

Rangkuman Materi Perkuliahan Machine learning Kelompok 21

Muhammad Shaadam haidar Yuwono¹, Muhammad labib²

1.Deskripsi singkat

1. Distribusi Gaussian (Konsep Dasar)

Distribusi Gaussian atau **Distribusi Normal** adalah pola persebaran data yang paling umum ditemukan di alam. Data dianggap "normal" jika sebagian besar nilai berkumpul di tengah (rata-rata) dan semakin sedikit saat menjauhi pusat, membentuk kurva lonceng.

- **Dua Kunci Utama:**

- **Mu (μ):** Titik tengah atau rata-rata data. Ia menentukan di mana "puncak gunung" berada.
- **Sigma (σ):** Standar deviasi yang menentukan lebar kurva. Sigma kecil berarti data sangat kompak; sigma besar berarti data sangat tersebar.

- **Pentingnya:** Banyak algoritma AI bekerja lebih cepat dan akurat jika data sudah "dinormalkan" (memiliki $\mu=0$ dan $\sigma=1$) melalui proses prapemrosesan.

2. Support Vector Machine / SVM (Algoritma Pemisah)

SVM adalah algoritma yang digunakan untuk **klasifikasi**. Tugas utamanya adalah mencari "garis pemisah" terbaik untuk membedakan dua kelompok data.

- **Prinsip Kerja:**

- **Maximum Margin:** SVM tidak hanya asal membuat garis, tetapi mencari garis yang memiliki jarak (margin) paling lebar antara kelompok A dan kelompok B agar tidak terjadi salah prediksi di masa depan.
- **Support Vectors:** Titik-titik data yang paling dekat dengan garis pemisah. Titik-titik inilah yang menjadi "penopang" model.
- **Kernel Trick:** Kemampuan SVM untuk memisahkan data yang tercampur secara kompleks dengan cara memetakan data ke dimensi yang lebih tinggi (misal: dari gambar 2D datar menjadi 3D yang memiliki kedalaman).

3. Hubungan Keduanya

Hubungan paling erat terjadi pada **Kernel RBF** di SVM. Kernel ini menggunakan prinsip distribusi Gaussian untuk menghitung seberapa dekat atau jauh hubungan antar

data. Selain itu, prapemrosesan agar data mengikuti distribusi Gaussian adalah langkah wajib sebelum menjalankan algoritma SVM agar model tidak bingung oleh skala data yang berbeda-beda.

Kesimpulan Singkat:

Jika Gaussian memberi tahu kita bagaimana data "duduk" atau tersebar, maka SVM adalah alat untuk menarik "pagar" yang paling aman untuk memisahkan data-data tersebut secara akurat.

Langkah selanjutnya yang bisa saya bantu:

Apakah kamu ingin saya membuatkan Tabel Perbandingan antara K-Means (yang berbasis jarak) dengan SVM & GMM (yang berbasis batas dan distribusi) agar kamu punya materi tambahan untuk diskusi di kelas?

2.Mindmap



3.penjelasan detail

Analisis Mendalam: Distribusi Gaussian dan Mekanisme Support Vector Machine (SVM)

Dalam dunia *Data Science*, memahami bagaimana data tersebar dan bagaimana mesin memisahkan data tersebut adalah kunci utama. Distribusi Gaussian memberikan kita pemahaman tentang karakteristik alami data, sementara SVM memberikan metodologi rekayasa untuk mengklasifikasikan data tersebut dengan presisi tinggi.

1. Distribusi Gaussian (Distribusi Normal)

Distribusi Gaussian, yang sering disebut sebagai Distribusi Normal atau kurva lonceng (*bell curve*), adalah fungsi probabilitas yang menunjukkan bahwa data cenderung berkumpul di sekitar nilai rata-rata (*mean*). Dosen sebelumnya menekankan bahwa banyak fenomena alam—seperti tinggi badan, nilai ujian, hingga *error* pada mesin—mengikuti pola ini.

A. Karakteristik Utama Gaussian

Sebuah distribusi dikatakan Gaussian jika memiliki ciri-ciri berikut:

1. **Simetris:** Sisi kiri dan kanan dari nilai rata-rata adalah cermin satu sama lain.
2. **Puncak Tunggal (Unimodal):** Memiliki satu titik tertinggi yang merupakan nilai rata-rata (*mean*), median, dan modus secara bersamaan.
3. **Asimtotik:** Ekor kurva mendekati sumbu horizontal tetapi tidak pernah benar-benar menyentuhnya.

B. Parameter Vital: μ (Mu) dan σ (Sigma)

Seperti yang dijelaskan dalam materi Pertemuan 3, identitas matematis dari sebuah distribusi Gaussian ditentukan oleh dua parameter:

- **Mean (μ):** Menentukan lokasi pusat atau puncak dari kurva. Jika μ bergeser, seluruh kurva akan bergeser ke kanan atau ke kiri pada sumbu X.
- **Standar Deviasi (σ):** Menentukan "kerampingan" kurva.
 - Jika σ kecil, data sangat rapat di sekitar rata-rata, menghasilkan gunung yang lancip.
 - Jika σ besar, data sangat tersebar, menghasilkan gunung yang lebar dan landai.

C. Pentingnya Gaussian dalam Machine Learning

Mengapa kita harus peduli dengan Gaussian?

1. **Teorema Limit Pusat:** Menyatakan bahwa jika kita mengambil sampel yang cukup besar, distribusinya akan mendekati normal, terlepas dari bentuk distribusi populasi aslinya.

2. **Optimasi Algoritma:** Banyak algoritma ML (seperti Linear Regression dan LDA) berasumsi bahwa data berdistribusi normal. Jika data tidak normal, proses prapemrosesan (seperti yang dibahas pada P2: Normalisasi) menjadi wajib dilakukan agar model bisa belajar dengan efisien.

2. Support Vector Machine (SVM)

Jika Gaussian membantu kita memahami "bentuk" data, maka **Support Vector Machine (SVM)** adalah algoritma *supervised learning* yang digunakan untuk mencari "garis pemisah" terbaik antar kelompok data. SVM sangat kuat terutama untuk dataset berdimensi tinggi.

A. Konsep Hyperplane dan Margin

Tujuan utama SVM adalah menemukan **Hyperplane** (bidang pemisah) yang optimal untuk membagi dataset ke dalam kelas-kelas yang berbeda.

- **Hyperplane:** Dalam ruang 2D, ini adalah garis. Dalam ruang 3D, ini adalah bidang datar.
- **Margin:** Jarak antara *hyperplane* dengan titik data terdekat dari masing-masing kelas. SVM bekerja dengan prinsip **Maximum Margin**, yaitu mencari garis yang memberikan jarak paling lebar antara dua kelas. Semakin lebar margin, semakin rendah risiko salah klasifikasi pada data baru.

B. Support Vectors: Titik Kritis

Mengapa dinamakan "Support Vector"? Karena hanya titik-titik data yang paling dekat dengan garis pemisah (*hyperplane*) yang menentukan posisi garis tersebut. Titik-titik kritis inilah yang disebut *Support Vectors*. Titik data lain yang jauh dari garis pemisah tidak berpengaruh pada pembentukan model. Ini membuat SVM sangat efisien dalam penggunaan memori.

C. Penanganan Data Non-Linear: Kernel Trick

Dalam dunia nyata, data jarang sekali bisa dipisahkan hanya dengan garis lurus (linear). Seringkali data tercampur secara kompleks. Di sinilah SVM menggunakan "sihir" yang disebut **Kernel Trick**.

- **Cara Kerja:** Kernel mengubah data dari dimensi rendah yang tidak bisa dipisahkan secara linear ke dimensi yang lebih tinggi di mana garis lurus (*hyperplane*) dapat ditemukan.
- **Jenis Kernel:**
 - **Linear:** Untuk data yang mudah dipisahkan garis lurus.
 - **Polynomial:** Untuk pola data yang lebih melengkung.
 - **RBF (Radial Basis Function):** Kernel yang paling populer. Menariknya, RBF memiliki kaitan erat dengan distribusi Gaussian

karena ia menggunakan fungsi basis radial yang mirip dengan kurva lonceng untuk memetakan jarak antar titik.

3. Korelasi Gaussian dan SVM dalam Analisis Data

Meskipun terlihat berbeda, terdapat benang merah antara distribusi Gaussian dan SVM, terutama saat kita berbicara tentang **RBF Kernel** pada SVM.

1. **Pengaruh Skala (Prapemrosesan):** SVM sangat sensitif terhadap skala data. Jika satu fitur memiliki rentang 1-10 (seperti nilai) dan fitur lain memiliki rentang 1jt-100jt (seperti harga), SVM akan gagal. Oleh karena itu, teknik **Min-Max Scaling** atau **Standardization** (mengubah data menjadi distribusi Gaussian dengan $\mu=0$ dan $\sigma=1$) yang dibahas pada Pertemuan 1 menjadi langkah wajib sebelum menjalankan SVM.
2. **Klasifikasi Berbasis Kerapatan:** Dalam GMM (Pertemuan 3), kita mengelompokkan data berdasarkan probabilitas Gaussian. Di sisi lain, SVM menggunakan fungsi mirip Gaussian (RBF) untuk menentukan batas keputusan berdasarkan kerapatan titik-titik *support vectors*.
3. **Ketahanan terhadap Outlier:** Distribusi Gaussian membantu kita mengidentifikasi pencilan (*outliers*). Dalam SVM, pencilan yang berada di dekat *hyperplane* bisa sangat merusak model (mengubah margin). Oleh karena itu, memahami prapemrosesan P3 (Cleaning) sangat krusial agar SVM tidak "terganggu" oleh data yang tidak valid.

4. pseudocode

```
// ALGORITMA SUPPORT VECTOR MACHINE (SVM) // Fokus:
Mencari Hyperplane Optimal dengan Margin Maksimum

BEGIN SVM_PROCESS

// 1. PENYIAPAN DATA
INPUT dataset (Features X, Labels Y)
SET Y = {-1, 1} // Label harus biner untuk klasifikasi
dasar

// 2. PRAPEMROSESAN (Sangat Penting untuk SVM)
FOR EACH feature IN X:
    APPLY Normalization (Min-Max Scaling)
    // Agar fitur dengan skala besar tidak mendominasi
    perhitungan jarak
END FOR
```

```
// 3. INISIALISASI PARAMETER
SET C = (Penalty parameter / Soft Margin)
SET Kernel_Type = "RBF" // Menggunakan fungsi dasar
Gaussian
SET Sigma ( $\sigma$ ) = Value // Parameter lebar kurva
Gaussian dalam kernel

// 4. DEFINISI KERNEL FUNCTION (RBF/Gaussian Kernel)
FUNCTION RBF_Kernel(x1, x2, sigma):
    distance = Squared_Euclidean_Distance(x1, x2)
    RETURN exp(-distance / (2 * sigma^2))
END FUNCTION

// 5. OPTIMASI (Mencari Lagrange Multipliers 'alpha')
// Tujuan: Memaksimalkan Margin  $W = \sum \alpha - 1/2 \sum \alpha_i * \alpha_j * y_i * y_j * \text{Kernel}(x_i, x_j)$ 
INITIALIZE alpha_vector = [0, 0, ..., 0]

WHILE NOT optimized (Until Karush-Kuhn-Tucker / KKT
conditions met):
    FOR EACH data_point_i, data_point_j IN dataset:
        // Gunakan metode seperti SMO (Sequential
Minimal Optimization)
        UPDATE alpha_i, alpha_j TO maximize objective
function

        // Pastikan alpha berada dalam rentang:  $0 \leq$ 
alpha  $\leq$  C
    END FOR
END WHILE

// 6. IDENTIFIKASI SUPPORT VECTORS
SUPPORT_VECTORS = Find points where alpha > 0
// Hanya titik-titik ini yang menentukan posisi pagar
pemisah (hyperplane)

// 7. MENGHITUNG BIAS (b)
CALCULATE bias (b) using Support Vectors and Alpha
values
```

```
// 8. FUNGSI PREDIKSI (Decision Function)
FUNCTION Predict(new_data_x):
    Sum = 0
    FOR EACH sv IN SUPPORT_VECTORS:
        Sum += alpha_sv * y_sv * RBF_Kernel(sv,
new_data_x, sigma)
    END FOR

    score = Sum + b

    IF score >= 0 THEN RETURN Class_1 (+1)
    ELSE RETURN Class_2 (-1)
END FUNCTION

END SVM_PROCESS
```