Nama : Michael Christopher
NIM : 1103210260
Analisis UTS ML Soal #3

Analisis Detail Teknik dan Tantangan Clustering

1. Inkonsistensi Antara Metode Elbow dan Silhouette Score pada K-Means

Inkonsistensi antara nilai silhouette score yang rendah (0,3) dan metode elbow yang menunjukkan K=5 sebagai nilai optimal berasal dari keterbatasan mendasar dalam pendekatan validasi ini. Metode elbow terutama mengukur pengurangan variansi dalam cluster (inertia), yang dioptimalkan oleh K-means. Namun, hal ini tidak selalu menghasilkan cluster yang terpisah dengan baik yang diukur oleh silhouette score.

Perbedaan ini biasanya terjadi karena:

- **Bentuk cluster non-spherical**: K-means mengasumsikan cluster berbentuk bulat (spherical) dengan variansi yang mirip, tetapi data transaksi retail seringkali membentuk pola memanjang dan tidak beraturan berdasarkan pola pembelian. K-means akan membagi struktur alami ini secara artifisial, menciptakan batas-batas yang tidak alami sehingga menghasilkan silhouette score yang rendah.
- **Variasi kepadatan cluster**: Jika beberapa segmen pelanggan lebih terkelompok dengan rapat sementara yang lain lebih tersebar (misalnya, pembeli rutin vs pembeli sesekali), K-means akan kesulitan menentukan batas yang tepat, terutama jika ukuran cluster tidak seimbang.
- **Struktur korelasi fitur**: Jarak Euclidean yang digunakan K-means menimbang semua dimensi secara sama, mengabaikan korelasi antara fitur seperti kuantitas dan total pengeluaran.

Strategi validasi alternatif yang dapat mengatasi masalah ini meliputi:

- **Analisis gap statistic**: Membandingkan dispersi dalam cluster dengan yang diharapkan dalam distribusi referensi null, mengidentifikasi titik di mana penambahan cluster tidak lagi memberikan peningkatan yang signifikan secara statistik. Tidak seperti metode elbow, teknik ini kurang sensitif terhadap asumsi bentuk cluster.
- **Validasi stabilitas melalui bootstrapping**: Dengan mengambil sampel ulang data secara berulang dan melakukan clustering, kita dapat menilai stabilitas penempatan cluster. Penempatan yang konsisten di seluruh sampel bootstrap menunjukkan cluster yang kuat terlepas dari bentuknya, memberikan validasi di luar metrik jarak sederhana.
- **Consensus clustering**: Menggabungkan hasil dari beberapa algoritma clustering dapat mengungkapkan struktur dasar yang stabil yang mungkin terlewatkan oleh metode individual, terutama dengan data non-spherical.

Untuk data transaksi retail, implementasi pendekatan validasi alternatif ini kemungkinan akan mengungkapkan bahwa K optimal berbeda dari yang disarankan oleh metode elbow, atau bahwa algoritma non-K-means (seperti DBSCAN atau Spectral Clustering) akan lebih tepat untuk menangkap segmen pelanggan alami.

2. Preprocessing Tipe Data Campuran untuk Clustering

Preprocessing yang efektif untuk dataset retail dengan fitur numerik (Quantity, UnitPrice) dan fitur kategorikal berkardinali tinggi (Description) memerlukan penanganan yang cermat terhadap skala dan representasi yang berbeda:

Untuk fitur numerik:

- **Robust scaling**: Menggunakan median dan IQR (interquartile range) alih-alih standard scaling lebih baik menangani outlier yang umum dalam data transaksi, di mana beberapa pembelian besar dapat membuat distribusi menjadi miring.
- **Transformasi log**: Menerapkan transformasi logaritma pada fitur seperti UnitPrice atau Quantity dapat menormalkan distribusi yang miring ke kanan yang tipikal dalam data retail.

Untuk fitur kategorikal berkardinali tinggi seperti Description (yang bisa memiliki ribuan nilai unik):

Risiko One-Hot Encoding:

- **Curse of dimensionality**: Dengan ribuan deskripsi produk, one-hot encoding menciptakan ruang berdimensi tinggi yang sangat jarang (sparse) di mana metrik jarak menjadi tidak bermakna.
- **Inefisiensi komputasi**: Matriks yang dihasilkan menjadi sulit dikelola, memperlambat algoritma clustering dan berpotensi menyebabkan masalah memori.
- **Kehilangan kesamaan semantik**: Produk serupa diperlakukan sebagai dimensi yang sepenuhnya berbeda, kehilangan hubungan inheren antara, misalnya, "RED HANGING HEART" dan "WHITE HANGING HEART".

Alternatif yang lebih robust meliputi:

- **Representasi TF-IDF**: Ini memberikan bobot pada deskripsi produk berdasarkan frekuensinya dalam transaksi dan frekuensi inverse di seluruh katalog, menangkap kepentingan relatif dari produk yang berbeda dalam pola pembelian pelanggan.
- **Embedding berdimensi rendah**: Teknik seperti UMAP mempertahankan hubungan lokal sambil secara drastis mengurangi dimensi, memungkinkan algoritma clustering mengidentifikasi pengelompokan alami dalam ruang yang lebih mudah dikelola.
- **Category embedding**: Untuk deskripsi produk, model bahasa pra-terlatih dapat menghasilkan embedding semantik yang menangkap hubungan produk, memungkinkan produk serupa diposisikan lebih dekat dalam ruang fitur.
- **Feature agglomeration**: Melakukan clustering hierarkis pada fitur itu sendiri sebelum clustering pelanggan dapat mengurangi dimensi sambil mempertahankan struktur informasi yang bermakna.

Pendekatan ini mengatasi tantangan mendasar dengan mengubah ruang kategorikal berdimensi tinggi menjadi representasi berdimensi lebih rendah yang padat (dense) sambil mempertahankan struktur semantik yang relevan untuk segmentasi pelanggan, menghasilkan cluster yang lebih koheren dan dapat diinterpretasi.

3. Pemilihan Parameter Adaptif untuk DBSCAN

Sensitivitas DBSCAN terhadap parameter epsilon menjadi tantangan khusus dengan data transaksi retail yang tidak seimbang (seperti 90% pelanggan dari UK). Menentukan nilai epsilon optimal secara adaptif memerlukan pemahaman tentang distribusi kepadatan dataset:

Pendekatan k-distance graph:

1. Menghitung jarak ke tetangga ke-k untuk setiap titik
2. Mengurutkan jarak ini dan memplotnya
3. Mencari "elbow" atau titik infleksi di mana kurva meningkat tajam

Peran k-distance graph sangat penting karena mengungkapkan distribusi kepadatan di seluruh dataset. Peningkatan tajam menunjukkan transisi dari cluster padat ke wilayah jarang atau noise. Dalam data retail, ini sering memisahkan pelanggan reguler dari outlier atau pembeli satu kali.

Pendekatan kuartil ketiga (persentil ke-75) menyediakan cara otomatis dan statistik robust untuk menetapkan epsilon:

- Menggunakan Q3 dari distribusi k-distance sebagai epsilon memastikan bahwa sekitar 75% titik memiliki k-neighborhood dalam radius ini
- Ini secara adaptif memperhitungkan pola kepadatan keseluruhan tanpa memerlukan inspeksi manual

Namun, dengan data yang tidak seimbang secara regional (90% pelanggan UK), nilai epsilon global menjadi problematik karena:

- Wilayah dominan (UK) akan mendikte profil kepadatan
- Wilayah minoritas mungkin memiliki pola kepadatan alami yang berbeda

Adaptasi kepadatan regional oleh karena itu sangat penting:

- **Pemilihan parameter stratifikasi**: Menghitung nilai epsilon terpisah untuk setiap negara atau wilayah
- **Penyesuaian MinPts berbasis kepadatan**: Meningkatkan MinPts untuk wilayah yang lebih padat (UK) dan menurunkan untuk wilayah yang lebih jarang
- **Algoritma OPTICS**: Pertimbangkan menggunakan OPTICS alih-alih DBSCAN, karena bekerja dengan rentang nilai epsilon daripada nilai global tunggal

Pendekatan optimal adalah menerapkan penskalaan berbasis kepadatan di mana:

1. Mengelompokkan data berdasarkan wilayah (negara)
2. Menghitung distribusi k-distance terpisah untuk setiap wilayah
3. Memilih nilai epsilon spesifik wilayah pada persentil yang sesuai

4. Menskalakan MinPts berdasarkan kepadatan relatif setiap wilayah (lebih tinggi untuk wilayah lebih padat)
5. Menerapkan pendekatan hierarkis: clustering global diikuti oleh penyempurnaan spesifik wilayah

Strategi adaptif ini mencegah basis pelanggan UK yang dominan dari menutupi pola menarik di segmen regional yang lebih kecil, memungkinkan DBSCAN menemukan cluster yang bermakna di seluruh profil kepadatan yang bervariasi.

4. Mengatasi Tumpang Tindih Cluster dengan Teknik Lanjutan

Ketika analisis pasca-clustering mengungkapkan tumpang tindih signifikan antara "high-value customers" dan "bulk buyers" berdasarkan total pengeluaran, tantangannya terletak pada membedakan antara perilaku agregat yang serupa dengan pola dasar yang berbeda:

Pendekatan clustering semi-supervised:

- **Constrained clustering** dengan batasan must-link dan cannot-link dapat secara efektif memasukkan pengetahuan domain. Untuk data retail, ini mungkin melibatkan spesifikasi bahwa pelanggan tertentu dengan pengeluaran serupa tetapi pola frekuensi pembelian berbeda harus berada dalam cluster terpisah:
    - Must-link: Pelanggan yang berulang kali membeli item premium dalam jumlah kecil
    - Cannot-link: Pelanggan dengan total pengeluaran serupa tetapi frekuensi pesanan yang sangat berbeda
- **Seeded K-means** dengan exemplar segmen yang diidentifikasi oleh pakar domain dapat mengarahkan algoritma menuju partisi yang relevan secara bisnis yang mungkin terlewatkan oleh pendekatan unsupervised murni.

Teknik metric learning:

- **Jarak Mahalanobis** meningkatkan pemisahan dengan memperhitungkan struktur kovarians antara fitur. Ini mengatasi keterbatasan utama jarak Euclidean, yang memperlakukan semua dimensi sebagai independen. Jarak Mahalanobis didefinisikan sebagai:
  $d(x,y) = \sqrt{[(x-y)^T S^{-1} (x-y)]}$
  di mana S adalah matriks kovarians dari fitur.
- **Supervised metric learning** dapat diterapkan dengan membuat subset kecil pelanggan berlabel, kemudian mempelajari metrik jarak yang memaksimalkan pemisahan antar-cluster sambil meminimalkan jarak dalam-cluster.

Tantangan dengan pendekatan non-Euclidean:

1. **Hambatan interpretabilitas**: Jarak non-Euclidean menciptakan batas keputusan kompleks yang sulit dijelaskan dalam istilah bisnis yang sederhana.
2. **Ketidakjelasan kepentingan fitur**: Dengan metrik jarak yang ditransformasi, kontribusi relatif fitur individual menjadi kurang transparan, membuat lebih sulit untuk mendeskripsikan cluster melalui ambang batas fitur sederhana.

3. **Kompleksitas implementasi**: Teknik lanjutan ini memerlukan penyetelan dan validasi yang cermat, meningkatkan beban teknis.

Solusi untuk mempertahankan interpretabilitas bisnis sambil menggunakan teknik lanjutan ini melibatkan:

- **Ekstraksi aturan post-hoc**: Setelah clustering dengan metrik canggih, ekstrak aturan keputusan yang disederhanakan yang mendekati batas kompleks
- **Analisis kepentingan fitur**: Mengukur kontribusi setiap fitur asli terhadap clustering akhir
- **Visualisasi berpusat pada bisnis**: Membuat penjelasan visual yang menerjemahkan perbedaan teknis ke dalam istilah yang relevan dengan bisnis
- **Pendekatan hybrid**: Menerapkan teknik clustering lanjutan terlebih dahulu, kemudian menyempurnakan menggunakan metode yang lebih dapat diinterpretasi untuk segmentasi akhir

Pendekatan seimbang ini memanfaatkan kekuatan pemisahan teknik lanjutan sambil mempertahankan relevansi bisnis dan kemampuan penjelasan yang diperlukan untuk wawasan yang dapat ditindaklanjuti.

5. Rekayasa Fitur Temporal untuk Data Transaksi

Merancang fitur temporal yang efektif dari InvoiceDate memerlukan keseimbangan antara penemuan pola dan validitas statistik. Untuk data transaksi retail, fitur-fitur ini dapat mengungkapkan pola perilaku penting:

Fitur temporal yang efektif:

Pengkodean siklikal dari hari-dalam-seminggu dan jam-dalam-hari menggunakan transformasi sinus/kosinus:
hour_sin = sin(2π * hour/24)

- hour_cos = cos(2π * hour/24)
  Ini mempertahankan hubungan siklikal di mana jam 23 dekat dengan jam 0, menghindari batas artifisial dalam representasi linier.
- **Rasio waktu pembelian**: Proporsi pembelian di pagi/siang/malam, menangkap preferensi waktu belanja pelanggan.
- **Metrik periodisitas**: Standar deviasi waktu pembelian untuk membedakan pembeli rutin (SD rendah) dari pembeli tidak teratur (SD tinggi).
- **Interval antar-pembelian**: Rata-rata dan varians waktu antara pembelian berurutan, mengungkapkan ritme dan konsistensi pembelian.

Risiko kebocoran data dengan agregasi temporal:

1. **Bias forward-looking**: Menggunakan fitur seperti "rata-rata pembelian bulanan" menciptakan kebocoran ketika informasi ini tidak akan tersedia pada waktu prediksi.
2. **Inkonsistensi jendela waktu**: Pelanggan yang berbeda memiliki panjang sejarah yang berbeda, membuat agregasi tidak konsisten di seluruh dataset.

3. **Confounding musiman**: Agregat temporal mungkin menangkap efek musiman daripada perilaku spesifik pelanggan.

Untuk mencegah masalah ini, validasi silang berbasis waktu sangat penting:

- Gunakan pembagian berbasis waktu alih-alih pengambilan sampel acak
- Pastikan semua perhitungan fitur hanya menggunakan informasi yang tersedia sebelum titik cutoff
- Validasi stabilitas cluster di berbagai periode waktu

Masalah dengan fitur lag:

Fitur lag (misalnya, jumlah pembelian dari 7 hari sebelumnya) dapat memperkenalkan noise karena:

1. **Pola pembelian tidak teratur**: Sebagian besar pelanggan retail tidak mengikuti pola mingguan yang ketat, membuat lag 7 hari berpotensi menjadi arbitrer.
2. **Masalah sparsitas**: Banyak pelanggan tidak memiliki pembelian tepat 7 hari sebelumnya, menciptakan banyak nilai yang hilang.
3. **Peluruhan korelasi temporal**: Nilai prediktif fitur lag berkurang dengan cepat seiring waktu, sering kali memperkenalkan lebih banyak noise daripada sinyal.
4. **Kebingungan musiman**: Fitur lag mungkin menangkap pola musiman daripada perilaku spesifik pelanggan.

Pendekatan yang lebih robust untuk clustering adalah menggunakan fitur temporal yang menangkap perilaku pelanggan lebih komprehensif, seperti distribusi waktu pembelian dan pola periodisitas, daripada mengandalkan nilai lag spesifik yang mungkin dipengaruhi oleh kebetulan atau noise.