BTLT - Data Mining - Tuần 3

21280099 - Nguyễn Công Hoài Nam

Ngày 30 tháng 3 năm 2024

I. Đề bài

Bài tập tuần 03: Các em phân biệt sự khác nhau và điểm mạnh/điểm yếu của 03 phương pháp: LSA, PCA, và SVD trong việc giảm số chiều dữ liệu.

II. Bài làm

1. PCA (Principal Component Analysis)

a) Sự khác nhau

Phân tích thành phần chính (PCA) là một kỹ thuật giảm số chiều dữ liệu được sử dụng rộng rãi trong phân tích dữ liệu và học máy. Mục tiêu chính là biến đổi dữ liệu nhiều chiều thành dữ liệu ít chiều hơn mà vẫn dữ được hầu hết thông tin quan trọng. PCA tập trung vào việc tìm ra các thành phần chính (principal components) của dữ liệu
Dưới đây là các bước thực hiện

1. **Chuẩn hoá dữ liệu**: Trước khi áp dụng PCA, dữ liệu thường cần được chuẩn hóa để đảm bảo rằng các đặc trưng có cùng phạm vi.
2. **Tính ma trận hiệp phương sai**: Ma trận này cho biết mức độ tương quan giữa các cặp đặc trưng trong dữ liệu.
3. **Tính các vector riêng và giá trị riêng của ma trận hiệp phương sai**: Các vector riêng đại diện cho các hướng trong không gian dữ liệu, còn các giá trị riêng đại diện cho mức độ biến thiên của dữ liệu theo các hướng đó.
4. **Sắp xếp các vector riêng và giá trị riêng**: Các giá trị riêng thường được sắp xếp theo thứ tự giảm dần. Điều này có nghĩa là vector riêng có giá trị riêng lớn nhất sẽ đứng đầu danh sách và đó sẽ là thành phần chính đầu tiên của PCA.
5. **Chọn các thành phần chính**: Người ta thường chọn một số lượng thành phần chính (principal components) dựa trên tỉ lệ phương sai giải thích (explained variance ratio) mà họ muốn giữ lại. Tổng số thành phần chính được chọn thường phụ thuộc vào mức độ giảm số chiều mà bạn mong muốn và mức độ giải thích phương sai mà bạn muốn đạt được.
6. **Biến đổi dữ liệu (transform)**: biến đổi dữ liệu ban đầu sử dụng các vector riêng.

b) Điểm mạnh

- PCA giảm chiều dữ liệu một cách hiểu quả và dữ được phần lớn nội dụng, điều này có ích cho các model gặp vấn đề với dữ liệu nhiều chiều
- Các thành phần chính là trực giao (không tương quan), có nghĩa là chúng chứa thông tin độc lập, đơn giản hóa việc diễn giải các đặc trưng được giảm.
- PCA có thể giúp giảm nhiễu bằng cách tập trung vào các thành phần giải thích phương sai lớn nhất trong dữ liệu.
- Dễ dàng cho trực quan hoá dữ liệu, hỗ trợ hiểu cấu trúc và khai phá dữ liệu

PCA được sử dụng khi có dữ liệu nhiều chiều, các đặc trưng có tương quan cao, cần trực quan hoá dữ liệu, mối quan hệ giữa các biến là tuyến tính.

c) Điểm yếu

- PCA giả định rằng các mối quan hệ giữa biến là tuyến tính, điều này có thể không đúng trong tất cả các trường hợp.
- PCA nhạy cảm với tỷ lệ của các đặc trưng, vì vậy thường cần chuẩn hóa.
- Các ngoại lai có thể ảnh hưởng đáng kể đến kết quả của PCA, vì nó tập trung vào việc thu thập phương sai lớn nhất, điều này có thể bị ảnh hưởng bởi các giá trị cực đoan.
- Các thành phần chính là sự kết hợp tuyến tính của các đặc trưng ban đầu nên trong quá trình đó có thể mất một số tính đặc trưng ban đầu

2. SVD (Singular Value Decomposition)

a) Sự khác nhau

Singular Value Decomposition (SVD) là một kỹ thuật phân rã ma trận được sử dụng rộng rãi trong nhiều ứng dụng, bao gồm đại số tuyến tính, xử lý tín hiệu và máy học. Nó phân rã một ma trận thành ba ma trận khác, cho phép biểu diễn ma trận ban đầu dưới dạng rút gọn. Các bước trong SVD như sau:

1. **Phân ra ma trận**:
   - SVD phân rã một ma trận M kích thước m x n thành ba ma trận: $M = U\Sigma V^T$
   - U là ma trận trực giao m x m, Σ là ma trận chéo kích thước m x r, V là ma trận trực giao kích thước r x n. Trong đó r là hạng của ma trận M.
   - Các phần tử trên đường chéo của Σ là các giá trị đơn thức của ma trận ban đầu M, và chúng được sắp xếp theo thứ tự giảm dần. Các cột của U là các vector đơn thức trái của M, còn các cột của V là các vector đơn thức phải của M.
2. **Dạng rút gọn (truncate SVD)**:
   - Chọn k giá trị đơn thức lớn nhất trong Σ. Các cột này có thể được chọn từ Σ và các hàng có thể được chọn từ V^T. Một ma trận mới B có thể được tái tạo từ ma trận ban đầu M bằng công thức sau:

$$B = U\Sigma$$

$$B = V^T A$$

   - Trong đó, Σ chỉ chứa các cột hàng đầu tiên trong Σ ban đầu dựa trên các giá trị đơn thức và V^T chứa các hàng đầu tiên của V^T tương ứng với các giá trị đơn thức.

b) Điểm mạnh

- **Giảm chiều dữ liệu:** SVD cho phép giảm số chiều dữ liệu bằng cách chỉ giữ lại các giá trị và vector quan trọng nhất.
- **Nén dữ liệu:** SVD được sử dụng trong các nhiệm vụ nén dữ liệu, giảm yêu cầu lưu trữ của một ma trận.
- **Giảm nhiễu:** Bằng cách chỉ sử dụng các giá trị đơn thức quan trọng nhất, SVD có thể giúp giảm tác động của nhiễu trong dữ liệu.
- **Ổn định số học:** SVD ổn định số học và thích hợp cho việc giải phương trình tuyến tính trong các hệ thống không ổn định.
- **Trực giao:** Các ma trận U và V trong phân rã SVD là trực giao, bảo toàn mối quan hệ giữa các hàng và cột của ma trận ban đầu.

Sử dụng SVD khi cần giảm chiều dữ liệu, nén giữ liệu, xử lý tín hiệu (giảm nhiễu, trích xuất đặc trưng), mô hình chủ đề (topic modeling) như LSA (Latent Semantic Analysis)

c) Điểm yếu

- **Độ phức tạp tính toán:** Tính toán SVD đầy đủ cho các ma trận lớn có thể tốn kém về mặt tính toán.
- **Yêu cầu bộ nhớ:** Lưu trữ các ma trận đầy đủ U, Σ, và V có thể tốn kém về mặt bộ nhớ, đặc biệt là đối với các ma trận lớn.

- **Nhạy cảm với giá trị khuyết:** SVD nhạy cảm với các giá trị bị thiếu trong dữ liệu, và xử lý các giá trị bị thiếu yêu cầu các kỹ thuật đặc biệt.

3. LSA (Latent Semantic Analysis)

a) Sự khác nhau

LSA thường được sử dụng cho giảm chiều dữ liệu trong xử lý ngôn ngữ tự nhiên (NLP).
LSA (Latent Semantic Analysis) giảm chiều dữ liệu bằng cách biểu diễn một tập hợp các văn bản dưới dạng ma trận tần số từ và xác định các khái niệm hoặc chủ đề tiềm ẩn trong dữ liệu. Cụ thể, quá trình thực hiện như sau:

1. **Tiền xử lý văn bản**: Loại bỏ stop words, dấu câu và nhiễu khác, cũng như thực hiện stemming hoặc lemmatization để chuyển các từ còn lại về dạng cơ bản.
2. **Ma trận từ-văn bản**: Chuyển đổi các văn bản đã được tiền xử lý thành ma trận từ-văn bản, trong đó mỗi hàng đại diện cho một văn bản và mỗi cột đại diện cho một từ trong từ vựng. Ma trận được điền với tần số của mỗi từ trong mỗi văn bản.
3. **Đánh trọng số cho từng từ**: Sử dụng TF-IDF (tần suất từ-nghịch đảo tần suất văn bản) để giảm ảnh hưởng của các từ có tần suất cao. Trọng số TF-IDF cho một từ trong một văn bản là tích của TF và IDF của từ đó.
4. **Phân rã giá trị đơn nhất (SVD)**: Phân rã ma trận từ-văn bản đã được đánh trọng số bằng SVD thành ba ma trận thành phần: U, Σ, và V^T, trong đó U và V^T là các ma trận trực giao và Σ là một ma trận đường chéo chứa các giá trị đơn nhất.
5. **Biểu diễn chiều thấp**: Giảm chiều của ma trận từ-văn bản bằng cách chọn các giá trị đơn nhất hàng đầu và cột tương ứng trong U và hàng trong Vt. Điều này giảm số lượng cột trong ma trận từ n (số từ duy nhất) xuống k, với k là một tham số do người dùng xác định.
6. **Lựa chọn số chiều sử dụng (k)**: Chọn một k thấp có thể đại diện cho cấu trúc tiềm ẩn của n từ trong các văn bản. Có thể sử dụng kỹ thuật ngưỡng giá trị đơn nhất để chọn k, hoặc parametrized k là một phần của các nhiệm vụ dưới luồng và sử dụng các kỹ thuật như Cross-validation để đến giá trị k tối ưu cho các nhiệm vụ dưới luồng.

b) Điểm mạnh

- **Tính tổng quát:** LSA giúp trích xuất thông tin ẩn từ dữ liệu văn bản một cách tổng quát, giúp cải thiện hiểu biết về các mối quan hệ ngữ nghĩa giữa các từ và tài liệu.
- **Giảm chiều dữ liệu:** LSA giảm chiều dữ liệu một cách hiệu quả bằng cách chọn ra các thành phần chính, giúp giảm độ phức tạp và tăng hiệu suất tính toán trong quá trình xử lý và phân tích dữ liệu.
- **Tính linh hoạt:** LSA có thể được áp dụng trong nhiều lĩnh vực khác nhau như phân loại văn bản, tìm kiếm thông tin, và gợi ý, làm cho nó trở thành một công cụ linh hoạt cho việc xử lý và phân tích dữ liệu văn bản.

LSA thích hợp khi cần giảm chiều dữ liệu văn bản và trích xuất thông tin tiềm ẩn.

c) Điểm yếu

- **Mất mát thông tin:** Quá trình giảm chiều dữ liệu có thể dẫn đến mất mát thông tin quan trọng, đặc biệt là trong việc trích xuất và biểu diễn ngữ nghĩa của dữ liệu văn bản.
- **Đối mặt với nhiễu:** LSA có thể bị ảnh hưởng bởi nhiễu trong dữ liệu, đặc biệt là trong các tập dữ liệu lớn hoặc không cấu trúc.
- **Phụ thuộc vào tham số:** Việc lựa chọn số lượng thành phần chính trong LSA có thể ảnh hưởng đến kết quả cuối cùng, và việc này đôi khi khá khó khăn và tốn thời gian.