CSCE 990 Lecture 5: Regularization*

Stephen D. Scott

February 2, 2006

Introduction

- In the previous lecture, we discussed how the VC dimension of high- (or infinite-) dimensional hyperplanes can be controlled by maximizing the margin

- I.e. we further restrict the class of functions $\mathcal{F}$ (from general hyperplanes to large-margin hyperplanes) we choose from when minimizing $R_{\text{emp}}[f]$

- Thus rather than simply look for a hyperplane f that minimizes $R_{\text{emp}}[f]$, we look for an f that minimizes $R_{\text{emp}}[f]$ plus a regularization term

  - Typically, we'll use $\|\mathbf{w}\|^2$

Regularization

- Define a regularization term $\Omega[f]$ to our original objective function $R_{\text{emp}}[f]$ and get

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda\Omega[f] \quad ,$$

where $\Omega[f]$ quantifies the "complexity" of f and λ weights the tradeoff between the two optimization objectives

- Choosing convex $R_{\text{emp}}[f]$ (e.g. squared loss) and convex $\Omega[f]$ (e.g. $\|\mathbf{w}\|^2$) yields a convex $R_{\text{reg}}[f]$
  - We'll use this in the next lecture

CSCE 990 Lecture 6: Optimization*

Stephen D. Scott

February 7, 2006

Introduction

- In general, all machine learning algorithms focus on optimizing some function
  - E.g. $R_{\text{emp}}[f]$ or $R_{\text{reg}}[f]$
  - Main differences come from the representation of examples, choice of function to optimize, and choice of optimization method
- SVMs focus on optimizing functions that are convex
  - No local optima (in contrast to e.g. backpropagation for ANNs)
  - Well-studied problem with many algorithms, even when constraints added

Outline

- Convex sets and convex functions

- Unconstrained optimization

- Constrained optimization

- Sections 1.4, 6.1–6.2.2, 6.3, 6.6 (also read 6.2.3–6.2.4)

Convex Sets and Functions

D6.1 A set X in a vector space is convex if for all $x, x' \in X$ and any $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)x' \in X$$

– I.e. the shortest path from x to x' is entirely in X

D6.2 A function f defined on (possibly non-convex) set X is convex if for all $x, x' \in X$ and any $\lambda \in [0, 1]$ s.t. $\lambda x + (1 - \lambda)x' \in X$,

$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x')$$

– I.e. while moving point x'' in a straight line from x to x', $f(x'')$ lies below the line connecting $f(x)$ to $f(x')$

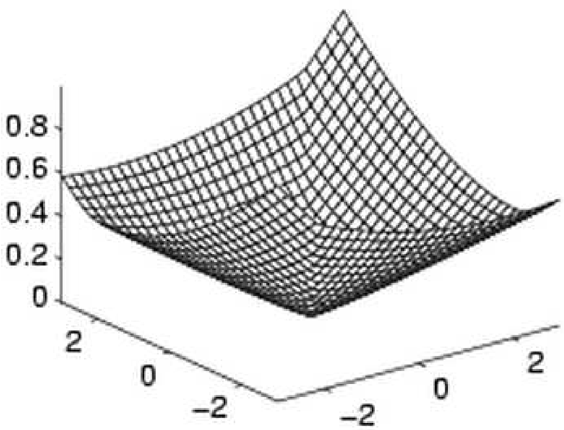
– I.e. $f(x)$ is shaped like a bowl

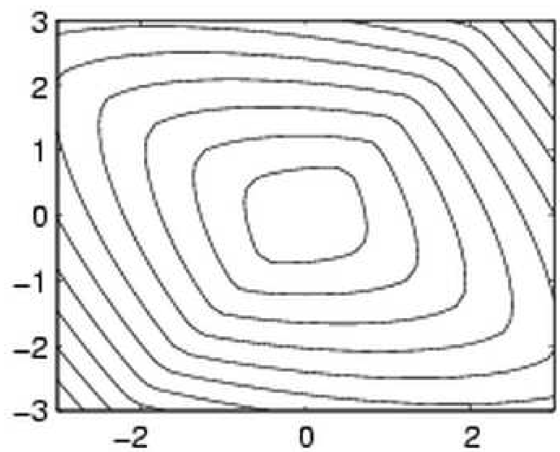
Properties of Convex Functions and Sets

L6.3 If f is a convex function on $\mathcal{X}$, then the convex level sets

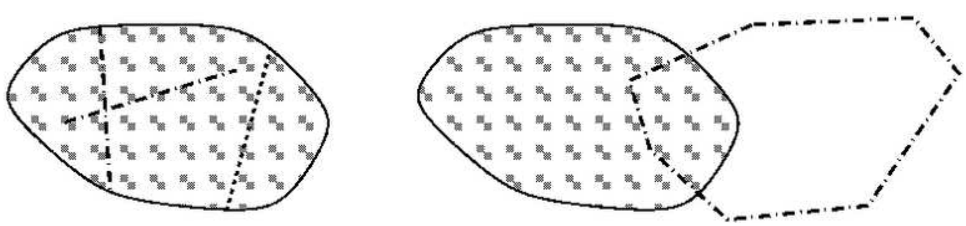
$$X_c := \{x \mid x \in \mathcal{X} \text{ and } f(x) \leq c\} \; \forall c \in \mathbb{R}$$

are convex





L6.4 If $X, X' \subset \mathcal{X}$ are both convex, then $X \cap X'$ is also convex



Constrained Convex Minimization

- Let $X \subset \mathcal{X}$ be convex, $f : \mathcal{X} \to \mathbb{R}$ be convex, and let c be the minimum value of f on X

- Then
$$X_m := \{x \mid x \in \mathcal{X} \text{ and } f(x) \leq c\}$$
is convex, as is $X_m \cap X$, and $f(x) = c$ for all $x \in X_m \cap X$

- Thus the set $X' \subseteq X$ on which f takes its minimum value over X is itself a convex set
  - Further, if f is strictly convex, then $|X'| = 1$

C6.6 If functions $f, c_1, \ldots, c_n$ are convex and if their domain $\mathcal{X}$ is convex, then the optimization problem
$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & c_i(x) \leq 0 \;\; \forall i \in \{1, \ldots, n\} \end{array}$$
has as its solution a convex set, if a solution exists. This solution is unique if f is strictly convex

Unconstrained Convex Minimization
Functions of One Variable
Interval Cutting

- Assume f is convex and differentiable

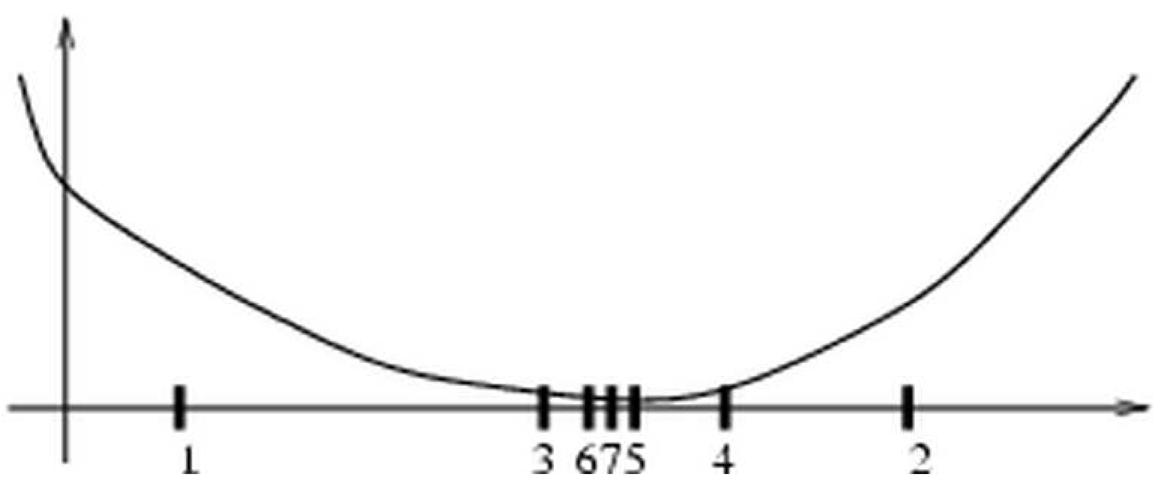
- Given an interval $[A, B] \subset \mathbb{R}$, look at $(A+B)/2$ and check if f is "going down" or "going up" at that point

  – If going up (i.e. $f'((A+B)/2) > 0$) then set $B = (A+B)/2$

  – Else set $A = (A+B)/2$

  – Repeat until $(B-A) \min\left(|f'(A)|, |f'(B)|\right) \leq \epsilon$

  – Called the Interval Cutting algorithm (Alg. 6.1, p. 155)



Unconstrained Convex Minimization

Functions of One Variable

Newton's Method

- We can do better if f twice differentiable

- Via Taylor series expansion of f around some fixed x_0:

  $f(x) \approx f(x_0)+(x-x_0)f'(x_0)+(x-x_0)^2 f''(x_0)/2$

- Minimize RHS by differentiating wrt x (so x_0 is a constant) and setting $= 0$:

$$x = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

- Thus Newton's Method starts at some point x_0 and repeatedly updates $x_{n+1} = x_n - f'(x_n)/f''(x_n)$ until $|f'(x_n)| \le \epsilon$

- Converges faster than Interval Cutting

Unconstrained Convex Minimization

Functions of Several Variables

Gradient Descent

- Very popular optimization technique
- Assume $f'(x)$ exists
- Like Newton's Method, we have a current solution x_n that we iterativly update
- At solution point x_n, compute the gradient* $g_n := f'(x_n)$, which gives the direction of steepest descent
- Then use line search (e.g. Newton's Method) to find γ that maximizes $f(x_n) - f(x_n - \gamma g_n)$
- Repeat until $|f'(x_n)| \leq \epsilon$
- Guaranteed to converge eventually

*Recall that the gradient of a function f over $\mathbb{R}^N$ is an N-dimensional vector of equations, where equation i is the partial derivative of f taken wrt the ith variable.

Constrained Optimization

- In SVMs, we will want to minimize $\|\mathbf{w}\|^2$, the squared length of the weight vector

- In general, this is trivially solved by $\mathbf{w} = \mathbf{0}$, so we need to constrain the set of solutions to choose from:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & c_i(x) \leq 0 \ \ \forall i \in \{1, \ldots, n\} \end{aligned} \tag{1}$$

- Can also convert equality constraint $e_j(x) = 0$ to pair of inequality constraints $c_j(x) \leq 0$ and $c'_j(x) \geq 0$

Constrained Optimization

Lagrange Multipliers

- Can integrate the constraints into the objective function using Lagrange multipliers: (1) becomes

$$L(x, \alpha) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x)$$

  - One Lagrange multiplier $\alpha_i \geq 0$ per constraint $c_i(x)$

- Goal is to now simultaneously minimize $L(x, \alpha)$ wrt primal variables x and maximize $L(x, \alpha)$ wrt dual variables α_i

  - Called a saddle point

- Intuition: if some $c_i(x) > 0$ (i.e. a constraint is violated) then $L(x, \alpha)$ can be increased by increasing α_i, which forces x to change to again decrease L

Constrained Optimization

Karush-Kuhn-Tucker Conditions

- Let $(\bar{x}, \bar{\alpha})$ (where $\bar{x} \in \mathbb{R}^m$ and $\bar{\alpha}_i \geq 0 \ \forall i$) be such that for all $x \in \mathbb{R}^m$ and $\alpha \in [0, \infty)^n$ we have

$$L(\bar{x}, \alpha) \leq L(\bar{x}, \bar{\alpha}) \leq L(x, \bar{\alpha}) \qquad (2)$$

- The first inequality implies that $L(\bar{x}, \alpha) - L(\bar{x}, \bar{\alpha}) \leq 0$, i.e.

$$\sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_i) c_i(\bar{x}) \leq 0$$

- Since (2) holds for all $\alpha_i \geq 0$, set $\alpha_i = \bar{\alpha}_i$ for all $i \neq j$ and $\alpha_j = \bar{\alpha}_j + 1$. Then $c_j(\bar{x}) \leq 0$ for all j, i.e. $\bar{x}$ satisfies the constraints

Constrained Optimization

Karush-Kuhn-Tucker Conditions
(cont'd)

- Further, when instead $\alpha_j = 0$ then $\bar{\alpha}_j c_j(\bar{x}) \geq 0$, which is only possible if $\bar{\alpha}_j c_j(\bar{x}) = 0 \ \forall j$ (this is the KKT condition)

- Combining this with the second inequality of (2):

$$f(\bar{x}) \leq f(x) + \sum_{i=1}^{n} \bar{\alpha}_i c_i(x)$$

- If x is feasible, then $c_i(x) \leq 0$ for all i, implying that $f(\bar{x}) \leq f(x)$ for all feasible $x \Rightarrow \bar{x}$ is optimal

- Thus if (2) holds then $\bar{x}$ is an optimal feasible solution of (1)

  – I.e. satisfying (2) in the Lagrangian yields an optimal solution to the original problem (1) (Thrm 6.21)

  – (2) is also necessary if f and c_i convex and if Lemma 6.23 satisfied

Constrained Optimization

Karush-Kuhn-Tucker Conditions
(cont'd)

T6.26 A solution to (1) with convex, differentiable f and c_i is given by $\bar{x}$ if $\exists\, \bar{\alpha} \in \mathbb{R}^n$ with $\bar{\alpha}_i \geq 0$ s.t. the following are satisfied:

$$\partial_x L(\bar{x}, \bar{\alpha}) = \partial_x f(\bar{x}) + \sum_{i=1}^{n} \bar{\alpha}_i \partial_x c_i(\bar{x}) = 0$$
$$\partial_{\alpha_i} L(\bar{x}, \bar{\alpha}) = c_i(\bar{x}) \leq 0$$
$$\sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x}) = 0$$

Proof ($\top$ means matrix transpose)

$$\begin{aligned} f(x) - f(\bar{x}) &\geq (\partial_x f(\bar{x}))^\top (x - \bar{x}) \\ &= -\sum_{i=1}^{n} \bar{\alpha}_i (\partial_x c_i(\bar{x}))^\top (x - \bar{x}) \\ &\geq -\sum_{i=1}^{n} \bar{\alpha}_i (c_i(x) - c_i(\bar{x})) \\ &= -\sum_{i=1}^{n} \bar{\alpha}_i c_i(x) \geq 0 \end{aligned}$$

- Thus of those x that satisfy c_i, $\bar{x}$ minimizes f

Constrained Optimization

Karush-Kuhn-Tucker Conditions
(cont'd)

- I.e. a solution to the set of equations of T6.26 is a solution to (1)

- Another useful tidbit (T6.27): For any point x that is a feasible solution to (1),

$$f(x) \geq f(\bar{x}) \geq f(x) + \sum_{i=1}^{n} \alpha_i c_i(x)$$

  where $\bar{x}$ is the optimal solution

- I.e. given any feasible point x, we can bound $f(\bar{x})$ in terms of $f(x)$ and $\sum_{i=1}^{n} \alpha_i c_i(x)$

  - Useful stopping criterion for optimization algorithm

Constrained Optimization

Duality

- Consider the following linear program:

$$
\begin{array}{lll}
\underset{x_1,x_2}{\text{minimize}} & 6x_1 + 8x_2 & \\
\text{s.t.} & -3x_1 - x_2 + 4 \leq 0 & \\
 & -5x_1 - 2x_2 + 7 \leq 0 & \quad (3) \\
 & -x_1, -x_2 \leq 0 &
\end{array}
$$

- Now find the Langrangian:

$$
\begin{aligned}
L(x,\alpha) &= 6x_1 + 8x_2 + \alpha_1(-3x_1 - x_2 + 4) \\
&\quad +\alpha_2(-5x_1 - 2x_2 + 7) - \alpha_3 x_1 - \alpha_4 x_2
\end{aligned}
$$

- T6.26 says that for an optimal solution:

$$
\partial_x L(x,\alpha) = \begin{bmatrix} 6 - 3\alpha_1 - 5\alpha_2 - \alpha_3 \\ 8 - \alpha_1 - 2\alpha_2 - \alpha_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
$$

- Which we substitute back into $L(x,\alpha)$ to get $4\alpha_1 + 7\alpha_2$

Constrained Optimization

Duality

(cont'd)

- Recall that we want to maximize wrt α, so equivalent to (3) is

$$\begin{array}{ll} \underset{\alpha_1,\alpha_2}{\text{maximize}} & 4\alpha_1 + 7\alpha_2 \\ \text{s.t.} & 6 - 3\alpha_1 - 5\alpha_2 - \alpha_3 = 0 \\ & 8 - \alpha_1 - 2\alpha_2 - \alpha_4 = 0 \\ & \alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0 \end{array}$$

  (note that we can drop α_3, α_4 and change "$=$" to "$\geq$" in first two inequalities)

- This is the dual (or Wolfe dual) of (3)

- Important properties:
  - Constraints in one correspond to variables in other
  - Value of obj function in primal $\leq$ that for dual; equality at optimal solution
  - We've eliminated the x variables from the primal (we'll use this when applying kernels for SVMs)

Constrained Optimization
Duality
(cont'd)

- Can also find dual of convex quadratic optimization problems:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \frac{1}{2}x^\top K x + c^\top x \\ \text{s.t.} & Ax + d \leq 0 \end{array} \tag{4}$$

where K is $m \times m$ PD matrix, $x, c \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$ and $d \in \mathbb{R}^n$

- Lagrangian is

$$L(x, \alpha) = \frac{1}{2}x^\top K x + c^\top x + \alpha^\top (Ax + d)$$

- Apply T6.26:

$$\partial_x L(x, \alpha) = K^\top x + A^\top \alpha + c = 0 \tag{5}$$
$$\partial_\alpha L(x, \alpha) = Ax + d \leq 0 \tag{6}$$
$$\alpha^\top (Ax + d) = 0 \tag{7}$$
$$\alpha \geq 0 \tag{8}$$

Constrained Optimization
Duality
(cont'd)

- Applying (7) gives $L(x,\alpha) = \frac{1}{2}x^\top Kx + c^\top x$ and further applying (5) and again (7) yields

$$\begin{aligned} L(x,\alpha) &= \frac{1}{2}x^\top Kx + \left(-K^\top x - A^\top \alpha\right)^\top x \\ &= -\frac{1}{2}x^\top Kx - \alpha^\top Ax \\ &= -\frac{1}{2}x^\top Kx + \alpha^\top d \end{aligned}$$

(in book, recall that when PD, $K = K^\top$)

- Now use $x = -K^{-1}(c + A^\top \alpha)$ from (5) and get

$$\begin{aligned} L(x,\alpha) = &-\frac{1}{2}\alpha^\top A^\top K^{-1} A\alpha + \left[d - c^\top K^{-1} A^\top\right]\alpha \\ &-\frac{1}{2}c^\top K^{-1} c \end{aligned}$$

- Last term is constant, so get

$$\begin{array}{ll} \underset{\alpha}{\text{maximize}} & -\frac{1}{2}\alpha^\top A^\top K^{-1} A\alpha + \left[d - c^\top K^{-1} A^\top\right]\alpha \\ \text{s.t.} & \alpha \geq 0 \end{array}$$

as dual to (4)

Topic summary due in 1 week!