Optimization Methods in Machine Learning
Lecture 10: Newton method and Interior Point Method for SVM

Katya Scheinberg

Lehigh University

Spring 2016

Slides from L. Vandenberghe http://www.ee.ucla.edu/ vandenbe/ee236c.html

Newton step

$$\Delta x_{\rm nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

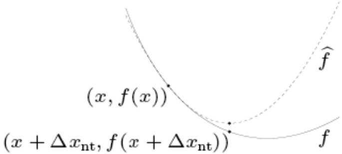
interpretations

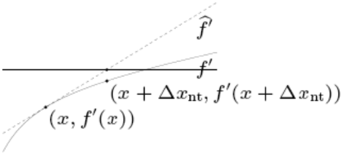
- $x + \Delta x_{\rm nt}$ minimizes second order approximation

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

- $x + \Delta x_{\rm nt}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$





Quadratic Approximation Model

- The problem we deal with is $f(x)$
- The quadratic approximation model is

$$q(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$$

- Newton Step. $\Delta x_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$
- $x_{k+1} = x_k + \Delta x_k$
- Damped Newton step: not the full step.
- If full step is not good, use line search.

Convergence Analysis

- Assumption:

$$\|\nabla^2 f(x) + \nabla^2 f(y)\|_2 \leq L\|y - x\|_2$$

- f is strongly convex with constant m
- We want Hessian to be Lipschitz continuous.
- There exists constants $\eta \in (0, m^2/L)$ and $\gamma > 0$ such that
  - if $\|\nabla f(x^k)\|_2 \geq \eta$, then

$$f(x^{k+1}) - f(x^k) \leq -\gamma$$

  - if $\|\nabla f(x^k)\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^{k+1})\|_2 \leq (\frac{L}{2m^2}\|\nabla f(x^k)\|_2)^2 \qquad \text{(fast converge stage)}$$

where m is the smallest eigenvalue of $\nabla^2 f(x)$ for all x.

Self-concordant

- Newton method is invariant under linear transformation! But the convergence analysis isn't!!
- To have a better analysis, self-concordant functions have been introduced.
- In optimization, a self-concordant function is a function $f : \mathbb{R} \to \mathbb{R}$ such that
$$|f'''(x)| \leq 2f''(x)^{\frac{3}{2}}$$
- $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if
$$g(t) = f(x + t\nu)$$
is self-concordant for all $x \in \mathrm{dom} f, \nu \in \mathbb{R}^n$
- Examples: linear, convex quadratic, logarithm.

Interior Point Method

- Rewrite the quadratic model

$$\begin{array}{c} \min\ \frac{1}{2}x^TQx + c^Tx \\ Ax = b \\ x \geq 0 \end{array} \quad \Rightarrow \quad \begin{array}{c} \min\ \frac{1}{2}x^TQx + c^Tx - \mu\sum_{i=1}^n \text{In } x_i \\ Ax = b \end{array}$$

- KKT conditions are:

$$\begin{aligned} Ax &= b \\ -Qx + A^Ty + s &= c \\ Xs &= \mu e \\ X, s &> 0 \end{aligned}$$

where X =diag(x)

Interior Point Method

- Given (x, y, s), find the Newton step $(\Delta x, \Delta y, \Delta s)$,

$$
\begin{aligned}
A(x + \Delta x) &= b \\
-Q(x + \Delta x) + A^T(y + \Delta y) + s + \Delta s &= c \\
s\Delta X + X\Delta s + Xs &= \mu e
\end{aligned}
$$

- Then we have

$$
\begin{aligned}
S\Delta x + X\Delta s &= \mu e - Xs \\
A\Delta x &= b - Ax = r_p \\
-Q\Delta x + A^T\Delta y + \Delta s &= c - Qx - A^T y - s = r_d
\end{aligned}
$$

Interior Point Method

- Augmented system

$$\begin{array}{rcl} A\Delta x & = & r_p \\ A^T\Delta y - (X^{-1}S + Q)\Delta x & = & r_d - X^{-1}(\mu e - Xs) \end{array}$$

- Eventually, we have the following *Normal Equation*

$$A(X^{-1}S + Q)^{-1}A^T\Delta y = r$$

- Important: $A(X^{-1}S + Q)^{-1}A^T$ is positive definite if A is full row rank.

Optimality conditions for SVM

Consider dual form of SVM as following:

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\
s.t. \quad & \\
& y^T\alpha = 0 \\
& 0 \leq \alpha \leq c
\end{aligned}
$$

Now consider KKT conditions:

$$
\begin{aligned}
& \alpha_i s_i = 0 && i = 1, 2, ..., n \\
& (c - \alpha_i)\xi_i = 0 && i = 1, 2, ..., n \\
& y^T\alpha = 0 && \\
& -Q\alpha + y\beta + s - \xi = -e && \\
& 0 \leq \alpha \leq c && \\
& s \geq 0, \xi \geq 0 &&
\end{aligned}
$$

Relaxed KKT conditions

We want to solve this problem by interior point method, so we rewrite the KKT conditions as following:

$$
\begin{aligned}
&\alpha_i s_i = \mu && i = 1, 2, ..., n \\
&(c - \alpha_i)\xi_i = \mu && i = 1, 2, ..., n \\
&y^T \alpha = 0 \\
&-Q\alpha + y\beta + s - \xi = -e \\
&0 < \alpha < c \\
&s > 0 \\
&\xi > 0
\end{aligned}
$$

A Newton step of IPM

- Let $\mathcal{A} = diag(\alpha)$, $\mathcal{S} = diag(s)$ and $\Xi = diag(\xi)$

$$\begin{pmatrix} y^T & 0 \\ -(Q + \mathcal{A}^{-1}S + (C - \mathcal{A})^{-1}\Xi) & y \end{pmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta\beta \end{pmatrix} = \begin{pmatrix} -y^T\alpha \\ -e + Q\alpha - y\beta - \mathcal{A}^{-1}\mu e + (C - \mathcal{A})^{-1}\mu e \end{pmatrix}$$

- Doing some algebra, we have:

$$y^T(Q + D)^{-1}\Delta\beta = \gamma$$

where

$$D = \mathcal{A}^{-1}\mathcal{S} + (C - \mathcal{A})^{-1}\Xi$$
$$\gamma = -y^T\alpha + y^T(Q + D)^{-1}(-e + Q\alpha - y\beta - \mathcal{A}^{-1}\mu e + (C - \mathcal{A})^{-1}\mu e)$$

Kernel operation

- As we defined before

$$Q_{i,j} = y_i y_j \mathcal{K}(x_i, x_j) \tag{1}$$

where $\mathcal{K}(x_i, x_j)$ is kernel operation of x_i and x_j.

- Some examples for kernel operation:

Linear kernel: $\mathcal{K}(x_i, x_j) = x_i^T x_j$

Quadratic kernel: $\mathcal{K}(x_i, x_j) = [a + b(x_i^T x_j]^2$

RBF kernel: $\mathcal{K}(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma})$

- Q is a Positive Semi-Definite (p.s.d) matrix.

Computation complexity

Back to SVM, consider Q.

- To solve the Newton system, we have to solve

$$y^T(Q+D)^{-1}\Delta\beta = \gamma$$

  where

$$D = \mathcal{A}^{-1}\mathcal{S} + (C-\mathcal{A})^{-1}\Xi$$
$$\gamma = -y^T\alpha + y^T(Q+D)^{-1}(-e + Q\alpha - y\beta - \mathcal{A}^{-1}\mu e + (C-\mathcal{A})^{-1}\mu e)$$

- Q is typically a dense matrix.

$$Q = Y^T X X^T Y$$

  where Y is a $n \times 1$ and X is a $n \times d$, so rank of Q is at most d.
- We need $\mathcal{O}(n^3)$ operations to invert $(Q+D)$.

Scherman-Morrison-Woodbury formula

- Let $Q = VV^T$.
- We can find $(Q+D)^{-1}$ as following

$$\begin{aligned}(Q+D)^{-1} &= (VV^T + D)^{-1} \\ &= D^{-1} - D^{-1}V(I + V^T D^{-1} V)^{-1} V^T D^{-1}\end{aligned}$$

which needs $\mathcal{O}(nd^2)$ operations and $\mathcal{O}(nd)$ storage amount.

Reminder: Matrix calculus, Symmetric matrix

Definition

Matrix $M_{n\times n}$ is symmetric, if and only if $M = M^T$.

- Consider $Q_{n\times n}$ where n is size of training data. We can define Q as following
$$Q = Y^T X X^T Y$$
- Q is symmetric because:
$$Q^T = (Y^T X X^T Y)^T = Y^T X X^T Y = Q$$
- If a matrix is symmetric, its eigen values are real number.

Positive semi-definite and positive definite

Definition

Symmetric matrix $M_{n\times n}$ is positive semi-definite if for all $z_{n\times 1}$ we have:

$$z^T M z \geq 0$$

M is positive definite if $z^T M z > 0$ for all $z_{n\times 1}$.

- Q is positive semi-definite, because

$$z^T Q z =^T Y^T X X^T Y z = (X^T Y z)^T (X^T Y z) = \|X^T Y z\|^2 \geq 0$$

Definition

Symmetric matrix $M_{n\times n}$ is positive semi-definite if all of its eigenvalues are nonnegative.

Eigenvalue decomposition

Definition

Consider symmetric matrix $M_{n \times n}$. We have:

$$M = P \Lambda P^T$$

where Λ is a diagonal matrix that elements on its main diagonal corresponds to eigenvalues of M, and P is an orthogonal matrix. Columns of P correspond to eigenvectors of M.

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & ... & 0 & 0 \\ 0 & \lambda_2 & 0 & ... & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & ... & 0 & \lambda_n \end{bmatrix}$$

$$P = \begin{bmatrix} P_1 \mid & P_2 \mid & ... & \mid P_n \end{bmatrix}$$

Eigenvalue decomposition (continued)

Definition

Rank of a matrix is the number of linear independent columns or linear independent rows of the matrix.

- The rank of p.s.d. matrix is the number of positive eigen values.

$$Q = P\Lambda P^T = \sum_i \lambda_i P_i P_i^T$$

where $P_i P_i^T (i = 1, 2, ..., n)$ are rank-one matrices.