Исследование алгоритмов уменьшения размерности данных для задачи классификации

Руслан Сайфутдинов, 344 гр.
Научный руководитель: Константин Невоструев
Кафедра системного программирования

Введение

- Задача: распознавать что-нибудь на изображениях и не только на изображениях
- Изображения слишком большие, чтобы запустить обучения прямо на них
- Можно просто уменьшить разрешение
- А можно попытаться сохранить больше важной информации

Уменьшение размерности

Постановка задачи

- Реализовать несколько интересных алгоритмов уменьшения размерности данных
- Сравнить их работу

Данные

- MNIST

Тренировочная выборка: 60,000

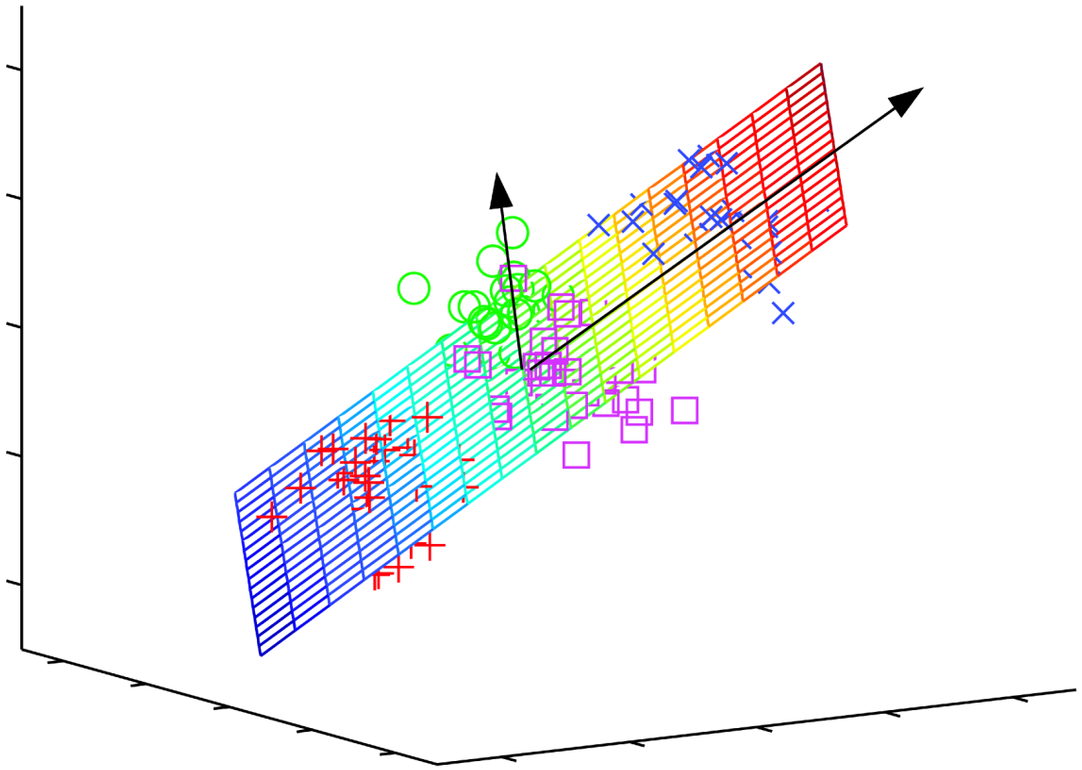
Тестовая: 10,000

784 признака

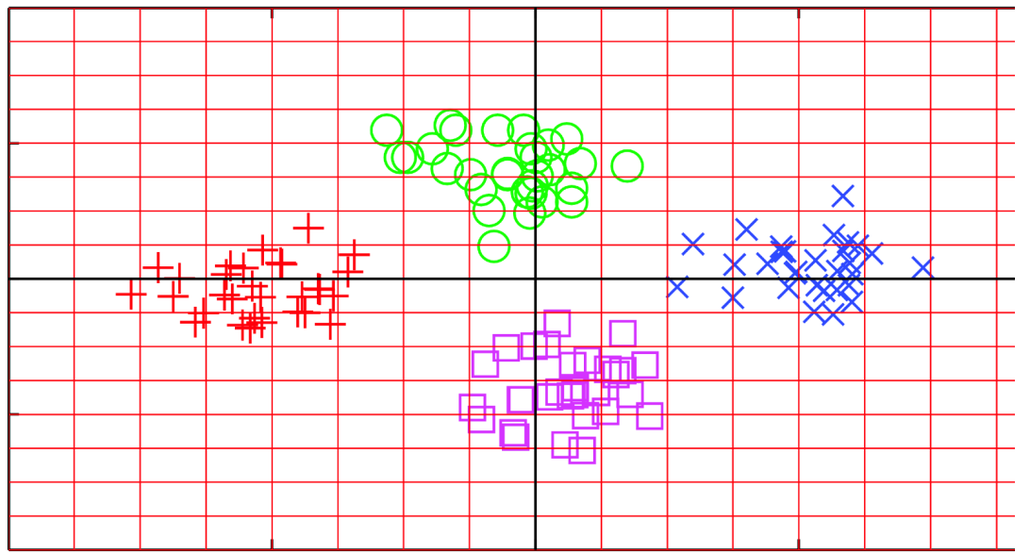
Алгоритмы

- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)
- SVD (Singular Value Decomposition)

PCA



PCA



исходное пространство данных

преобразованное пространство

ICA

- Пытается разделить данные на максимально независимые компоненты
- Критерий независимости — негауссовость
- Мера негауссовости — коэффициент эксцесса:

$$\mathrm{kurt}(v) = \mathbb{E}[v^4] - 3(\mathbb{E}[v^2]^2)$$

SVD

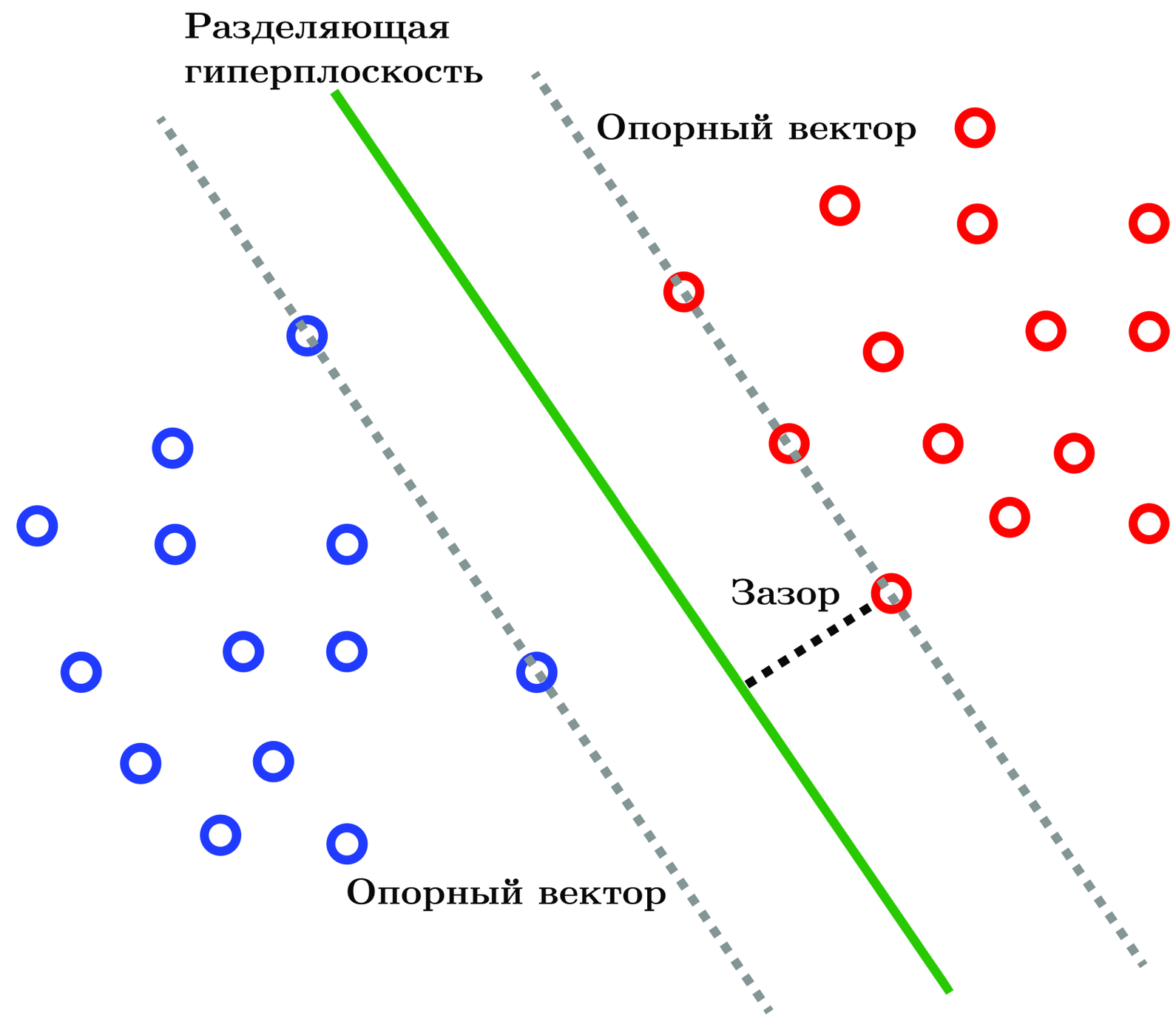
- Раскладываем матрицу X:

$$X = U\Sigma V^*$$

- Приближаем матрицей меньшего ранга:

$$X_k = U_k\Sigma_k V_k^*$$

SVM


Разделяющая
гиперплоскость
Опорный вектор
Зазор
Опорный вектор

Методика сравнения

- Python 2.7, numpy, scipy, libsvm
- Intel Core i7 2.9 ГГц, 8 ГБ RAM
- $F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$, больше — лучше

Точность

F1 Score

0.990
0.985
0.980
0.975
0.970
0.965
0.960
0.955
0.950
0.945
0.940
0.935
0.930

0
50
100
150
200
250
300
350
400

количество компонент

PCA
ICA
SVD
None

Время работы

PCA
ICA
SVD

время, сек

800
700
600
500
400
300
200
100
0

0
50
100
150
200
250
300
350
400

количество компонент

Время обучения

PCA
ICA
SVD
None

время, сек

900
800
700
600
500
400
300
200
100
0

0
50
100
150
200
250
300
350
400

количество компонент

Время предсказания

PCA
ICA
SVD
None

время, сек

0
20
40
60
80
100
120
140
160
180

0
50
100
150
200
250
300
350
400

количество компонент

Суммарное время работы

PCA
ICA
SVD
None

время, сек

1,800
1,600
1,400
1,200
1,000
800
600
400
200
0

0
50
100
150
200
250
300
350
400

количество компонент

Полученные результаты

- Реализованы алгоритмы PCA, ICA, SVD
- Проведено сравнение их работы в зависимости от количества компонент