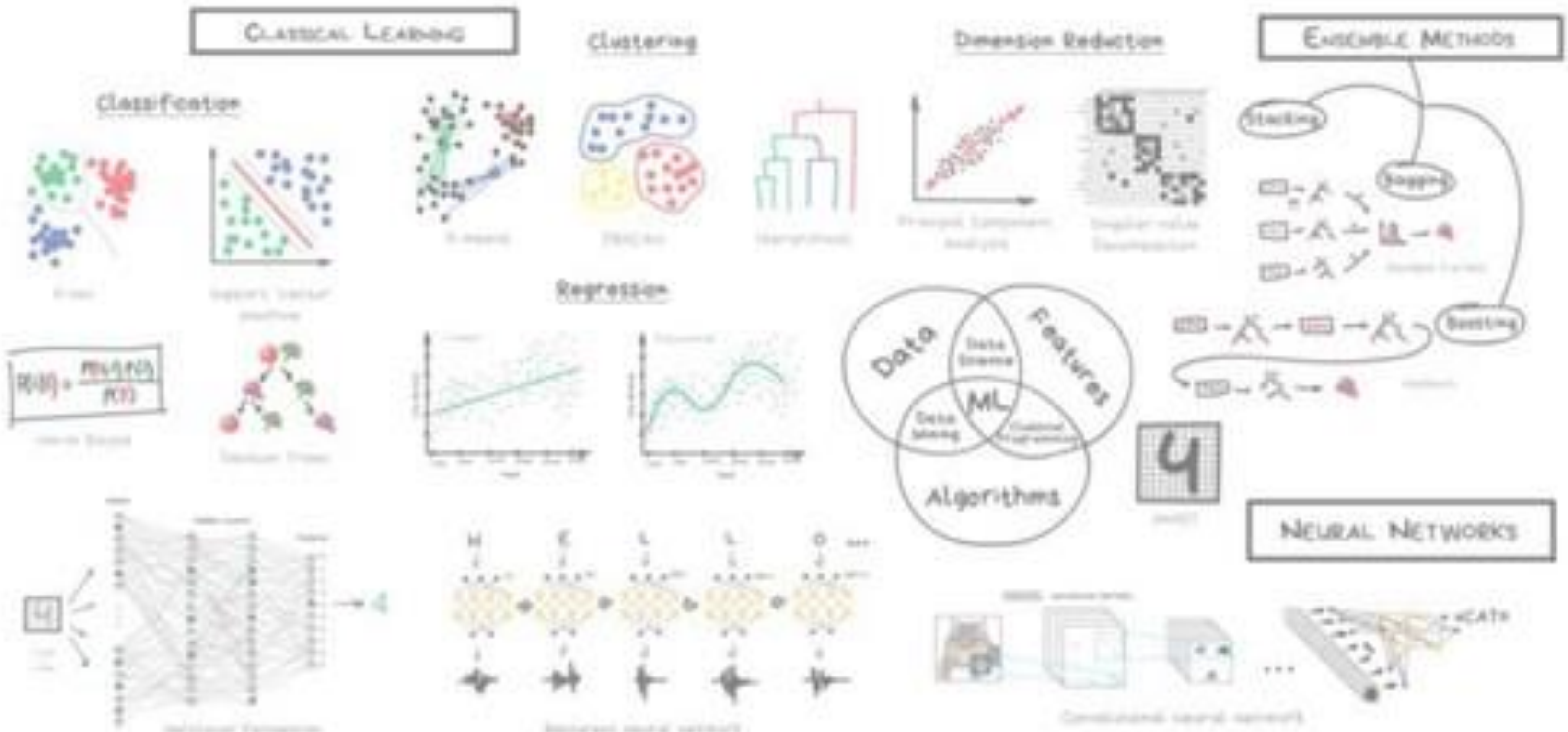
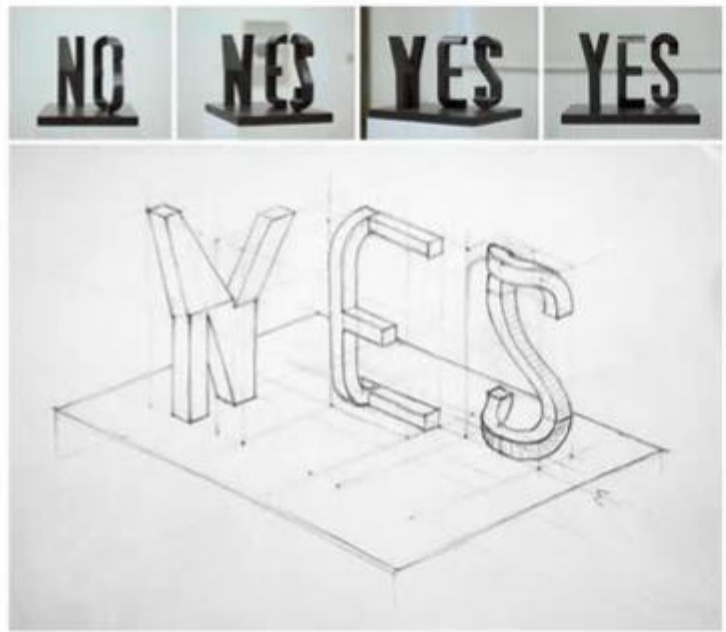
МАШИНЕ НАВЧАННЯ

Класичне навчання. Навчання без вчителя



Лекція №9

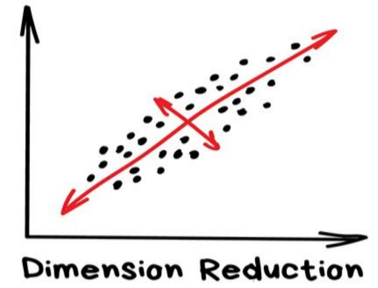
Метод сингулярного розкладання (SVD)



$$A_{x \times d} = U_{x \times x} \Sigma_{x \times d} V^T_{d \times d}$$

Зменшення розмірності

Метод сингулярного розкладання (SVD)



Сингулярним розкладанням будь-якої матриці X розміру $n \times d$ є її подання у вигляді добутку трьох матриць: U, Σ, V:

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

Є обмеження на U, Σ, V:

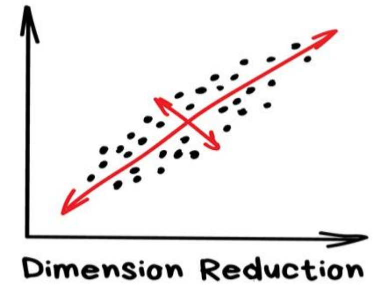
- U має розміри $n \times n$, Σ має розмір $n \times d$, а V має розмір $d \times d$.
- U і V – ортогональні матриці. Тобто $U^T U = I$ і $V^T V = I$.
- Σ є діагональною матрицею: $\Sigma = diag(\sigma_1, \dots, \sigma_d)$ – сингулярні значення матриці X. Тобто всі елементи в Σ дорівнюють нулю, якщо вони не лежать на діагоналі. Крім того, діагональні елементи в Σ розташовані від найбільшого до найменшого.

Стовбці матриці U – ліві сингулярні вектори;

Стовбці матриці V – праві сингулярні вектори;

Зменшення розмірності

Метод сингулярного розкладання (SVD)



Математичний сенс трьох матриць U, Σ, V в SVD

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

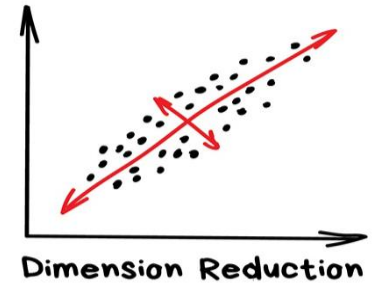
Встановимо,
як значення $\sigma_1, \dots, \sigma_d$ діагональної матриці залежать від даних
Знайдемо матрицю коваріації

$$S = \frac{X^T X}{n} = \frac{1}{n} V\Sigma U^T U\Sigma V^T = [U^T U = I] = \frac{1}{n} V\Sigma^2 V^T$$

Таким чином

- Стовбці $v_1, \dots, v_d$ матриці V являють собою власний базис матриці коваріацій S;
- $\sigma_1^2/n, \dots, \sigma_d^2/n$ – власні значення матриці коваріацій S;
- Вектори $v_1, \dots, v_d$ – головні (основні) компоненти (principal components) для матриці даних X

Зменшення розмірності
Метод сингулярного розкладання (SVD)



Математичний сенс трьох матриць U, Σ, V в SVD

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

Для будь-якого вектору $\widetilde{x_i} = Xv_i$ маємо

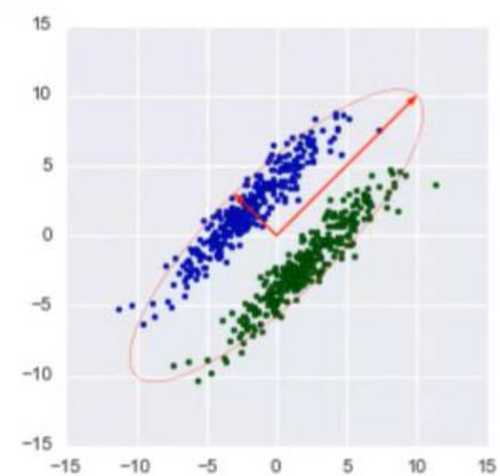
$$Var(\widetilde{x_i}) = Var(Xv_i) = \frac{\sigma_i^2}{n}, \qquad \widetilde{x_i} = Xv_i = \sigma_i^2 u_i$$

Властивість матриці Σ:

$$\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_d^2$$

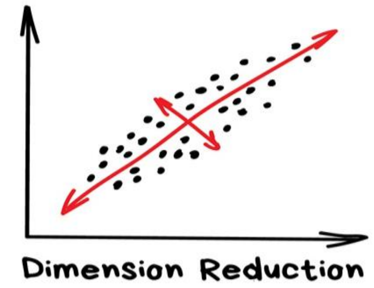
Таким чином:

- Перша головна компонента u_1 характеризується тією властивістю, що $\widetilde{x_i}$ має максимальну дисперсію серед всіх нормованих лінійних комбінацій стовбців матриці X;
- $\widetilde{x_d}$ має мінімальну дисперсію.



Зменшення розмірності

Метод сингулярного розкладання (SVD)



SVD із середини

У якості прикладу розглянемо результати гри в гольф. Гравець 1, гравець 2 та гравець 3 грають разом на 9-ти лунках. Їх карта результатів, яку можна розглядати як матрицю (лунка/гравець) має наступний вигляд

Hole	Par	Player 1	Player 2	Player 3
1	4	4	4	4
2	5	5	5	5
3	3	3	3	3
4	4	4	4	4
5	4	4	4	4
6	4	4	4	4
7	4	4	4	4
8	3	3	3	3
9	5	5	5	5

Par – нормативна кількість ударів за які гравець має пройти дану лунку

Зменшення розмірності

Метод сингулярного розкладання (SVD)

Давайте розглянемо проблему, пов'язану зі спробою передбачити, який рахунок набере кожен гравець на певній лунці. Одна з ідей полягає в тому, щоб надати кожній лунці фактор складності лунки (HoleDifficulty), а кожному гравцеві - коефіцієнт майстерності гравця (PlayerAbility). Фактична оцінка передбачається шляхом множення цих двох факторів разом: PredictedScore = HoleDifficulty * PlayerAbility

Player 1	Player 2	Player 3
4	4	4
5	5	5
3	3	3
4	4	4
4	4	4
4	4	4
4	4	4
3	3	3
5	5	5

=

HoleDifficulty
4
5
3
4
4
4
4
3
5

*

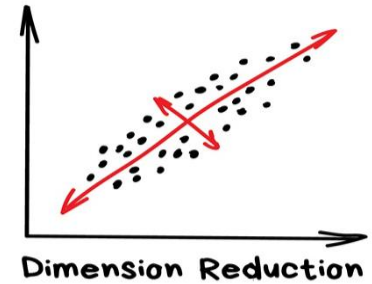
PlayerAbility		
Player 1	Player 2	Player 3
1	1	1

Для першої спроби давайте покладемо HoleDifficulty рівним номінальному рахунку для лунки, а майстерність гравця PlayerAbility прирівняємо до 1. Таким чином, на першій лунці з HoleDifficulty = 4, ми очікуємо, що гравець з PlayerAbility = 1 отримає рахунок:

$$PredictedScore = HoleDifficulty * PlayerAbility = 4 * 1 = 4$$

Зменшення розмірності

Метод сингулярного розкладання (SVD)



Player 1	Player 2	Player 3
4	4	4
5	5	5
3	3	3
4	4	4
4	4	4
4	4	4
4	4	4
3	3	3
5	5	5

=

HoleDifficulty
4
5
3
4
4
4
4
3
5

*

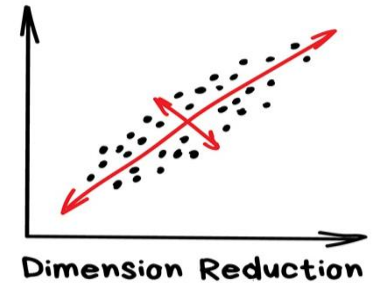
PlayerAbility		
Player 1	Player 2	Player 3
1	1	1

Для всієї нашої системи показників або матриці все, що нам потрібно зробити, це помножити PlayerAbility (передбачається, що вона дорівнює 1 для всіх гравців) HoleDifficulty (варіюється від пар 3 до пар 5), і ми можемо точно передбачити всі результати в нашому прикладі .

Фактично, це одновимірна (1-D) SVD-факторизація системи показників. Ми можемо представити нашу систему показників або матрицю як добуток двох векторів: вектор складності лунок і вектор коефіцієнта майстерності гравця. Щоб передбачити будь-який рахунок, необхідно помножити відповідний коефіцієнт складності лунки на відповідний фактор майстерності гравця.

Зменшення розмірності

Метод сингулярного розкладання (SVD)



Відмасштабуємо вектора так, щоб їх довжина дорівнювала 1. Наприклад, вектор PlayerAbility змінений так, що сума квадратів його елементів дорівнює 1, а не поточним $1^2 + 1^2 + 1^2 = 3$. Щоб зробити це, ми повинні розділити кожен елемент на квадратний корінь із 3. Так само розділимо кожен елемент HoleDifficulty на квадратний корінь із 148. Квадратний корінь із 3, помножений на квадратний корінь із 148, є нашим коефіцієнтом масштабування (ScaleFactor) 21,07. Повна факторизація 1-D SVD:

Player 1	Player 2	Player 3
4	4	4
5	5	5
3	3	3
4	4	4
4	4	4
4	4	4
4	4	4
3	3	3
5	5	5

=

HoleDifficulty
0.33
0.41
0.25
0.33
0.33
0.33
0.33
0.25
0.41

×

ScaleFactor
21.07

×

PlayerAbility		
Player 1	Player 2	Player 3
0.58	0.58	0.58

HoleDifficulty - лівий сингулярний вектор.
ScaleFactor - **сингулярне** значення
PlayerAbility - правий сингулярний вектор.

Якщо ми точно представимо ці 3 частини та перемножимо їх разом, ми отримаємо точні вихідні оцінки. Це означає, що наша матриця є матрицею рангу 1, тобто вона має простий і передбачуваний шаблон.

Зменшення розмірності

Метод сингулярного розкладання (SVD)

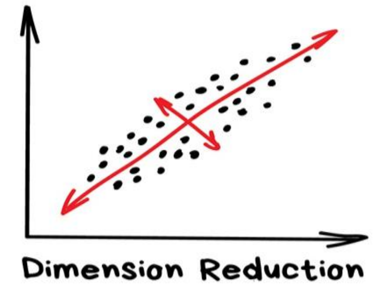
Більш складні матриці неможливо повністю передбачити використовуючи лише один набір факторів ScaleFactor. У такому випадку треба уводити додатковий набір факторі з метою уточнення прогнозу. З цією метою від реальних результатів віднімаються прогнозовані, отримуємо залишкові бали і знаходимо інший набір чисел HoleDifficulty2 і PlayerAbility2, які найкраще передбачають залишкові бали

Реальні результати перших 9-ти лунок Чемпіонату з гольфу 2007 року

Hole	Par	Player1	Player2	Player3
1	4	4	4	5
2	5	4	5	5
3	3	3	3	2
4	4	4	5	4
5	4	4	4	4
6	4	3	5	4
7	4	4	4	3
8	3	2	4	4
9	5	5	5	5

Зменшення розмірності

Метод сингулярного розкладання (SVD)



Hole	Par	Player1	Player2	Player3
1	4	3.95	4.64	4.34
2	5	4.27	5.02	4.69
3	3	2.42	2.85	2.66
4	4	3.97	4.67	4.36
5	4	3.64	4.28	4.00
6	4	3.69	4.33	4.05
7	4	3.33	3.92	3.66
8	3	3.08	3.63	3.39
9	5	4.55	5.35	5.00

=

HoleDifficulty
4.34
4.69
2.66
4.36
4.00
4.05
3.66
3.39
5.00

*

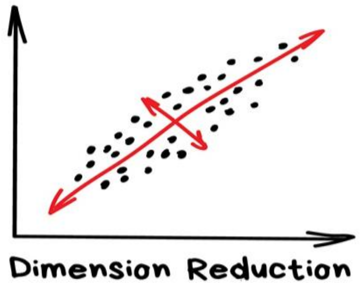
PlayerAbility		
Player1	Player2	Player3
0.91	1.07	1.00

1-D SVD !!!

Hole	Par	Player1	Player2	Player3
1	4	4	4	5
2	5	4	5	5
3	3	3	3	2
4	4	4	5	4
5	4	4	4	4
6	4	3	5	4
7	4	4	4	3
8	3	2	4	4
9	5	5	5	5

Коефіцієнт складності лунки HoleDifficult майже дорівнює середньому значенню цієї лунки для всіх трьох гравців: лунка 6 де середній бал 4 має 4.05; майстерність гравця PlayerAbility майже дорівнює його відсотку від номіналу: гравець 2 зробив 39 ударів з 36: 39/36=1.08

Зменшення розмірності



Метод сингулярного розкладання (SVD)

Визначимо різницю між фактичними оцінками та нашим одномірним наближенням. Позитивна різниця означає, що фактична оцінка вище за прогнозовану, мінусова різниця означає, що фактична оцінка нижче за прогнозовану. Наприклад, першій лунці Player2 отримав 4, а прогнозований рахунок був 4,64 (різниця -0,64). Іншими словами, ми маємо додати -0,64 до нашого прогнозу, щоб отримати фактичну оцінку.

Як тільки ці відмінності будуть знайдені, ми можемо зробити те саме знову і передбачити ці відмінності, використовуючи формулу HoleDifficulty2 * PlayerAbility2.

	Differences		
Hole	Player1	Player2	Player3
1	0.05	-0.64	0.66
2	-0.28	-0.02	0.31
3	0.58	0.15	-0.66
4	0.03	0.33	-0.36
5	0.36	-0.28	0.00
6	-0.69	0.67	-0.05
7	0.67	0.08	-0.66
8	-1.08	0.37	0.61
9	0.45	-0.35	0.00

=

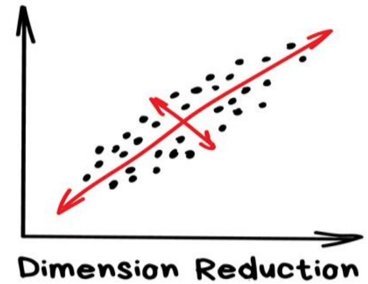
HoleDifficulty2
-0.18
-0.38
0.80
0.15
0.35
-0.67
0.89
-1.29
0.44

×

PlayerAbility2		
Player1	Player2	Player3
0.82	-0.20	-0.53

Зменшення розмірності

Метод сингулярного розкладання (SVD)



Повне SVD для цього прикладу

Player1	Player2	Player3
4	4	5
4	5	5
3	3	2
4	5	4
4	4	4
3	5	4
4	4	3
2	4	4
5	5	5

HoleDifficulty 1-3

4.34	-0.18	-0.90
4.69	-0.38	-0.15
2.66	0.80	0.40
4.36	0.15	0.47
4.00	0.35	-0.29
4.05	-0.67	0.68
3.66	0.89	0.33
3.39	-1.29	0.14
5.00	0.44	-0.36

✖

PlayerAbility 1-3

Player1	Player2	Player3
0.91	1.07	1.00
0.82	-0.20	-0.53
-0.21	0.76	-0.62

HoleDifficulty 1-3

0.35	0.09	-0.64
0.38	0.19	-0.10
0.22	-0.40	0.28
0.36	-0.08	0.33
0.33	-0.18	-0.20
0.33	0.33	0.48
0.30	-0.44	0.23
0.28	0.64	0.10
0.41	-0.22	-0.25

✖

ScaleFactor 1-3

21.07	0	0
0	2.01	0
0	0	1.42

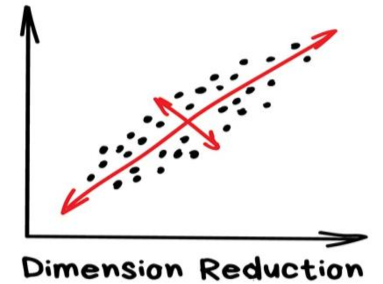
✖

PlayerAbility 1-3

Player1	Player2	Player3
0.53	0.62	0.58
-0.82	0.20	0.53
-0.21	0.76	-0.62

Зменшення розмірності

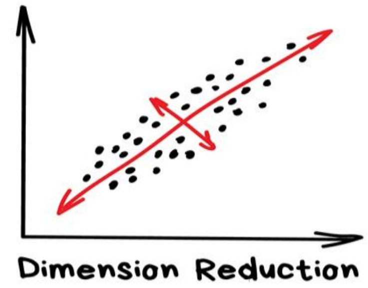
Метод сингулярного розкладання (SVD)



Одна дуже корисна властивість SVD полягає в тому, що він завжди знаходить оптимальний набір факторів, які найкраще передбачають оцінки, відповідно до стандартної міри подібності до матриць (нормою Фробеніуса). Тобто якщо ми використовуємо SVD для пошуку факторів матриці, це найкращі фактори, які можна знайти. Ця властивість оптимальності означає, що нам не потрібно ставити питання, чи може інший набір чисел краще передбачати результати.

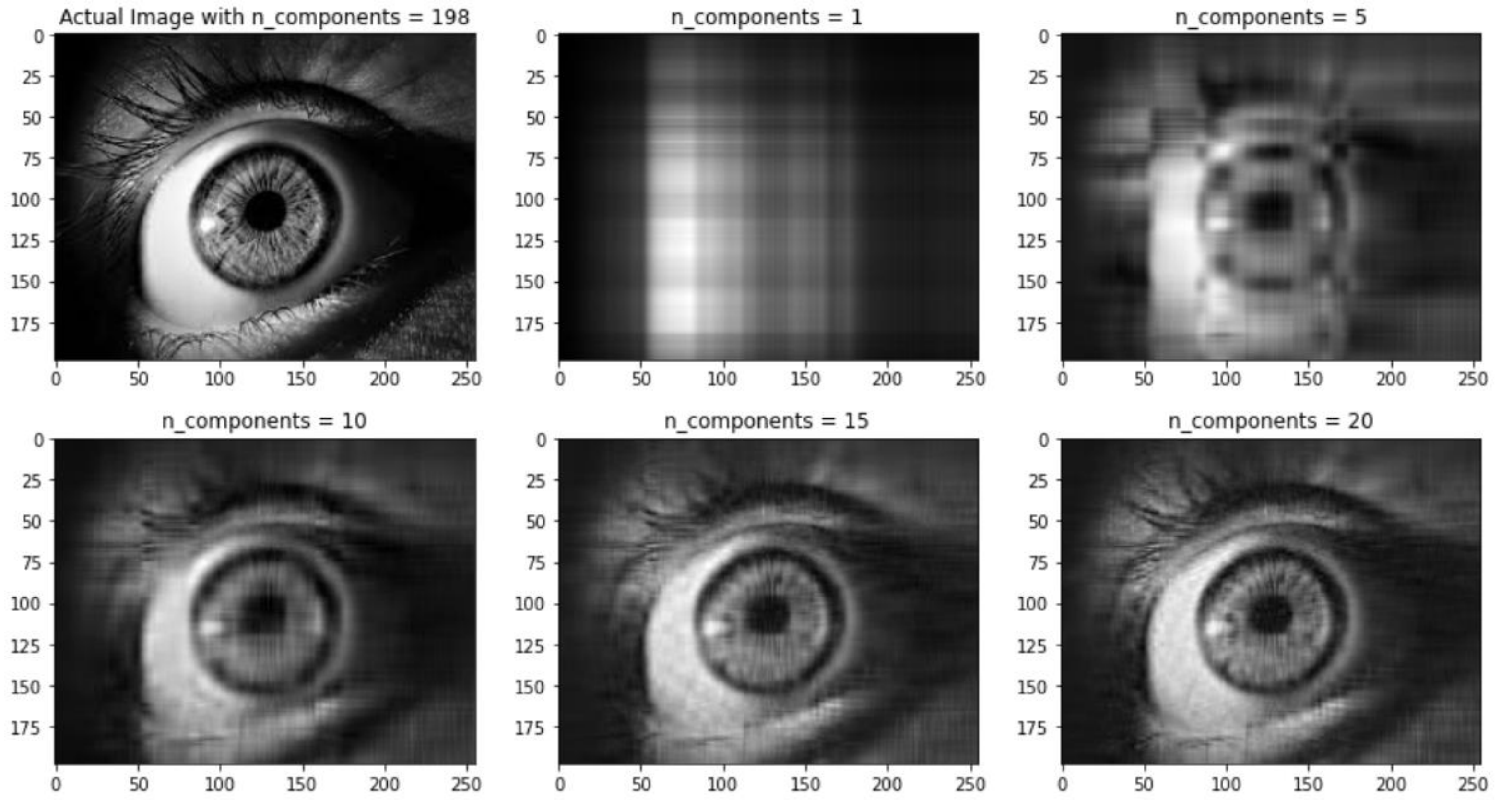
Зменшення розмірності

Метод сингулярного розкладання (SVD)



Стиснення зображення з використанням методу SVD

1. Відцифрувати чорнобілий малюнок.
2. Провести розкладання SVD
3. Відтворити малюнок для 1, 5, 15, … головних компонент



Дякую за увагу