Міністерство освіти і науки України
Львівський національний університет імені Івана Франка
Факультет електроніки та комп’ютерних технологій
Кафедра системного проектування

Звіт
Про виконання лабораторної роботи №8
З курсу «Системи машинного навчання»
Аналіз головних компонент (PCA)

Виконала:
Студентка групи ФЕС-32
Філь Дарина

Перевірив:
Доцент Колич І.І.

Львів 2024

Мета: навчитися проводити аналіз головних компонент (PCA) для зменшення розмірності даних та виявлення структурних взаємозв'язків між змінними.

Теоретичні відомості

Аналіз головних компонент (PCA)

Аналіз головних компонент (PCA) є статистичним методом, який використовується для зменшення розмірності даних шляхом перетворення їх на новий набір змінних (головних компонент), які є ортогональними між собою. Головні компоненти обираються таким чином, щоб максимізувати дисперсію даних.

Алгоритм:

1. Центрувати дані шляхом віднімання середнього значення.
2. Обчислити коваріаційну матрицю даних.
3. Виконати сингулярне розкладання або обчислити власні вектори та власні значення коваріаційної матриці.
4. Відсортувати власні вектори за спаданням власних значень.
5. Вибрати кількість головних компонент, які пояснюють більшу частину дисперсії даних.

Переваги:

- Зменшує розмірність даних з мінімальною втратою інформації.
- Полегшує візуалізацію багатовимірних даних.

Недоліки:

- Працює лише з лінійними взаємозв'язками.
- Не дає зрозумілої інтерпретації головних компонент.

Оцінка PCA

Власні значення (Eigenvalues)
Опис: Власні значення показують кількість дисперсії, що пояснюється кожною головною компонентою. Більші значення вказують на більший вклад у пояснення дисперсії даних.

Інтерпретація:

- Компоненти з власними значеннями більше 1 зазвичай вважаються значущими.

Переваги:

- Дозволяє зрозуміти, які компоненти мають найбільший внесок у пояснення дисперсії даних.

Недоліки:

- Вибір порогу для значущих ейгензначень може бути суб'єктивним.

Дисперсія (Explained Variance)
Опис: Дисперсія показує частку загальної дисперсії даних, яка пояснюється кожною головною компонентою.

Інтерпретація:

- Вибираються компоненти, які пояснюють більшу частину дисперсії даних.

Переваги:

- Дозволяє зрозуміти, скільки інформації зберігається після зменшення розмірності.

Недоліки:

- Може бути важко вирішити, скільки компонент залишити для аналізу.

Оцінка кумулятивної дисперсії:

- Кумулятивна дисперсія показує накопичену частку загальної дисперсії, пояснюваної головними компонентами.
- Використовується для визначення кількості головних компонент, які слід залишити, щоб зберегти більшу частину інформації.

Хід роботи

Завдання

1. **Підготовка даних**
   1.1. Використайте набір даних Wine.
   1.2. Нормалізуйте дані для покращення продуктивності PCA.
2. **Аналіз головних компонент (PCA)**
   2.1. Виконайте PCA на даних.
   2.2. Визначте кількість головних компонент, які пояснюють більшу частину дисперсії даних.
   2.3. Візуалізуйте головні компоненти.
3. **Оцінка PCA**
   3.1. Оцініть результати PCA за допомогою власних значень та дисперсії.
   3.2. Побудуйте графік кумулятивної поясненої дисперсії для визначення оптимальної кількості головних компонент.

4. **Використання PCA для класифікації**
   4.1. Використайте головні компоненти як вхідні дані для побудови класифікаційної моделі (наприклад, логістична регресія).
   4.2. Побудуйте класифікаційну модель з такою з кількістю ознак, як і попередня модель, але без використання головних компонент
   4.3. Оцініть продуктивність класифікаційної моделі на основі головних компонент та без використання головних компонент.
5. **Оформлення звіту**
   5.1. Оформіть звіт з результатами лабораторної роботи, включаючи графіки, таблиці та аналіз результатів

```
data = load_wine()
X = data.data
y = data.target


scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)


pca = PCA()
X_pca = pca.fit_transform(X_scaled)
```

Рис. 1 Завантаження дата-сету, нормалізація даних та використання PCA

```
explained_variance = pca.explained_variance_ratio_
cumulative_variance = np.cumsum(explained_variance)
```

Рис. 2 Знаходимо кумулятивну дисперсію для PCA

```
n_components = np.argmax(cumulative_variance >= 0.95) + 1
print(f"Number of components that explain 95% of the variance: {n_components}")


X_pca_n = X_pca[:, :n_components]
X_train_pca, X_test_pca, y_train, y_test = train_test_split(X_pca_n, y, test_size=0.3, random_state=42)

model_pca = LogisticRegression(max_iter=500)
model_pca.fit(X_train_pca, y_train)
y_pred_pca = model_pca.predict(X_test_pca)
accuracy_pca = accuracy_score(y_test, y_pred_pca)
print(f"Classification accuracy with PCA components: {accuracy_pca:.4f}")


X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)

model_orig = LogisticRegression(max_iter=500)
model_orig.fit(X_train, y_train)
y_pred_orig = model_orig.predict(X_test)
accuracy_orig = accuracy_score(y_test, y_pred_orig)
print(f"Classification accuracy without PCA components: {accuracy_orig:.4f}")
```

Рис. 3 Визначення компонент, які пояснюють основну частину дисперсії; класифікація із застосуванням цих компонент та методу PCA, а також порівняння з класифікацією без використання PCA

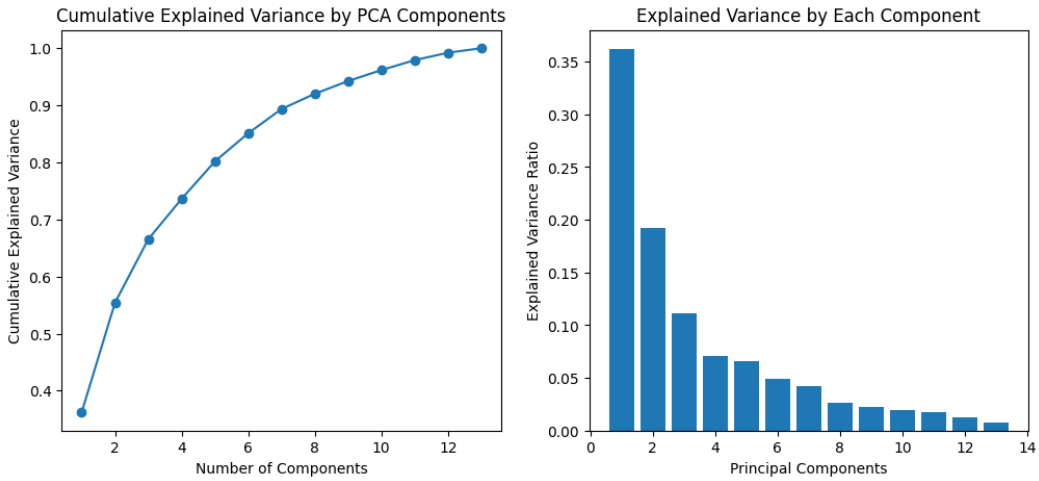

Рис. 4 Графіки що пояснюють кумулятивну дисперсію та візуалізація компонент, що пояснюють більшу частину дисперсії.

```
Number of components that explain 65% of the variance: 3
Classification accuracy with PCA components: 0.9630
Classification accuracy without PCA components: 0.9815
```

Рис. 5 Можемо побачити, що кількість компонент, які пояснюють більшу частину дисперсії, становить три, що узгоджується з візуалізацією на графіку

Також помітно, що точність класифікації із застосуванням методу PCA дещо нижча порівняно з точністю без використання цього методу. Можливо, це пояснюється тим, що набір даних Wine вже містить добре розділені за класами дані, що дозволяє досягати високої точності класифікації навіть без використання методу головних компонент.

```
Number of components that explain 95% of the variance: 10
Classification accuracy with PCA components: 0.9815
Classification accuracy without PCA components: 0.9815
```

Рис. 6 Використовуючи 10 компонент, які майже повністю пояснюють дисперсію набору даних, можемо помітити, що точність методів для набору даних Wine практично однакова.

Раніше у 7-й лабораторній виникала проблема з низькою точністю кластеризації, що при візуалізації показувало відсутність чітких кластерів у даних. Застосування методу PCA дозволило збільшити точність кластеризації втричі, що підкреслює його ефективність у завданнях кластеризації.

Висновок: у цій лабораторній роботі я засвоїла навички використання методу PCA та визначила кількість головних компонент, які пояснюють більшу частину дисперсії в датасеті. Під час виконання помітила, що точність класифікації із

застосуванням методу PCA та без нього фактично однакова для даних, які вже мають чітку класифікацію.