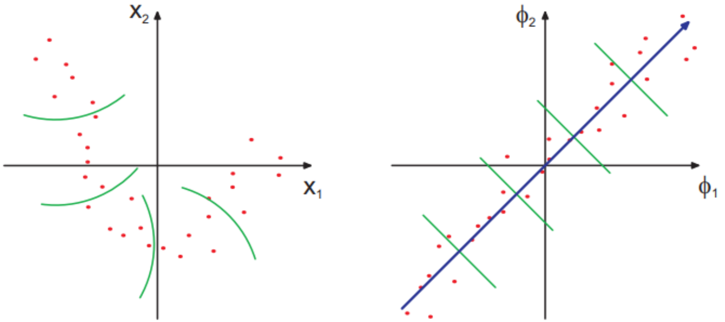
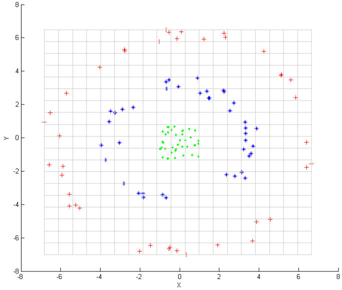
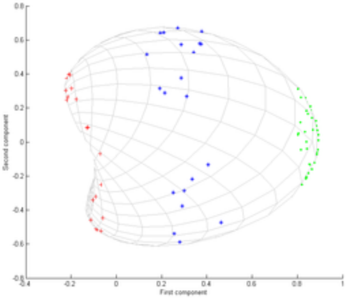



Рис. 8: Иллюстрация ядрового метода главных компонент.

# kernel pca

$$k(a, b) = e^{-||a-b||^2}$$



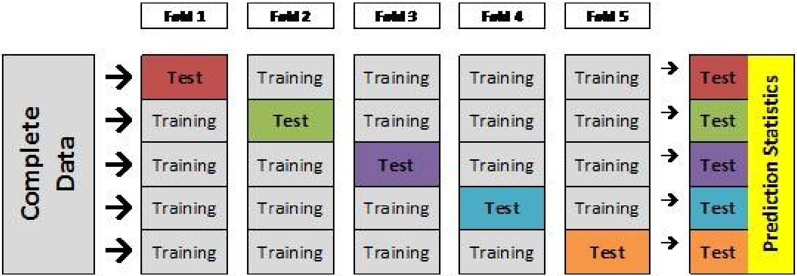



# Отбор признаков

- Обучаемся на подмножестве, выбираем лучший скор
- Жадный перебор
- ADD-DEL
- Из данных самой модели

# Кросс-валидация

Делим на k частей (фолдов).
adf Тренируем на k-1 и оцениваем качество модели на одном из них.



# Ссылки

*Статья на Хабре*
*Визуализация собственных векторов*
*Визуализация PCA*
*Лекция от Стенфорда(pdf)*
*Лекция от Стенфорда*