Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh

TRUNG TÂM TIN HỌC

Big Data in Machine Learning

Bài 13: Unsupervised Learning – *PCA(Principal Component Analysis)*

Ngành LT & CSDL

https://csc.edu.vn/lap-trinh-va-csdl/Big-Data-in-Machine-Learning_198

2020

Nội dung

1. **Dimensionality Reduction**
2. Giới thiệu PCA
3. Xây dựng PCA

Dimensionality Reduction

- Nhiều vấn đề về Machine Learning bao gồm hàng nghìn hoặc thậm chí hàng triệu tính năng/ thuộc tính cho mỗi cá thể huấn luyện. Điều này không chỉ làm cho việc đào tạo cực kỳ chậm, mà còn có thể làm cho việc thực hiện trở nên khó hơn nhiều để tìm ra một giải pháp tốt.

- May mắn thay, trong các vấn đề thực tế, có thể làm giảm số lượng các tính năng đáng kể, chuyển một vấn đề phức tạp thành một vấn đề có thể xử lý được.

Dimensionality Reduction

- Ví dụ: Hãy xem xét các hình ảnh trong dữ liệu MNIST: các điểm ảnh trên viền ảnh hầu như luôn luôn màu trắng, vì vậy ta có thể loại bỏ hoàn toàn các pixel này từ tập huấn luyện mà không mất nhiều thông tin

- Ngoài việc tăng tốc đào tạo, việc giảm kích thước cũng vô cùng hữu ích cho việc trực quan hóa dữ liệu. Giảm số lượng kích thước xuống còn hai chiều (hoặc ba chiều) giúp có thể vẽ một tập huấn luyện nhiều chiều (highdimensional training) trên biểu đồ và thường thu được một số thông tin quan trọng bằng cách phát hiện các mẫu trực quan, chẳng hạn như các cụm.

Dimensionality Reduction

❑ Ghi chú

- Giảm kích thước sẽ làm mất một số thông tin (giống như nén một hình ảnh sang JPEG có thể làm giảm chất lượng của nó), vì vậy mặc dù nó sẽ tăng tốc độ đào tạo, nó cũng có thể làm cho hệ thống của ta hoạt động kém hơn một chút.
- Vì vậy, ta nên cố gắng đào tạo hệ thống với dữ liệu gốc trước khi cân nhắc việc sử dụng giảm kích thước nếu đào tạo quá chậm.
- Tuy nhiên, trong một vài trường hợp, việc giảm kích thước của dữ liệu đào tạo có thể lọc ra một số nhiễu và chi tiết không cần thiết và do đó dẫn đến hiệu suất cao hơn (thường là không!!!; nó sẽ chỉ tăng tốc độ đào tạo)

Dimensionality Reduction

❑ Tại sao phải giảm kích thước?

- Khám phá các mối tương quan / chủ đề ẩn (vd: Các từ thường đi cùng nhau)
- Loại bỏ các tính năng dư thừa và nhiễu vì không phải tất cả các tính năng đều hữu ích
- Giải thích và trực quan hóa
- Lưu trữ và xử lý dữ liệu dễ dàng hơn

Nội dung

1. Dimensionality Reduction
2. **Giới thiệu PCA**
3. Xây dựng PCA

Giới thiệu PCA

- **PCA (*Principal Component Analysis*), *phân tích thành phần chính*, là một thuật toán dimensionality reduction) phổ biến nhất hiện nay**

- **Thuộc nhóm: Unsupervised Learning**

Giới thiệu PCA

- Kích thước tuyệt đối của dữ liệu hiện nay không chỉ là một thách thức đối với phần cứng máy tính mà còn là một nút cổ chai chính cho hiệu suất của nhiều thuật toán Machine Learning.

- Mục tiêu chính của phân tích PCA là xác định các mẫu trong dữ liệu; PCA giúp phát hiện môi tương quan giữa các biến. Nếu có mối tương quan chặt chẽ giữa các biến tồn tại, nỗ lực giảm kích thước mới có ý nghĩa.

- PDA thực hiện việc tìm các hướng của phương sai tối đa (maximum variance) trong dữ liệu high-dimensional và chuyển vào một không gian con có chiều nhỏ hơn và giữ lại hầu hết các thông tin.

Giới thiệu PCA

- PCA
  - Là một công cụ giảm kích thước có thể được sử dụng để giảm một tập hợp lớn các biến thành một tập nhỏ mà vẫn chứa hầu hết thông tin trong tập hợp lớn.
    - Là một thủ tục toán học biến đổi một số (có thể) các biến tương quan thành một số (nhỏ hơn) các biến không tương quan được gọi là các thành phần chính.
    - Thành phần chính đầu tiên chiếm phần lớn dữ liệu có thể thay đổi và mỗi thành phần kế tiếp chiếm phần lớn các biến còn lại càng tốt.

Giới thiệu PCA

- PCA là phương pháp nén giảm kích thước hoặc phương pháp nén dữ liệu. Mục tiêu là giảm kích thước nhưng không đảm bảo rằng các chiều có thể diễn giải được
- Để chọn một tập con của các biến từ một tập lớn hơn, dựa trên các biến ban đầu có mối tương quan cao nhất với thành phần chính

Các ứng dụng

- Trực quan hóa dữ liệu
- Nếu thuật toán Machine Learning quá chậm vì kích thước đầu vào quá cao, thì việc sử dụng PCA để tăng tốc là một lựa chọn hợp lý
- Nếu bộ nhớ hoặc dung lượng ổ đĩa bị giới hạn, PCA cho phép tiết kiệm không gian để đổi lấy một chút thông tin của dữ liệu. Đây có thể là một sự cân bằng hợp lý.

Ưu/ khuyết điểm

❑ Ưu điểm

- Giảm kích thước, tăng tốc độ
- Trực quan hóa dữ liệu

Ưu/ khuyết điểm

❑ Khuyết điểm

- PCA không thể chỉnh hằng số (scale invariant).
- Các hướng có phương sai lớn nhất được giả định là quan trọng nhất
- Chỉ xem xét các phép biến đổi trực giao (các phép quay) của các biến gốc
- PCA chỉ dựa trên vectơ trung bình và ma trận hiệp phương sai. Một số phân phối (multivariate normal) đặc trưng bởi điều này, nhưng một số khác thì không.
- Nếu có các biến tương quan, PCA có thể giảm kích thước. Nếu không, PCA không giảm được kích thước.

Nội dung

1. Dimensionality Reduction
2. Giới thiệu PCA
3. **Xây dựng PCA**

Xây dựng PCA

- **Sử dụng pyspark.ml.feature.PCA để thực hiện:**
  - Việc trực quan hóa dữ liệu
  - Việc tăng tốc độ cho thuật toán Machine Learning

http://localhost:8888/notebooks/Chapter13/demo_PCA_visualization.ipynb#PCA---Visualization
http://localhost:8888/notebooks/Chapter13/demo_PCA_with_basic_Model.ipynb