1. Giới thiệu

Với thời đại dữ liệu bùng nổ như ngày nay, dữ liệu ta thu thập được rất lớn. Trong thực tế, các**vector đặc trưng**(*vectơ đặc trưng*) có thể có số chiều rất lớn, tới vài nghìn. Đồng thời, lượng điểm dữ liệu cũng rất lớn. Điều đó sẽ gây khó khăn cho việc lưu trữ và tính toán. Vì vậy, một trong những bước quan trọng trong nhiều bài toán học máy là ta phải**giảm chiều dữ liệu**(*giảm độ mờ*).

Bạn đang xem: Principal component analysis là gì

Giảm chiều dữ liệucòn là phương pháp được sử dụng để giảm vấn đề**quá khớp**(*quá mức*),nó có hai hướng là hướng**lựa chọn đặc trưng**(*lựa chọn tính năng*) và hướng**trích xuất đặc trưng**(*khai thác tính năng*). Hôm nay ta sẽ tìm hiểu về một thuật toán theo hướng**trích xuất đặc trưng**là Phân tích Thành phần Chính (PCA).

Xem thêm: Đánh Giá Camera A9 Pro Từ Người Đã Mua, Đánh Giá Galaxy A9 Pro: Tạm Quên Nỗi Lo Hết Pin

1.1 Giới thiệu về PCA

1.2 Cở sở toán học

Kỳ vọng (mean)

Là giá trị mong muốn, nó đơn giản là trung bình cộng của toàn bộ các giá trị

ChoNgiá trị

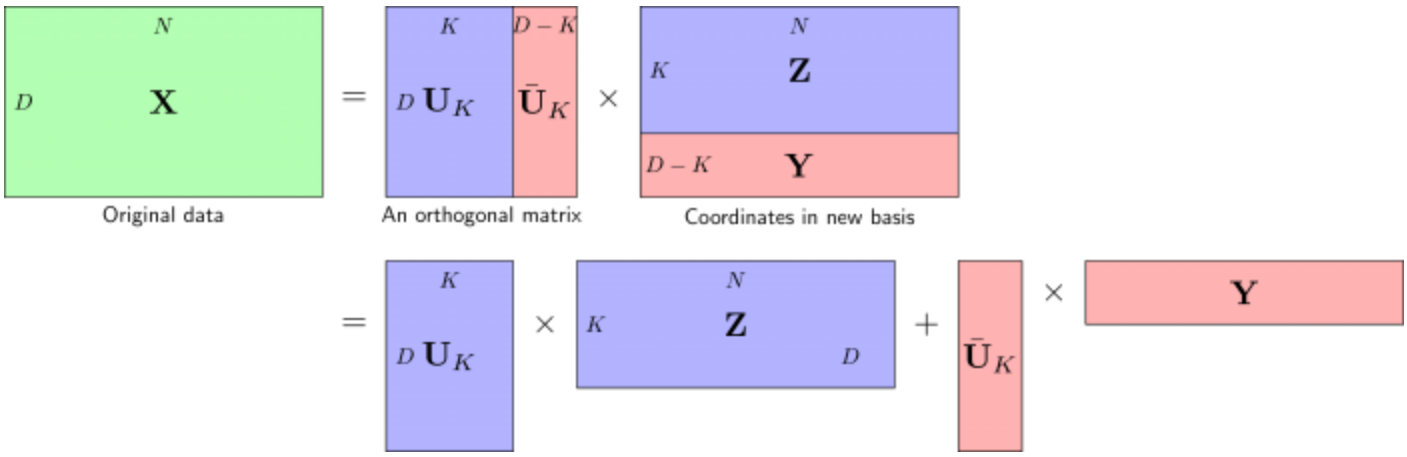
$x_1, x_2, \ldots, x_N$ $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ $COV(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N}$ $x_1, \ldots, x_N$

$S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N}\hat{X}\hat{X}^T$ $S_{ij}, i \neq j$

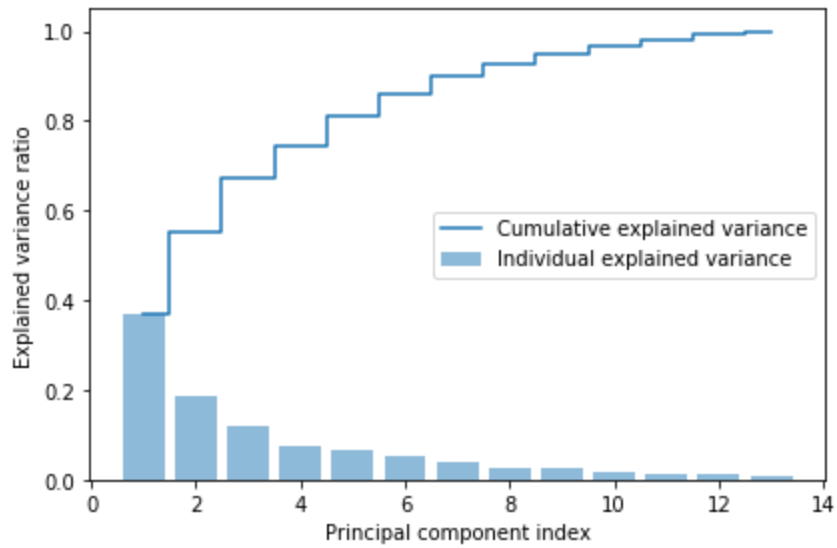
$S = \begin{bmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{bmatrix}$ $A \in R^{n\times n}$ $x \neq 0 \in R^n$ $Ax = \lambda x$ x

$kx, k \neq 0$ λ_i $(A - \lambda I)u = 0$ $S = \frac{1}{N}\hat{X}\hat{X}^T$ u $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ $u_1, u_2, \ldots, u_n$ $\hat{X}$

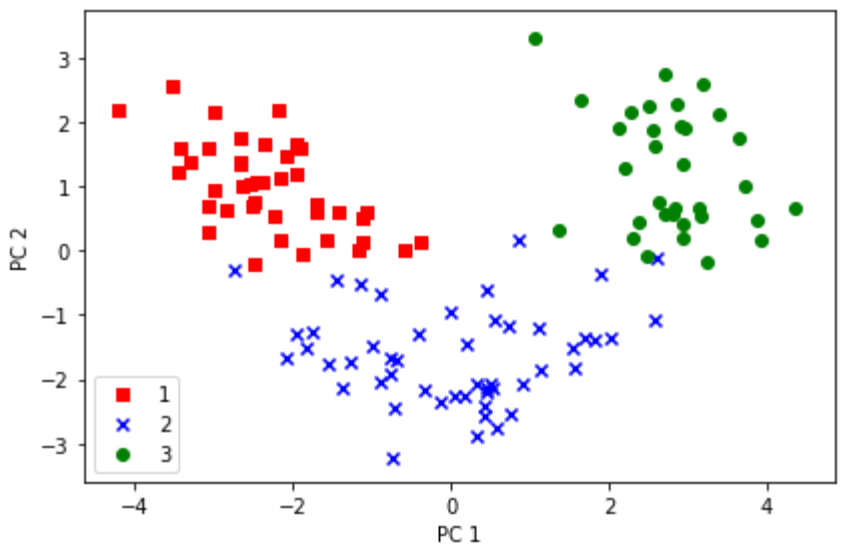
$Z = U^T\hat{X}$



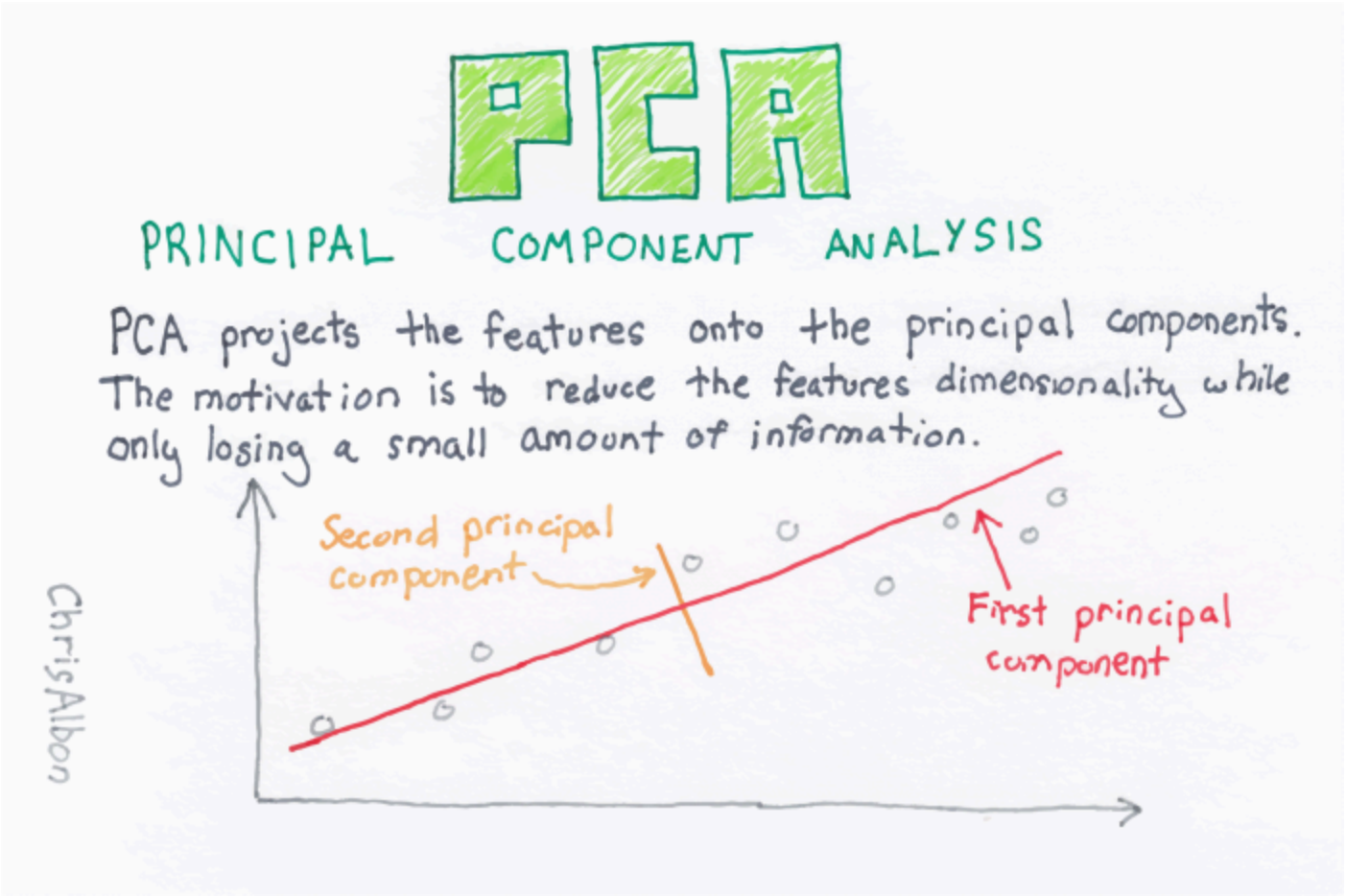
3. Demo đơn giản $\text{variance ratio} = \frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}$



D $K \leq D$ $x' = xU$



$X' = XU$



Dữ liệu được trải rộng hơn ở trục PC 1 – thành phần chính đầu tiên (the first principal component)Dữ liệu được trải rộng it hơn ở trục PC 2 – thành phần chính thứ hai (the second principal component)***Phần này ta áp dụng thư viện sklearn để giảm chiều dữ liệu và training***

từ matplotlib.colors nhập khẩu ListedColormap**phản đối** plot_decision_regions (X, y, phân loại, độ phân giải = 0.02): *# thiết lập trình tạo điểm đánh dấu và bản đồ màu* markers = (“s”, “x”, “o”, “^”, “v”) color = (“red”, “blue”, “lightgreen”, “gray”, “cyan”) cmap = ListedColormap (màu sắc plot_decision_regions (X_train_pca, y_train, classifier = lr) plt.xlabel (“PC 1”) plt.ylabel (“PC 2”) plt.legend (loc = “phía dưới bên trái”) plt.tight_layout () plt.show ()

Phân loại dữ liệu trên tập test chỉ với dữ liệu 2 chiều

plot_decision_regions (X_test_pca, y_test, classifier = lr) plt.xlabel (“PC1”) plt.ylabel (“PC2”) plt.legend (loc = “phía dưới bên trái”) plt.tight_layout () plt.show ()

Ta có thể thấy chỉ có một vài điểm bị phân loại sai, logistic regression hoạt động khá tốt trên không gian hai chiều này.

Xem thêm: Bà Bầu Ăn Mít Thường Xuyên Có Tốt Không, Ăn Mít Có Nóng Không

4. Kết luận 4.1 Ưu điểm của PCA

Loại bỏ các đặc trưng tương quan (giảm các đặc trưng)Cải thiện hiệu suất thuật toánGiảm quá khớp (*quá mức*)Cải thiện trực quan hóa dữ liệu (dễ trực quan hóa khi có ít chiều)

4.2 Nhược điểm của PCA

Các biến độc lập trở nên khó hiểu hơn (các đặc trưng mới không dễ đọc và dễ hiểu như các đặc trưng ban đầu).Chuẩn hóa dữ liệu trước khi sử dụng PCA.Mất thông tin.

5. Tài liệu tham khảo
Sách học máy Python
Chuyên mục: Review tổng hợp
The post Giới Thiệu Principal Component Analysis Là Gì, Phân Tích Thành Phần Chính (Pca) first appeared on CALLOFDUTYMOBILEPC.COM.

via CALLOFDUTYMOBILEPC.COM
https://callofdutymobilepc.com/gioi-thieu-principal-component-analysis-la-gi-phan-tich-thanh-phan-chinh-pca/