Optimization Methods
Confronto tra SMO e DCD per l'addestramento di SVMs

Carlo Lucchesi
Matricola: 7176837

Febbraio 2025

1 Obbiettivo dell'Elaborato

Questo elaborato contiene i risultati del Laboratorio di Optimization Methods in merito a gli algoritmi di ottimizzazione utilizzati per l'addestrameto di SVM lineari: il Sequential Minimal Optimization (SMO) il Dual Cordinate Decend (DCD).

L'obbiettivo consiste nell'implementazione degli algoritmi, e la loro validazione con un modello di confronto. Stabilita la correttezza, eseguire dei test con dataset di dimensioni diverse e a precisioni diverse al fine di confrontarne le performance.

1.1 Strumenti utilizzati

Per la realizzazione del codice sono stati utilizzati i seguenti strumenti:

- **numpy**:
- **scipy**: dataset recuperati con la funzione *load_svmlight_file()* restituivano una matrice sparsa, che doveva essere convertita in una array bidimensionale denso.
- **scikit-learn**: da questa libreria sono stati utilizzati io modello LinearSVC come implementazione di riferimento, e altre funzioni per la gestione di train e test set.
- **pandas**: questa libreria è stata utilizzata per poter condurre analizzare velocemente le caratteristiche salienti dei dataset utilizzati.
- **matplotli**: questa libreria è stata utilizzata per la generazione di grafici.
- **jupyter**: si è scelto di implementare il codice in una Notebook Jupyter per renderlo più chiaro e interattivo.

Per evitare conflitti con le librerie installate localmente, si è scelto si creare un ambiente virtuale mediante **conda**. Questo è stato realizzato con il comando.

```
conda create --name OPT numpy scikit-learn jupyter pandas\
    conda-forge::matplotlib
conda activate OPT
```

2 Algoritmi implementati

Nel contesto del machine learning, l'efficienza degli algoritmi diventa cruciale data al crescere dei dataset utilizzati. Nel caso delle SVM, la formulazione del problema duale è diventata sempre più popolare per via della possibilità di realizzare il *Kernel Trick*. Questo ha portato a notevoli sviluppi nella risoluzione del problema in forma duale.

Nel contesto di dati con un numero elevato di features, l'utilizzo del Kernel Trick diventa irrilevante. Infatti, un ulteriore aumento della dimensionalità dei dati potrebbe far perdere la capacità di generalizzazione del modello finale, e portare al problema dalla *Curse of Dimensionality*.

Limitandosi quindi ai kernel lineari, è possibile realizzare una formulazione semplificata del problema duale che richiede meno tempo per l'esecuzione di una singola iterazione.

2.1 Sequential Minimal Optimization (SMO)

Questo algoritmo permette di risolvere il problema duale di una SVM, ovvero:

$$min\frac{1}{2}\alpha^T Q\alpha - e^T\alpha \tag{1}$$

$$0 \leq \alpha \leq C \tag{2}$$

$$\alpha^T y = 0 \tag{3}$$

Sia $X \in \mathbf{R}^{n\times p}$ il dataset, e $Y \in \{-1,1\}^n$ le classi, si ottiene che $Q_{ij} = y^i y^j K(x^i, x^j)$, dove K è una Kernel Function valida. Noi vedremo il caso lineare dove $Q_{ij} = y^i y^j {x^i}^T x^j$.

Al crescere della dimensione del dataset calcolare esplicitamente Q potrebbe non essere fattibile, per questo motivo spesso vengono calcolate le singole colonne della matrice quando necessario. Infatti, nella maggior parte dei casi $\alpha_i = 0$, rendendo colonne inutili. Per via nelle limitate capacità di calcolo a disposizione, si dataset abbastanza piccoli tali da permette un calcolo esplicito di Q. I tempi sono stati confrontati con una implementazione che calcola le singole colonne (definizione *lazy* matrice Q), e non si sono notate significative differenze. Con n più grandi potrebbe non essere possibile il calcolo di Q.

Dato l'elevato numero di variabili, è conveniente ottimizzare il problema con una tecnica di decomposizione. In questo caso, dato il vincolo di eguaglianza all'equazione (3), è necessario utilizzare un working set $W = \{i, j | i, j\{1, 2, \ldots, n\}, i \neq j\}$, cosa che permette di individuare la soluzione ottima del sotto problema in modo forma chiusa.

Oltre a questo, bisogna associare a W una direzione che sia: ammissibile, di discesa, e che rispetti l'equazione (3). Sia $\alpha' = \alpha + d$ con d direzione dello spostamento, il mantenimento del vincolo (3) sia ha con:

$$d_i = \frac{1}{y^i}, d_j = -\frac{1}{y^j}, d_k = 0 \ \forall k \neq i, j \tag{4}$$

Data questa forma della direzione, perché sia ammissibile si deve restringere la selezione a $W = \{i, j\}$ tale per cui:

$$\begin{aligned} i \in R(\alpha) &= \{i | (\alpha_i = 0 \ \wedge \ y^i = 1) \vee (\alpha_i = C \ \wedge \ y^i = -1) \ \vee (\alpha_i \in (0, C))\} \\ j \in S(\alpha) &= \{j | (\alpha_j = 0 \ \wedge \ y^j = -1) \vee (\alpha_j = C \ \wedge \ y^j = 1) \ \vee (\alpha_j \in (0, C))\} \end{aligned} \tag{5}$$

Infine, per garantire una direzione di discesa si seleziona, la coppia $W = \{i, j\}$ Deve rispettare il vincolo per cui:

$$\frac{\nabla_i f(\alpha)}{y_i} < \frac{\nabla_j f(\alpha)}{y_j} \tag{6}$$

Dato la scelta delle variabili espressa in (6), per garantire una terminazione in un numero finito di passi, la scelta delle variabili avviene con la coppia che realizza la distanza maggiore di queste quantità, ovvero:

$$i \in arg \max_{h\in R(\alpha)} \{-\frac{\nabla_h f(\alpha)}{y^h}\}, \ j \in arg \min_{h\in S(\alpha)} \{-\frac{\nabla_h f(\alpha)}{y^h}\} \tag{7}$$

Data questa scelta specifica dalle variabili, sfruttando le condizioni KKT per l'individuazione di un ottimo, si può dimostrare (sezione 7.2.2 [1]) che $\forall\epsilon$ la seguente condizione viene soddisfatta in un numero finito di passi:

$$\max_{h\in R(\alpha)} \{-\frac{\nabla_h f(\alpha)}{y^h}\} + \epsilon > \min_{h\in S(\alpha)} \{-\frac{\nabla_h f(\alpha)}{y^h}\} \tag{8}$$

2.2 DCD

Se ci limitiamo al caso di una SVM con Kernel lineare (ovvero dove $K(x^i, x^j) = {x^i}^T x^j$), è possibile riformulare il problema di ottimizzazione senza il bias b, aggiungendo a ogni elemento una feature fittizia sempre uguale ad 1:

$$\min_w \frac{1}{2}||w||^2 + C\sum_i^n \max\{0, 1 - y^i(w^T x^i)\} \tag{9}$$

Oppure, in modo analogo, si può passare alla formulazione:

$$\min_w \frac{1}{2}||w||^2 + C\sum_i^n \max\{0, 1 - y^i(w^T x^i)\}^2 \tag{10}$$

Da queste si può derivare la formulazione del duale:

$$\begin{gathered}\min_\alpha \frac{1}{2}\alpha^T \bar{Q}\alpha - e^T\alpha \\ 0 \leq \alpha \leq U\end{gathered} \tag{11}$$

Dove $\bar{Q} = Q + D$, con D matrice diagonale. In (9) $U = C$ e $D_{ii} = 0$, mentre in (10) $U = \infty$ e $D_{ii} = \frac{1}{2C}$. In ogni caso, è stato rimosso il vincolo di uguaglianza (3), permettendo di lavorare con working set di una singola variabile.

In questo elaborato, si è esplorata soltanto la formulazione derivata da (9) in modo da poterne confrontare il duale con (1).

Questa formulazione semplificata riduce il costo per l'ottimizzazione in un singolo passo, ma discapito di richiede più iterazioni.

La scelta della variabile è indifferente, purché ogni variabile venga scelta ciclicamente. Come condizione di terminazione, è stata utilizzata quella descritta nella sezione 3.3 dell'articolo [2].

2.3 Analisi di Complessità

Analizzando le operazioni richieste per eseguire un singolo passo si ottengono questi risultati. Quindi, in questi esperimenti si vorrà verificare se il ridotto costo di aggiornamento di DCD

Aggiornamento	SMO	DCD (caso lineare)
Variabili	$O(1)$	$O(p)$
Gradiente	$O(np)$	NA

Table 1: Costo asintotico delle operazioni di aggiornamento

rispetto a SMO compensa il numero maggiore di operazioni eseguite.

Come detto in sezione 2.1, la matrice Q viene calcolata esplicitamente, comportando un costo asintotico complessivo di $O(pn^2)$. Questo non può essere confrontato direttamente con il costo di una singola iterazione, in quanto il calcolo della matrice avviene una sola volta all'inizio dell'algoritmo. Inoltre, si è notato che fintanto sia possibile mantenere Q interamente in memoria, non c'è nessun peggioramento delle performance rispetto a una valutazione singola delle colonne necessarie.

3 Dataset Utilizzati

In entrambi i dataset è stata aggiunta una feature fittizia con valore sempre uguale a 1 nei casi di test con DCD per poter modellare il bias come un parametro del modello, cosa richiesta dalla formulazione del problema.

Per entrambi i dataset è stato necessario accertarsi che le classi fossero identificate con 1 e -1, dato che questi valori vengono utilizzati nelle operazioni aritmetiche degli algoritmi.

3.1 Breast Cancer

Questo dataset è stato scelto per via della sua ridotta dimensione, così da verificare la correttezza delle implementazioni sotto analisi. Tra le versioni disponibili è stata scelta quella "scalata" tra [-1, 1], in modo da semplificare l'addestramento.

Di questo dataset non è stato possibile recuperare ulteriori informazioni.

3.2 Covtype

Questo dataset è stato utilizzato per i test di performance. Data l'elevata dimensione (581.012 elementi per 54 feature ciascuno), non è stato possibile utilizzarlo interamente per 2 motivi. I primo dall'implementazione di SMO utilizzata che calcola esplicitamente Q (anche se è già presente una implementazione con il calcolo *lazy* della matrice). Il secondo per le limitate capacità di calcolo a disposizione. Quindi, per ovviare a questi problemi, ne sono stati estratti dei campioni.

Analizzando svariati dataset disponibili su questa pagina, escludendo quelli di dimensioni inadeguati, con una preliminare analisi con i DataFrame di **pandas**, si può notare che molti di essi avessero features con le stesse metriche di tendenza centrale. Questo fenomeno suggeriva una scarsa qualità del dataset. Questo dataset non risultava affetto dal fenomeno.

Maggiori informazioni sul dataset sono disponibili qui. Notare che il dataset che è stato precedentemente convertito per realizzare una classificazione binaria, e non multi classe come originariamente era stato pensato.

4 Risultati

Inizialmente si è verificato che le implementazioni fossero corrette, andando ad analizzare la accuratezza ottenuta sia da SMO che DCD, confrontandola con quella del modello LinearSVC della libreria **scikit-learn**.

Una volta accertata la correttezza, i test si sono focalizzati sulle performance dei due algoritmi al variare della dimensione del dataset. Data la scarsità di risorse computazionali, e di un numero di dataset comparabili, si è scelto di eseguire i vari test con sottoinsiemi del dataset covertype descritto nella sezione 3.2.

Infine, si è misurato il comportamento degli algoritmi al cambiamento del numero di features.

4.1 Correttezza

Come detto, per prima cosa si è verificata la correttezza degli algoritmi implementati. Per questo scopo si è utilizzato il dataset descritto nella sezione 3.1.

Classe	Precision			Recall			F1		
	LinearSVC	SMO	DCD	LinearSVC	SMO	DCD	LinearSVC	SMO	DCD
-1	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
1	0.97	0.97	0.97	0.97	0.97	0.7	0.97	0.97	0.97
	LinearSVC			SMO			DCD		
accuracy	0.97			0.96			0.97		

Table 2: Risultati dei modelli ottenuti dai vari algoritmi sul training set.

I risultati riportati in 2 sono stati ottenuti ponendo il coefficiente di regolarizzazione $C = 10$.

Da questa tabella si può notare come il modelli risultanti siano praticamente identici. In questo caso, i risultati fanno riferimento al training set in quanto ci interessava che i modelli fossero simili, e non le loro capacità di generalizzazione.

4.2 Cambiamento della dimensione

Gli esperimenti successivi si sono concentrati sul comportamento degli algoritmi al variare della dimensione del dataset.

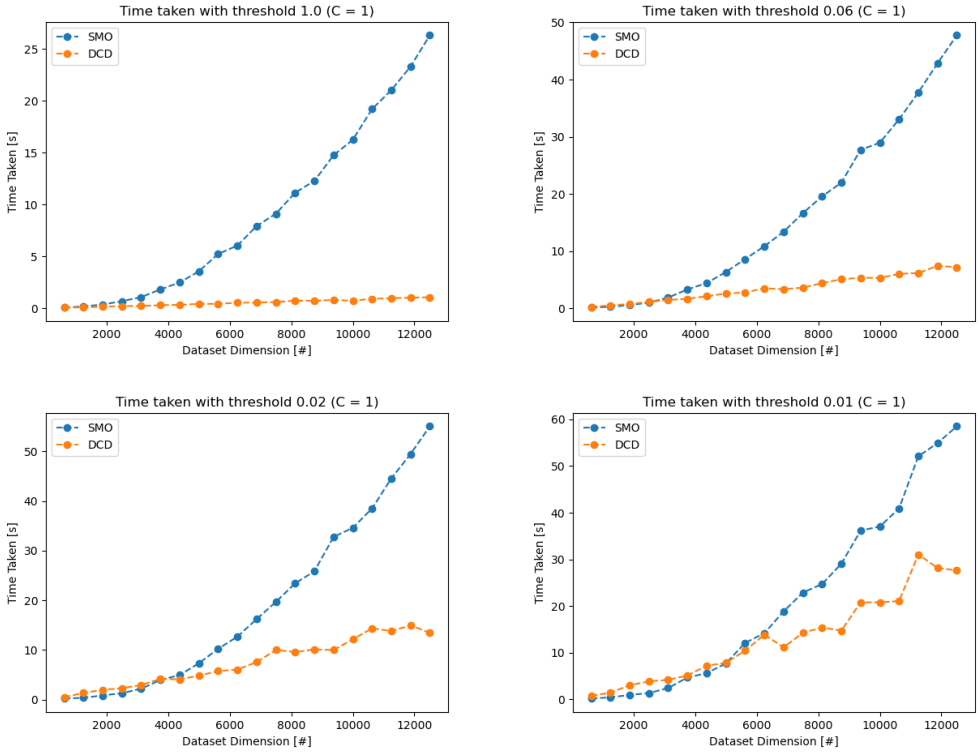

Figure 1: Andamento del tempo di addestramento al cambiare della dimensione del dataset.

Come si può notare dalle figure in 1, il tempo richiesto da DCD cresce più lentamente

rispetto a SMO. In generale, la complessità di DCD è lineare rispetto alla dimensione del dataset, mentre per SMO quadratica, cosa che sembra in linea con il tempo per il calcolo della matrice Q.

In generale si può notare come la differenza tra i due algoritmi sembra venir meno al crescere della precisione richiesta nella condizione di stop. Questo fenomeno è più facilmente descritto dalla figura 2.

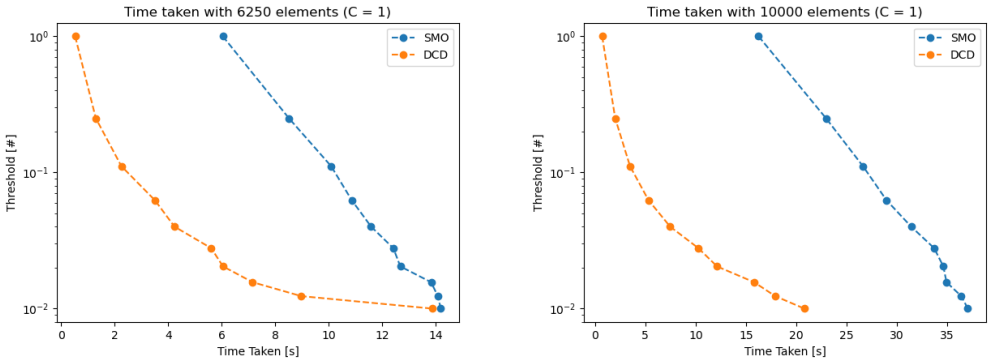

Figure 2: Rapporto tra tolleranza della condizione terminante e il tempo.

Per SMO, un aumento di una cifre significative della tolleranza sulla condizione d'arresto sembra far crescere linearmente il tempo di addestramento. Questo non vale per DCD, la cui relazione sembra esponenziale. Detto questo, è importante notare come la condizione di terminazione non sia strettamente correlata alla precisione ottenuta su train e test set, come mostrato in figura 3.

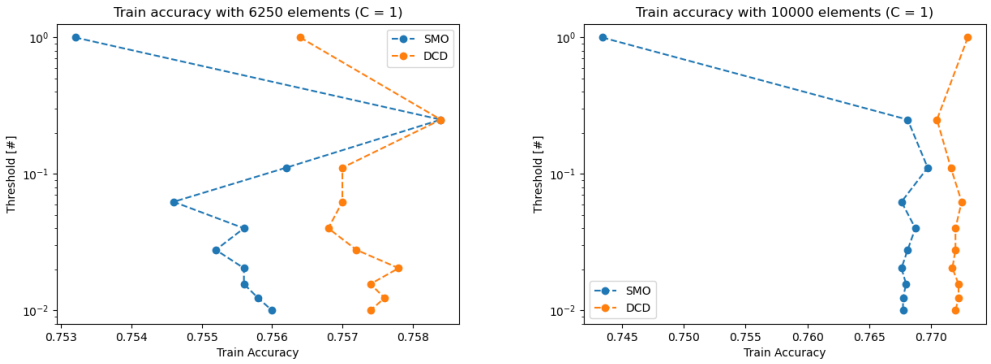

Figure 3: Rapporto tra tolleranza della condizione terminante e precisione su train set.

Fatta eccezione di condizioni terminanti troppo piccole, l'accuratezza sul train set è raggiunta anche con valori tolleranze piuttosto piccole, ammesso di utilizzare dataset di adeguate dimensioni.

Infatti, come mostrato in figura 4, sembra che la precisione sul test set dipenda principalmente dalla dimensione del dataset utilizzato, piuttosto che dalla tolleranza sulla condizione di terminazione.

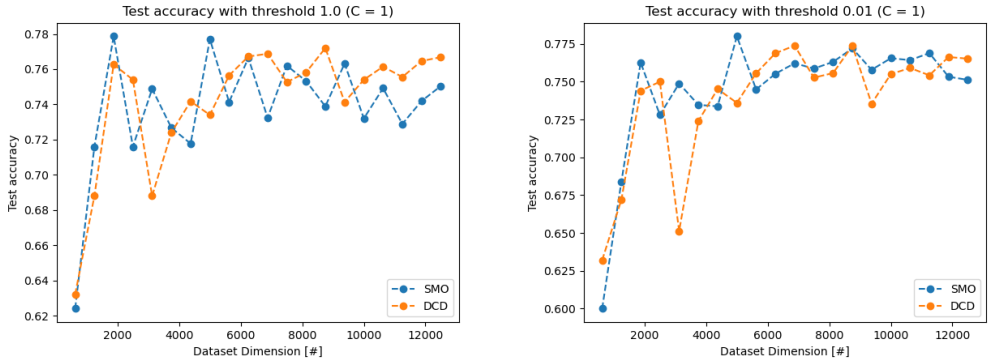

Figure 4: Rapporto tra dimensione del train set e precisione su test set.

4.3 Cambiamento del numero di features

Sono stati condotti anche dei test che verificavano il tempo impiegato per l'addestramento al variare del numero di features utilizzate. In ogni caso di test ci si è assicurati che entrambi gli algoritmi utilizzassero lo stesso sottoinsieme di features. Questo era volto a evitare che un algoritmo, in qualche caso ricevesse un insieme sfortunato che mal separava il dataset, falsando il confronto.

I risultati riportati in figura 5, mostrano un andamento lineare del tempo di addestramento rispetto al numero di features. Questo è in linea con le aspettative teoriche sul costo asintotico per operazione.

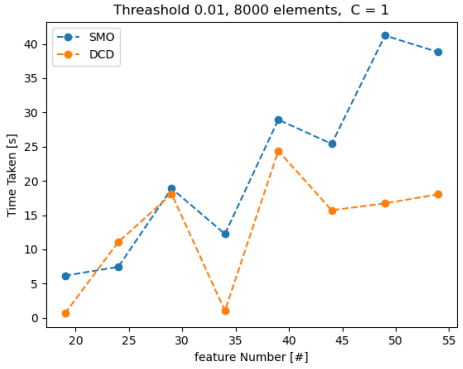

Figure 5: Tempo di addestramento in funzione delle feature.

5 Conclusioni

In questo laboratorio è stato possibile apprezzare come la maggior efficienza delle singole iterazioni di DCD ne compensi il numero maggiore.

In particolare, si è notato un comportamento asintotico per l'addestramento di DCD

pari a $O(pn)$, mentre per SMO di $O(pn^2)$. Notare che le aspettative teoriche riportate in 2.3 sono compatibili con i risultati. Questi indicano il costo computazionale per una singola iterazione, mentre il tempo di addestramento è influenzato dal numero complessivo di iterazioni. Sembra ragionevole assumere che il numero di iterazioni richiesto cresca linearmente con il numero di elementi nel dataset.

Oltre a questo, si conferma che l'utilizzo di una tolleranza più piccola aumenti il tempo complessivo di addestramento. Gli esperimenti mostrano come DCD sia maggiormente influenzato rispetto a SMO, mantenendo però un andamento lineare della complessità rispetto alla grandezza del dataset. Quindi, al netto di errori di calcolo numerico che impediscono di raggiungere elevate precisioni, DCD rimane più efficiente a livello asintotico. Inoltre, come mostrato dalla figura 3, per ottenere un modello adeguato è sufficiente l'utilizzo di tolleranze piuttosto piccole.

Infine, come mostrato dalla figura 4, la formulazione senza bias non ha comportato un peggioramento del modello, rendendo DCD una valida opzione per l'addestramento di SVM lineari su dataset di grandi dimensioni.

References

1. Lapucci, M. *Lecture Notes* (2024).
2. Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S. S. & Sundararajan, S. *A dual coordinate descent method for large-scale linear SVM* in *Proceedings of the 25th international conference on Machine learning* (2008), 408–415.