Nama : Adifa Syahira
Nim : 1103202067
Kelas : TK 44-G4

TUGAS 5 MACHINE LEARNING

Penjelasan mengenai PCA,LDA, dan SVD

Jawab :

dataset Iris dan dua metode analisis data, yaitu Principal Component Analysis (PCA) dan Linear Discriminant Analysis (LDA), yang dapat diterapkan pada dataset ini:

1. **Iris Dataset**:
    - Dataset Iris mewakili tiga jenis bunga Iris: Setosa, Versicolour, dan Virginica.
    - Masing-masing bunga diukur berdasarkan empat atribut: panjang kelopak (sepal length), lebar kelopak (sepal width), panjang mahkota (petal length), dan lebar mahkota (petal width).
    - Data ini digunakan untuk mengidentifikasi perbedaan dan pola antara tiga jenis bunga Iris berdasarkan atribut-atribut ini.
2. **Principal Component Analysis (PCA)**:
    - PCA adalah metode analisis yang digunakan untuk mengurangi dimensi data.
    - Tujuannya adalah untuk mengidentifikasi kombinasi atribut (komponen utama) yang menyumbang sebagian besar variasi dalam data.
    - Pada dataset Iris, PCA digunakan untuk mengidentifikasi kombinasi atribut yang paling penting dalam menjelaskan variasi antara sampel data. Dua komponen utama pertama yang dihasilkan adalah yang paling signifikan dalam menjelaskan variasi dalam data.
    - Kemudian, sampel data direpresentasikan dalam ruang yang terdiri dari dua komponen utama ini.
3. **Linear Discriminant Analysis (LDA)**:
    - LDA adalah metode analisis yang berusaha mengidentifikasi atribut-atribut yang menjelaskan sebagian besar variasi antara kelas atau grup data.
    - Berbeda dengan PCA, LDA adalah metode yang diawasi yang menggunakan label kelas yang diketahui.
    - Pada dataset Iris, LDA digunakan untuk mengidentifikasi atribut-atribut yang paling signifikan dalam memisahkan tiga jenis bunga Iris (Setosa, Versicolour, dan Virginica).
    - LDA mencoba menemukan proyeksi atribut yang memaksimalkan pemisahan antara kelas dengan mempertimbangkan label kelas dari setiap sampel.

Dengan menggunakan PCA, Anda mendapatkan pemahaman tentang variasi dalam data yang tidak memperhatikan label kelas. Di sisi lain, LDA berfokus pada perbedaan antara kelas dan mencoba mengidentifikasi atribut-atribut yang paling relevan untuk membedakan kelas. Kedua metode ini dapat membantu Anda dalam analisis data dan pengurangan dimensi dengan tujuan yang berbeda.

4. **IDF**

   Ini menggambarkan sebuah contoh penggunaan API scikit-learn untuk mengelompokkan dokumen berdasarkan topik mereka menggunakan pendekatan Bag of Words. Pendekatan Bag of Words mewakili setiap dokumen sebagai koleksi kata-kata, mengabaikan tata bahasa dan urutan kata.Contoh ini mendemonstrasikan dua algoritma pengelompokan: KMeans dan MiniBatchKMeans. KMeans adalah algoritma pengelompokan tradisional, sementara MiniBatchKMeans adalah varian yang lebih skalabel yang dapat menangani dataset besar dengan lebih efisien. Selain pengelompokan, contoh ini juga menerapkan analisis semantik laten (LSA) untuk mengurangi dimensi data dan menemukan pola-pola laten. LSA adalah teknik yang mengungkapkan hubungan tersembunyi antara istilah dan dokumen dengan menganalisis matriks istilah-dokumen.

   Contoh ini menggunakan dua vektorisasi teks yang berbeda: TfidfVectorizer dan HashingVectorizer. TfidfVectorizer menghitung nilai Term Frequency-Inverse Document Frequency (TF-IDF) untuk setiap istilah dalam dokumen, yang mengukur pentingnya suatu istilah dalam sebuah dokumen. HashingVectorizer, di sisi lain, menggunakan trik hashing untuk mengonversi teks menjadi representasi vektor berpanjang tetap.Jika Anda tertarik untuk membandingkan kinerja berbagai vektorisasi dan waktu pemrosesan mereka, Anda dapat merujuk ke notebook contoh yang disebut "FeatureHasher and DictVectorizer Comparison."
   Terakhir, jika Anda ingin menganalisis dokumen dengan pendekatan pembelajaran terawasi, ada skrip contoh yang tersedia untuk mengklasifikasikan dokumen teks menggunakan fitur yang sangat terurai. Pendekatan ini melibatkan pelatihan model pembelajaran mesin pada data berlabel untuk memprediksi topik atau kategori dari dokumen yang diberikan.

   Dalam konteks ini, penjelasan menyatakan bahwa data yang digunakan dalam contoh berasal dari "The 20 newsgroups text dataset." Dataset ini terdiri dari sekitar 18.000 posting dari berbagai grup berita yang mencakup 20 topik berbeda. Namun, untuk tujuan ilustrasi dan untuk mengurangi kompleksitas komputasi, hanya dipilih subset 4 topik. Subset ini terdiri dari sekitar 3.400 dokumen.
   Untuk memahami bagaimana topik-topik ini tumpang tindih, Anda dapat merujuk ke contoh yang disebut "Classification of text documents using sparse features."
   Penting untuk dicatat bahwa sampel teks dalam dataset ini mengandung informasi tambahan seperti "headers" (kepala), "footers" (tanda tangan), dan "quotes" (petikan) dari posting lain.

Namun, untuk membuat masalah pengelompokan yang lebih fokus, parameter "remove" digunakan dalam fungsi fetch_20newsgroups untuk menghapus fitur-fitur tambahan ini dari sampel teks. Dengan menghapus informasi tambahan ini, fokus ditempatkan pada teks inti yang berkaitan dengan topik-topik berita, yang membuat masalah pengelompokan menjadi lebih bersih dan lebih terfokus pada kontennya.