Part 12 - Klasifikasi dengan Support Vector Machines

Klasifikasi biner adalah tugas pembelajaran mesin di mana tujuannya adalah untuk memprediksi satu dari dua kemungkinan hasil diskrit. Untuk tugas ini, kita mencari sebuah prediktor dalam bentuk $f : \mathbb{R}^D \to +1, -1$. Seperti halnya regresi, ini adalah tugas pembelajaran terawasi di mana kita diberikan sebuah set data pelatihan yang terdiri dari pasangan contoh-label $(x_1, y_1), \ldots, (x_N, y_N)$. *Support Vector Machine* (SVM) adalah sebuah pendekatan untuk klasifikasi biner yang menggunakan intuisi geometris untuk merumuskan masalahnya, berbeda dengan pendekatan probabilistik seperti *maximum likelihood*. Pendekatan ini sangat bergantung pada konsep-konsep seperti perkalian dalam (*inner products*) dan proyeksi.

12.1 Hyperplane Pemisah

Ide utama di balik banyak algoritma klasifikasi adalah merepresentasikan data dalam ruang $\mathbb{R}^D$ dan kemudian mempartisi ruang ini. Untuk klasifikasi biner, kita membagi ruang menjadi dua bagian menggunakan sebuah *hyperplane*. Sebuah hyperplane didefinisikan sebagai himpunan titik x di mana sebuah fungsi linear-affine bernilai nol:

$$x \in \mathbb{R}^D : f(x) = 0 \quad \text{dengan} \quad f(x) := \langle w, x \rangle + b$$

di mana $w \in \mathbb{R}^D$ adalah vektor normal terhadap hyperplane dan $b \in \mathbb{R}$ adalah perpotongan (*intercept*). Untuk mengklasifikasikan sebuah contoh uji x_{test}, kita menghitung nilai $f(x_{\text{test}})$ dan mengklasifikasikannya sebagai +1 jika $f(x_{\text{test}}) \geq 0$ dan -1 jika sebaliknya. Selama pelatihan, kita ingin memastikan bahwa semua contoh data diklasifikasikan dengan benar, yang dapat dirangkum dalam satu persamaan:

$$y_n(\langle w, x_n \rangle + b) \geq 0 \quad \text{untuk semua } n = 1, \ldots, N$$

12.2 Primal Support Vector Machine

Untuk data yang dapat dipisahkan secara linear, terdapat tak terhingga banyaknya hyperplane yang dapat memisahkan kedua kelas tanpa kesalahan pelatihan. Ide dari SVM adalah untuk memilih hyperplane pemisah yang memaksimalkan *margin* antara contoh positif dan negatif. Margin yang besar cenderung memberikan generalisasi yang baik.

12.2.1 Konsep Margin

Margin secara intuitif adalah jarak dari hyperplane pemisah ke contoh terdekat dalam set data. Untuk meresmikannya, kita memaksimalkan jarak r sambil memastikan semua data berada pada sisi yang benar dari hyperplane, dengan kendala normalisasi pada w. Ini menghasilkan masalah optimisasi:

$$\max_{w,b,r} r \quad \text{dengan kendala} \quad y_n(\langle w, x_n \rangle + b) \geq r, \quad |w| = 1, \quad r > 0$$

12.2.2 Penurunan Margin Tradisional

Pendekatan alternatif adalah dengan memilih skala data sedemikian rupa sehingga nilai prediktor untuk contoh terdekat adalah 1, yaitu $\langle w, x \rangle + b = 1$. Dengan asumsi ini, jarak geometris r dari hyperplane ke contoh terdekat dapat diturunkan menjadi:

$$r = \frac{1}{|w|}$$

Memaksimalkan margin r ini setara dengan meminimalkan $|w|$. Untuk kemudahan matematis, kita meminimalkan $\frac{1}{2}|w|^2$. Ini mengarah pada masalah optimisasi *hard margin SVM*, yang tidak mengizinkan adanya pelanggaran margin:

$$\min_{w,b} \frac{1}{2}|w|^2 \quad \text{dengan kendala} \quad y_n(\langle w, x_n \rangle + b) \geq 1 \quad \text{untuk semua } n$$

Kedua formulasi margin ini (dari 12.2.1 dan 12.2.2) terbukti ekuivalen.

12.2.4 Soft Margin SVM: Pandangan Geometris

Untuk data yang tidak dapat dipisahkan secara linear, kita mengizinkan beberapa contoh berada di dalam margin atau bahkan di sisi yang salah dari hyperplane. Hal ini dicapai dengan memperkenalkan variabel-variabel kelonggaran (*slack variables*) $\xi_n \geq 0$ untuk setiap contoh. Masalah optimisasi menjadi *soft margin SVM*:

$$\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{n=1}^{N} \xi_n \quad \text{dengan kendala} \quad y_n(\langle w, x_n \rangle + b) \geq 1 - \xi_n, \quad \xi_n \geq 0$$

Parameter regularisasi $C > 0$ menyeimbangkan antara ukuran margin dan jumlah total kelonggaran.

12.2.5 Soft Margin SVM: Pandangan Fungsi Kerugian

Formulasi *soft margin* SVM juga dapat diturunkan dari prinsip minimisasi risiko empiris. Dalam pandangan ini, suku $\frac{1}{2}|w|^2$ adalah regularizer, dan maksimisasi margin dapat diinterpretasikan sebagai regularisasi. Fungsi kerugian (*loss function*) yang digunakan adalah *hinge loss*:

$$\ell(t) = \max 0, 1 - t \quad \text{di mana} \quad t = y(\langle w, x \rangle + b)$$

Ini mengarah pada masalah optimisasi tak terkendala yang ekuivalen:

$$\min_{w,b} \frac{1}{2}|w|^2 + C \sum_{n=1}^{N} \max 0, 1 - y_n(\langle w, x_n \rangle + b)$$

12.3 Dual Support Vector Machine

Formulasi primal SVM memiliki jumlah parameter yang bergantung pada dimensi fitur D. Formulasi dual yang ekuivalen memiliki jumlah parameter yang bergantung pada jumlah contoh data N, yang berguna ketika $D > N$.

12.3.1 Dualitas Konveks melalui Pengali Lagrange

Dengan menerapkan dualitas Lagrange pada masalah primal *soft margin*, kita dapat menurunkan masalah dualnya. Salah satu hasil kunci dari penurunan ini adalah *representer theorem*, yang menyatakan bahwa vektor bobot optimal adalah kombinasi linear dari contoh-contoh pelatihan:

$$w = \sum_{n=1}^{N} \alpha_n y_n x_n$$

Contoh-contoh x_n yang memiliki $\alpha_n > 0$ disebut *support vectors* karena mereka yang "menopang" hyperplane. Masalah optimisasi *dual SVM* dinyatakan secara eksklusif dalam variabel dual α_i:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i=1}^{N} \alpha_i \quad \text{dengan kendala} \quad \sum_{i=1}^{N} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C$$

12.3.2 Dual SVM: Pandangan Selubung Konveks (*Convex Hull*)

Sebuah cara alternatif untuk menurunkan SVM dual adalah dengan mempertimbangkan selubung konveks dari contoh-contoh kelas positif dan negatif. Masalahnya menjadi

mencari dua titik, satu di setiap selubung konveks, yang memiliki jarak terpendek satu sama lain. Meminimalkan jarak ini setara dengan masalah dual SVM *hard margin*.

12.4 Kernel

Formulasi dual SVM hanya bergantung pada perkalian dalam (*inner product*) antara contoh-contoh data, $\langle x_i, x_j \rangle$. Hal ini memungkinkan kita untuk mengganti perkalian dalam ini dengan sebuah fungsi *kernel* $k(x_i, x_j)$, yang secara implisit menghitung perkalian dalam pada ruang fitur berdimensi lebih tinggi yang dipetakan secara non-linear, $\langle \phi(x_i), \phi(x_j) \rangle$. Trik ini, yang dikenal sebagai *kernel trick*, memungkinkan SVM (yang merupakan pengklasifikasi linear) untuk membangun batas keputusan non-linear tanpa secara eksplisit merepresentasikan pemetaan fitur $\phi(x)$. Sebuah fungsi dapat menjadi kernel jika matriks Gram yang dihasilkannya simetris dan semidefinit positif.

12.5 Solusi Numerik

Masalah optimisasi SVM dapat diselesaikan dengan berbagai cara. Pandangan fungsi kerugian menghasilkan masalah optimisasi tak terkendala yang dapat diselesaikan dengan metode subgradien karena *hinge loss* tidak dapat diturunkan di semua titik. Formulasi primal dan dual SVM adalah masalah *quadratic programming* (QP) konveks dan dapat ditulis dalam bentuk standar untuk diselesaikan menggunakan perangkat lunak optimisasi.