

CauHoiPCA

May 8, 2024

Dưới đây là 20 câu hỏi và câu trả lời về Phân tích thành phần chính (PCA):

- 1. PCA là viết tắt của điều gì?**
 - PCA là viết tắt của Phân tích thành phần chính (Principal Component Analysis).
- 2. PCA được sử dụng để làm gì trong phân tích dữ liệu?**
 - PCA được sử dụng để giảm chiều dữ liệu bằng cách chuyển đổi tập dữ liệu ban đầu thành một tập dữ liệu mới, giữ lại các thành phần chính chứa phần lớn thông tin của dữ liệu.
- 3. Các bước chính trong quá trình thực hiện PCA là gì?**
 - Các bước chính trong quá trình thực hiện PCA bao gồm chuẩn hóa dữ liệu, tính ma trận hiệp phương sai, tính các vector riêng và giá trị riêng của ma trận hiệp phương sai, chọn các thành phần chính và chuyển đổi dữ liệu ban đầu thành các thành phần chính.
- 4. Tại sao cần chuẩn hóa dữ liệu trước khi áp dụng PCA?**
 - Chuẩn hóa dữ liệu trước khi áp dụng PCA giúp đảm bảo rằng mỗi biến đều có cùng độ lớn và trọng số trong quá trình phân tích.
- 5. Phương pháp nào được sử dụng để tính ma trận hiệp phương sai trong PCA?**
 - Ma trận hiệp phương sai được tính bằng cách lấy ma trận covariance của dữ liệu, hoặc tính tích vô hướng của dữ liệu với chính nó chia cho số lượng quan sát.
- 6. Cách tính toán các vector riêng và giá trị riêng trong PCA là gì?**
 - Các vector riêng và giá trị riêng của ma trận hiệp phương sai được tính bằng cách giải quyết bài toán eigenvalue decomposition hoặc singular value decomposition (SVD) của ma trận hiệp phương sai.
- 7. Thành phần chính là gì trong PCA?**
 - Thành phần chính là các hướng trong không gian dữ liệu mà biểu diễn phương sai lớn nhất.
- 8. Cách chọn số lượng thành phần chính trong PCA là gì?**
 - Số lượng thành phần chính thường được chọn sao cho tỷ lệ phương sai được giữ lại đạt một ngưỡng nhất định, ví dụ như 95% hoặc 99%.
- 9. Tại sao cần lựa chọn số lượng thành phần chính thích hợp trong PCA?**
 - Lựa chọn số lượng thành phần chính thích hợp giúp giảm chiều dữ liệu một cách hiệu quả mà vẫn giữ lại phần lớn thông tin của dữ liệu.
- 10. PCA có thể được sử dụng để giảm overfitting trong mô hình học máy không?**
 - Có, PCA có thể được sử dụng để giảm overfitting trong mô hình học máy bằng cách giảm số lượng biến đầu vào.
- 11. Sự khác biệt giữa PCA và t-SNE là gì?**
 - PCA là một phương pháp giảm chiều dữ liệu tuyến tính trong khi t-SNE là một phương pháp giảm chiều dữ liệu phi tuyến tính, thích hợp cho việc biểu diễn dữ liệu trong không gian có cấu trúc phức tạp hơn.
- 12. PCA có thể được sử dụng cho dữ liệu có tính chất dạng văn bản không?**

- Có, PCA có thể được sử dụng cho dữ liệu có tính chất dạng văn bản bằng cách biểu diễn các từ dưới dạng vector (ví dụ: TF-IDF) và áp dụng PCA lên các vector đó.
- PCA có thể được sử dụng cho dữ liệu có giá trị bị thiếu không?**
 - Có, PCA có thể được sử dụng cho dữ liệu có giá trị bị thiếu sau khi thực hiện tiền xử lý dữ liệu để điền vào các giá trị bị thiếu.
 - Tại sao PCA không phù hợp cho dữ liệu có tính chất dạng cây (hierarchical data)?**
 - PCA không phù hợp cho dữ liệu có tính chất dạng cây vì nó giả định rằng dữ liệu là tuyến tính và không thể biểu diễn được cấu trúc phân cấp trong dữ liệu.
 - PCA có ổn định không khi số lượng quan sát nhỏ so với số lượng biến?**
 - PCA có thể không ổn định khi số lượng quan sát nhỏ so với số lượng biến, vì nó có thể dẫn đến overfitting.
 - Cách đánh giá hiệu suất của PCA là gì?**
 - Hiệu suất của PCA thường được đánh giá bằng cách so sánh tỷ lệ phương sai được giữ lại giữa dữ liệu ban đầu và dữ liệu đã giảm chiều.
 - Sự khác biệt giữa PCA và LDA (Linear Discriminant Analysis) là gì?**
 - PCA là một phương pháp giảm chiều dữ liệu không giám sát trong khi LDA là một phương pháp giảm chiều dữ liệu giám sát, thường được sử dụng cho bài toán phân loại.
 - PCA có thể được sử dụng để giảm nhiễu trong dữ liệu không?**
 - Có, PCA có thể được sử dụng để giảm nhiễu trong dữ liệu bằng cách loại bỏ các thành phần chính ít quan trọng hoặc nhiễu.
 - PCA có thể được sử dụng cho dữ liệu có biến phụ thuộc (dependent variables) không?**
 - Không, PCA chỉ được sử dụng cho dữ liệu có các biến độc lập (independent variables) để giảm chiều dữ liệu.
 - Có bao nhiêu cách để biểu diễn kết quả của PCA sau khi giảm chiều dữ liệu?**
 - Kết quả của PCA sau khi giảm chiều dữ liệu có thể được biểu diễn bằng các biểu đồ phân tán 2D hoặc 3D để hiểu được sự phân phối của các quan sát trong không gian mới.

Dưới đây là 20 câu hỏi khác về Phân tích thành phần chính (PCA):

- PCA có thể được áp dụng cho dữ liệu ảnh không? Nếu có, cách thức áp dụng là gì?**
 - Có, PCA có thể được áp dụng cho dữ liệu ảnh bằng cách biểu diễn mỗi ảnh dưới dạng vector và sau đó áp dụng PCA lên các vector đó.
- Tại sao PCA được gọi là Phân tích thành phần chính?**
 - PCA được gọi là Phân tích thành phần chính vì nó giúp tìm ra các thành phần chính chứa phần lớn thông tin của dữ liệu.
- Sự khác biệt giữa PCA và SVD (Singular Value Decomposition) là gì?**
 - PCA và SVD đều được sử dụng để giảm chiều dữ liệu nhưng PCA giả định rằng dữ liệu đã được chuẩn hóa, trong khi SVD không có giả định này.
- Tại sao cần chuẩn hóa dữ liệu trước khi áp dụng PCA?**
 - Chuẩn hóa dữ liệu trước khi áp dụng PCA giúp đảm bảo rằng mỗi biến có cùng độ lớn và trọng số trong quá trình phân tích.
- Phương pháp nào được sử dụng để tính ma trận hiệp phương sai trong PCA?**
 - Ma trận hiệp phương sai thường được tính bằng cách tính ma trận covariance của dữ liệu.
- Tại sao PCA được gọi là một phương pháp giảm chiều dữ liệu không giám sát?**
 - PCA được gọi là một phương pháp giảm chiều dữ liệu không giám sát vì nó không cần thông tin về nhãn của các quan sát.

27. **Cách lựa chọn số lượng thành phần chính trong PCA là gì?**
- Số lượng thành phần chính thường được chọn dựa trên tỷ lệ phương sai được giữ lại, ví dụ như giữ lại 95% hoặc 99% phương sai của dữ liệu.
28. **Sự khác biệt giữa PCA và t-SNE là gì?**
- PCA là một phương pháp giảm chiều dữ liệu tuyến tính trong khi t-SNE là một phương pháp giảm chiều dữ liệu phi tuyến tính.
29. **Tại sao cần giảm chiều dữ liệu?**
- Giảm chiều dữ liệu giúp giảm độ phức tạp của mô hình và tăng tốc độ tính toán.
30. **PCA có ứng dụng trong lĩnh vực nào?**
- PCA có ứng dụng trong nhiều lĩnh vực như xử lý ảnh, nhận dạng khuôn mặt, nhận dạng chữ viết, sinh học, kinh tế học, v.v.
31. **Tại sao PCA không phù hợp cho dữ liệu có tính chất dạng cây (hierarchical data)?**
- PCA không phù hợp cho dữ liệu có tính chất dạng cây vì nó giả định rằng dữ liệu là tuyến tính và không thể biểu diễn được cấu trúc phân cấp trong dữ liệu.
32. **PCA có ổn định không khi số lượng quan sát nhỏ so với số lượng biến?**
- PCA có thể không ổn định khi số lượng quan sát nhỏ so với số lượng biến, vì nó có thể dẫn đến overfitting.
33. **Cách đánh giá hiệu suất của PCA là gì?**
- Hiệu suất của PCA thường được đánh giá bằng cách so sánh tỷ lệ phương sai được giữ lại giữa dữ liệu ban đầu và dữ liệu đã giảm chiều.
34. **Sự khác biệt giữa PCA và LDA (Linear Discriminant Analysis) là gì?**
- PCA là một phương pháp giảm chiều dữ liệu không giám sát trong khi LDA là một phương pháp giảm chiều dữ liệu giám sát.
35. **PCA có thể được sử dụng để giảm overfitting trong mô hình học máy không?**
- Có, PCA có thể được sử dụng để giảm overfitting trong mô hình học máy bằng cách giảm số lượng biến đầu vào.
36. **PCA có thể được sử dụng để giảm nhiễu trong dữ liệu không?**
- Có, PCA có thể được sử dụng để giảm nhiễu trong dữ liệu bằng cách loại bỏ các thành phần chính ít quan trọng hoặc nhiễu.
37. **PCA có thể được sử dụng cho dữ liệu có biến phụ thuộc (dependent variables) không?**
- Không, PCA chỉ được sử dụng cho dữ liệu có các biến độc lập (independent variables) để giảm chiều dữ liệu.
38. **Có bao nhiêu cách để biểu diễn kết quả của PCA sau khi giảm chiều dữ liệu?**
- Kết quả của PCA sau khi giảm chiều dữ liệu có thể được biểu diễn bằng các biểu đồ phân tán 2D hoặc 3D để hiểu được sự phân phối của các quan sát trong không gian mới.
39. **Tại sao cần sử dụng phân tích thành phần chính?**
- Phân tích thành phần chính giúp giảm chiều dữ liệu một cách hiệu quả và tìm ra các biểu diễn quan trọng của dữ liệu.