ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC KHOA HỌC

—————o0o—————

VŨ VĂN THỊNH

TỐI ƯU DC VÀ ỨNG DỤNG TRONG BÀI TOÁN PHÂN CỤM

LUẬN VĂN THẠC SĨ TOÁN HỌC

Thái Nguyên - 2017

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC KHOA HỌC

—————o0o—————

VŨ VĂN THỊNH

TỐI ƯU DC VÀ ỨNG DỤNG TRONG BÀI TOÁN PHÂN CỤM

Chuyên ngành: Toán ứng dụng
Mã số: 60.46.01.12

LUẬN VĂN THẠC SĨ TOÁN HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. TẠ MINH THỦY

Thái Nguyên - 2017

Mục lục

Danh mục các ký hiệu

$\mathbb{R}$	Tập hợp số thực
$\mathbb{R}^n$	Không gian số thực n-chiều
X^*	Không gian liên hợp của X
$x \in C$	x thuộc tập C
$x \notin C$	x không thuộc tập C
$x := y$	x được định nghĩa bằng y
$\exists x$	Tồn tại x
$\forall x$	Với mọi x
$\varnothing$	Tập hợp rỗng
$\cap$	Phép giao các tập hợp
$\cup$	Phép hợp các tập hợp
$\langle x, y \rangle$	Tích vô hướng của x và y
$\nabla_x f(x)$	Véc tơ đạo hàm của hàm f tại điểm x
A^T	Ma trận chuyển vị của ma trận A
A^*	Toán tử liên hợp của toán tử A
I	Ánh xạ đơn vị
$\|x\|$	Chuẩn của véc tơ x
$\arg\min\{f(x) : x \in C\}$	Tập các điểm cực tiểu của hàm f trên C

Mở đầu

Lý thuyết tối ưu đã được tìm hiểu và phát triển để giải quyết những vấn đề trong thực tế cuộc sống. Tuy nhiên với những bài toán có hàm mục tiêu không lồi, bài toán trở nên phức tạp hơn, và chính những bài toán thực tế lại thường dẫn đến hàm mục tiêu không lồi. Luận văn này sẽ tìm hiểu về lý thuyết tối ưu DC (hiệu 2 hàm lồi – difference of convex) và thuật toán DC (DCA - DC Algorithm) để giải quyết những vấn đề như vậy. Phân cụm là một trong những bài toán khó và được nghiên cứu nhiều trong lĩnh vực Tin học và Công nghệ thông tin. Bài toán phân cụm là chia dữ liệu thu thập được thành những cụm (nhóm) có cùng tính chất. Đây là bài toán NP – khó và đã được nghiên cứu từ lâu. Trong luận văn này, chúng tôi tìm hiểu phương pháp tối ưu DC và thuật toán DC để giải quyết bài toán phân cụm. Thuật toán được thử nghiệm trên các bộ dữ liệu thu được từ những vấn đề trong thực tế.

Luận văn gồm có 3 chương:

Chương 1: Giới thiệu các kiến thức cơ bản về giải tích lồi, đặc biệt chú trọng hàm lồi, hàm DC và một số tính chất của hàm DC; những kiến thức này được sử dụng làm nền tảng trong các chương tiếp theo.

Chương 2: Giới thiệu về phân cụm dữ liệu và một số vấn đề của phân cụm dữ liệu. Trong chương này, luân văn sẽ trình bày về khái niệm của phân cụm, các yêu cầu và giới thiệu một số kỹ thuật trong phân cụm dữ liệu. Chương này sẽ trình bày cụ thể một số cách tiếp cận theo hướng phân cụm phân hoạch.

Chương 3: Từ thuật toán phân cụm với trọng số của thuộc tính (SCAD) đã được trình bày trong chương 2, luận văn sẽ giới thiệu phương pháp tối ưu DC và giải thuật DC cho bài toán tối ưu không

lỗi đã được trình bày. Chương này cũng trình bày các kết quả thực nghiệm của thuật toán với các bộ dữ liệu thực tế.

Do thời gian có hạn nên luận văn này chủ yếu chỉ dừng lại ở việc tập hợp tài liệu, bước đầu tìm hiểu về lý thuyết tối ưu DC cũng như thuật toán DC. Luận văn cũng đưa ra được những kết quả thực nghiệm ban đầu minh họa cho thuật toán. Trong quá trình viết luận văn cũng như trong soạn thảo văn bản, luận văn chắc chắn không khỏi có những sai sót nhất định. Tác giả rất mọng nhận được sụ góp ý của các thầy cô, bạn bè đồng nghiệp để luận văn được hoàn thiện hơn.

Nhân dịp này em xin được bày tỏ lòng biết ơn sâu sắc tới thầy hướng dẫn TS. Tạ Minh Thủy đã tận tình giúp đỡ tác giả trong suốt quá trình làm luận văn. Em cũng xin chân thành cảm ơn các thầy, cô: GS, PGS, TS,... của khoa Toán - Tin trường Đại học Khoa học Thái Nguyên và Viện Toán học đã giảng dạy và tạo mọi điều kiện thuận lợi trong quá trình tác giả học tập và nghiên cứu.

Thái Nguyên, tháng 4 năm 2017

Tác giả luận văn

Vũ Văn Thịnh

Chương 1

Một số khái niệm cơ bản

1.1 Tập lồi

Định nghĩa 1.1.1 Tập $X \subseteq \mathbb{R}^n$ được gọi là tập lồi nếu $\forall x, y \in X$ và với mọi số thực $\lambda \in [0,1]$ thì $\lambda x + (1-\lambda)y \in X$.
Nghĩa là nếu $x, y \in X$ thì đoạn:

$$[x,y] := \{z \in \mathbb{R}^n, z = \lambda x + (1-\lambda)y \in X, 0 \leq \lambda \leq 1\} \subseteq X.$$

Ví dụ:

i) Cả không gian $\mathbb{R}^n$ và tập $\emptyset$ là các tập lồi.

ii) Các hình cầu mở hoặc đóng trong $\mathbb{R}^n$ tức là các tập:

$B(x_0, r) = \{x \in \mathbb{R}^n, \|x - x_0\| < r\}$ và $B(x_0, r) = \{x \in \mathbb{R}^n, \|x - x_0\| \leq r\}$

là tập lồi.

1.2 Hàm lồi

Định nghĩa 1.2.1 Hàm $f : X \to [-\infty, +\infty]$ xác định trên một tập lồi $X \subseteq \mathbb{R}^n$ được gọi là hàm lồi trên X nếu với mọi $x^1, x^2 \in X$ và mọi số thực $\lambda \in [0,1]$ ta có

$$f[(1-\lambda)x^1 + \lambda x^2] \leq (1-\lambda)f(x^1) + \lambda f(x^2).$$

Hàm $f : X \to [-\infty, +\infty]$ được gọi là lồi chặt trên tập lồi X nếu với mọi $x^1, x^2, x^1 \neq x^2$ và $\lambda \in (0,1)$ ta có

$$f[(1-\lambda)x^1 + \lambda x^2] < (1-\lambda)f(x^1) + \lambda f(x^2).$$

Một hàm lồi chặt là lồi, nhưng điều ngược lại không đúng.
Hàm $f : X \to [-\infty, +\infty]$ được gọi là lồi mạnh trên tập lồi X nếu tồn tại một số $\rho > 0$ sao cho với mọi $x^1, x^2 \in X, x^1 \neq x^2$ và $\lambda \in (0, 1)$ ta có

$$f[(1-\lambda)x^1 + \lambda x^2] \leq (1-\lambda)f(x^1) + \lambda f(x^2) + \rho\|x^1 - x^2\|^2.$$

Định nghĩa 1.2.2 Hàm $f : X \to [-\infty, +\infty]$ được gọi là lõm (lõm chặt) trên tập lồi X nếu $-f$ là lồi (lồi chặt) trên X. Hàm $f : X \to [-\infty, +\infty]$ được gọi là tuyến tính afin (hay afin) trên X nếu f nhận giá trị hữu hạn và vừa lồi vừa lõm trên X. Một hàm afin trên $\mathbb{R}^n$ có dạng $f(x) = \langle a, x\rangle + \alpha$ với $a \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ bởi vì $\forall x^1, x^2 \in \mathbb{R}^n$ và $\forall \lambda \in [0, 1]$, ta có

$$f[(1-\lambda)x^1 + \lambda x^2] = (1-\lambda)f(x^1) + \lambda f(x^2).$$

Tuy nhiên hàm afin không lồi chặt hay lõm chặt.

Ví dụ về hàm lồi:

i) Hàm chuẩn Euclid $\|x\| = \sqrt{\langle x, x\rangle}, x \in \mathbb{R}^n$

ii) Hàm khoảng cách từ điểm $x \in \mathbb{R}^n$ tới tập C ($C \subset \mathbb{R}^n$ là một tập lồi khác rỗng):

$$d_C(x) = \inf_{y \in C} \|x - y\|.$$

Định nghĩa 1.2.3 Cho hàm bất kỳ $f : X \to [-\infty, +\infty]$ với $X \subseteq \mathbb{R}^n$, các tập

$$dom\ f = \{x \in X : -\infty < f(x) < +\infty\},$$

$$epi\ f = \{(x, \alpha) \in X \times \mathbb{R} : f(x) \leq \alpha\}$$

được gọi lần lượt là miền hữu dụng (hữu hiệu) và tập trên đồ thị của hàm f.
Nếu $domf \neq \emptyset$ và $f(x) > -\infty \neq, \forall x \in X$ thì ta nói hàm f là chính thường.

Hàm lồi $f : X \to [-\infty, +\infty]$ có thể được mở rộng thành hàm lồi trên không gian $\mathbb{R}^n$ bằng cách đặt $f(x) = +\infty,\ \forall x \notin dom\ f$. Vì vậy để đơn giản ta thường xét f là hàm lồi trên toàn $\mathbb{R}^n$.

Định nghĩa 1.2.4 Một ma trận A được gọi là ma trận xác định dương nếu với vecto x bất kỳ ta có: $x^T Ax > 0$.
Ma trận A được gọi là ma trận nửa xác định dương nếu với vecto x bất kỳ ta có: $x^T Ax \geq 0$.

Định nghĩa 1.2.5 Cho $x^0 \in X \subseteq \mathbb{R}^n$. Hàm chính thường $f : X \to [-\infty, +\infty]$

i) được gọi là nửa liên tục dưới tại x^0 nếu $\limsup_{y \to x^0} f(y) \geq f(x^0)$.

ii) nửa liên tục trên tại x^0 nếu $\limsup_{y \to x^0} f(y) \leq f(x^0)$.

iii) Hàm f được gọi là liên tục tại điểm x^0 nếu nó vừa nửa liên tục trên và nửa liên tục dưới tại x^0.

Định lý 1.2.6 i) Một hàm thực một biến $\varphi(t)$ khả vi trong một khoảng mở (a, b) là lồi khi và chỉ khi đạo hàm của nó $\varphi'(t)$ là một hàm không giảm trên khoảng ấy.

ii) Một hàm thực một biến $\varphi(t)$ hai lần khả vi trong một khoảng mở là lồi khi và chỉ khi đạo hàm cấp hai của nó $\varphi''(t)$ không âm trên toàn bộ khoảng ấy.

Định lý 1.2.7 Cho một tập lồi $X \subset \mathbb{R}^n$ và một hàm $f : \mathbb{R}^n \to \mathbb{R}$ khả vi trên X. $\nabla f(x) = \left(\dfrac{\partial f}{\partial x_1}(x), ..., \dfrac{\partial f}{\partial x_n}(x) \right)$ là vectơ gradient của hàm f tại điểm x và $\dfrac{\partial f}{\partial x_i}$ là các đạo hàm riêng cấp một của f tính theo biến x_i. Khi đó:

i) Hàm f lồi trên X khi và chỉ khi:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \forall x, y \in X$$

ii) Nếu $f(y) > f(x) + \langle \nabla f(x), y - x \rangle$ với mọi $x, y \in X$ và $x \neq y$ thì hàm f lồi chặt trên X.

Định lý 1.2.8 Cho một tập lồi mở $X \subset \mathbb{R}^n$ và một hàm $f : \mathbb{R}^n \to \mathbb{R}$ hai lần khả vi liên tục trên X. Kí hiệu $\nabla^2 f(x)$ là ma trận các đạo hàm riêng cấp hai (hay hessian) của f tại x.