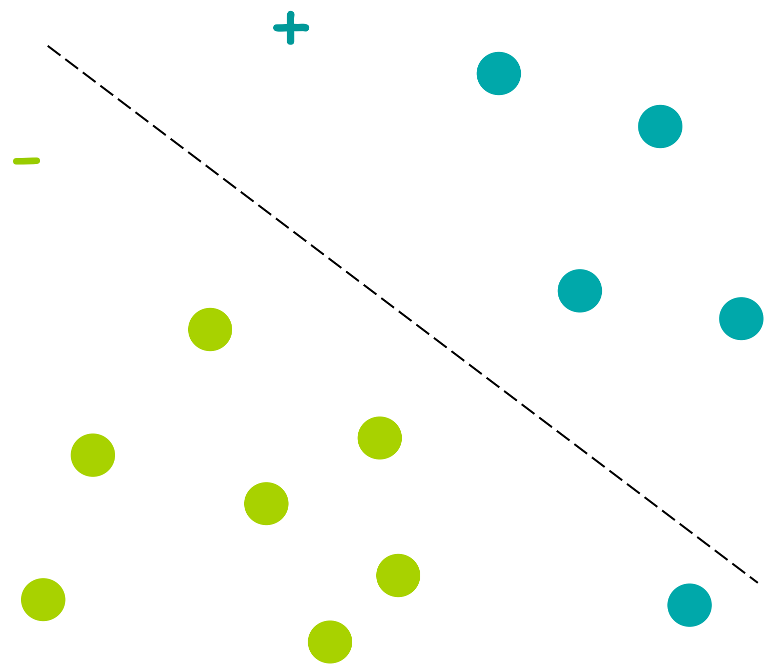
Optimization Methods in Machine Learning

Lecture 12,

IPMs for SVMs

An Interior Point Method for SVM

Support Vector Machines



Optimization Problem

Two sets of points: $X_+ \subset \mathbf{R}^p$ and $X_- \subset \mathbf{R}^p$
Total number of points: n

$$\begin{array}{ll} \min_{\xi, w, \beta} & \frac{1}{2} w^\top w + c \sum_{i=1}^{n} \xi_i \\ \text{s.t.} & -y_i(w^\top x_i + \beta) \leq -1 + \xi_i, \quad i = 1, \ldots, n \\ & \xi \geq 0, \qquad i = 1, \ldots, n. \end{array}$$

$$y_i = 1 \ \text{ if } \ x_i \in X_+$$

$$y_i = -1 \ \text{ if } \ x_i \in X_-$$

Optimization Problem

At optimality $w^* = \sum_{i=1}^n \alpha_i y_i x_i, \quad 0 \le \alpha_i \le c$

$$\begin{array}{ll} \min_{\alpha,\beta,\xi} & \frac{1}{2}\alpha^\top Q\alpha + c\sum_{i=1}^n \xi_i \\ \text{s.t.} & -Q\alpha + y\beta + s_i - \xi_i = -1, \quad i = 1,\ldots,n \\ & s_i \ge 0, \xi \ge 0, 0 \le \alpha_i \le c, \qquad i = 1,\ldots,n, \end{array}$$

$$Q := D_y X X^\top D_y \quad \Leftrightarrow \quad Q_{ij} = y_i y_j {x_i}^\top x_j$$

$$\begin{array}{ll} \min_{\alpha} & \frac{1}{2}\alpha^\top Q\alpha - e^\top \alpha \\ \text{s.t.} & y^\top \alpha = 0, \\ & 0 \le \alpha \le c, \end{array}$$

Kernel SVM

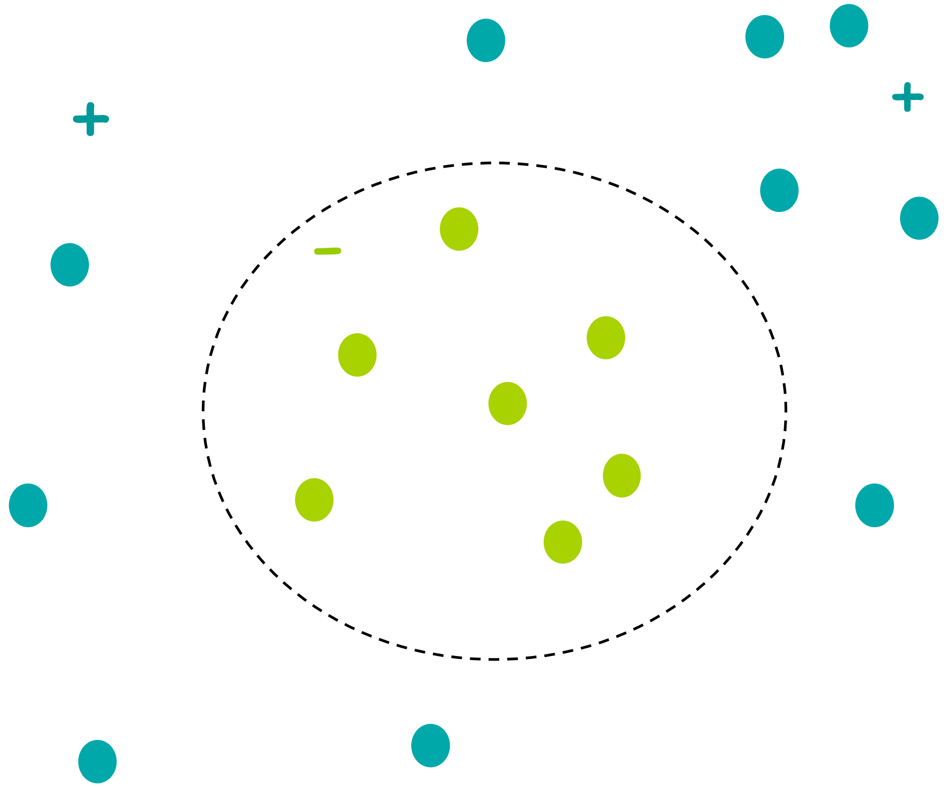
$Q_{ij} = y_i y_j {x_i}^\top x_j \;\rightarrow\; Q_{ij} = y_i y_j \phi(x_i)^\top \phi(x_j) = y_i y_j K(x_i, x_j)$

Kernel operation: $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$

Examples:

- $K(x_i, x_j) = \exp^{-||x_i - x_j||^2/2\sigma^2}$

- $K(x_i, x_j) = ({x_i}^\top x_j / a_1 + a_2)^d$



Optimality Conditions

Dual problem

$$
\begin{array}{rl}
\min_{\alpha} & \dfrac{1}{2}\alpha^\top Q\alpha - e^\top \alpha \\
\text{s.t.} & y^\top \alpha = 0, \\
 & 0 \le \alpha \le c,
\end{array}
$$

KKT conditions

$$
\begin{array}{l}
\alpha_i s_i = 0, \quad i = 1, \ldots, n, \\
(c - \alpha_i)\xi_i = 0, \quad i = 1, \ldots, n, \\
y^\top \alpha = 0, \\
-Q\alpha + y\beta + s - \xi = -e, \\
0 \le \alpha \le c, \ s \ge 0, \ \xi \ge 0.
\end{array}
$$

Relaxed KKT conditions

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

A Newton step of IPM

Linearize perturbed KKT conditions

$$\alpha_i \Delta s_i + s_i \Delta \alpha_i = \mu - s_i \alpha_i, \quad i = 1, \ldots, n,$$
$$(c - \alpha_i)\Delta \xi_i - \xi \Delta \alpha_i = \mu - (c - \alpha_i)\xi_i \quad i = 1, \ldots, n,$$
$$y^\top \Delta \alpha = -y^\top \alpha,$$
$$-Q\Delta\alpha + y\Delta\beta + \Delta s - \Delta\xi = -e + Q\alpha - y\beta - s + \xi,$$

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

$$\Delta s = \mathcal{A}^{-1}\mu e - s - \mathcal{A}^{-1}S\Delta\alpha,$$
$$\Delta\xi = (C - \mathcal{A})^{-1}\mu e - \xi + (C - \mathcal{A})^{-1}\Xi\Delta\alpha,$$
$$y^\top \Delta \alpha = -y^\top \alpha,$$
$$-Q\Delta\alpha + y\Delta\beta + \Delta s - \Delta\xi = -e + Q\alpha - y\beta - s + \xi,$$

Solving the linear system

$$
\begin{aligned}
&y^\top \Delta\alpha = -y^\top \alpha, \\
&-Q\Delta\alpha + y\Delta\beta + \mathcal{A}^{-1}\mu e - s - \mathcal{A}^{-1}S\Delta\alpha - (C-\mathcal{A})^{-1}\mu e - \xi - (C-\mathcal{A})^{-1}\Xi\Delta\alpha \\
&= -e + Q\alpha - y\beta - s + \xi,
\end{aligned}
$$

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

Solving the linear system

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

$$[y^\top (Q + \mathcal{A}^{-1}S + (C - \mathcal{A})^{-1}\Xi)^{-1} y]\Delta\beta = -y^\top \alpha +$$
$$y^\top (Q + \mathcal{A}^{-1}S + (C - \mathcal{A})^{-1}\Xi)^{-1}(-e + Q\alpha - y\beta - \mathcal{A}^{-1}\mu e - (C - \mathcal{A})^{-1}\mu e)$$

Forming storing inverting matrices

Need to compute Q+D and solve a system of lin equation with it (factorize)

$$Q_{ij} = y_i y_j K(x_i, x_j)$$

Q is n×n, typically dense matrix

- $K(x_i, x_j) = \exp^{-||x_i - x_j||^2/2\sigma^2}$
- $K(x_i, x_j) = ({x_i}^\top x_j/a_1 + a_2)^d$

Q+D has to be factorized at each iteration – $O(n^3)$ flops

Scherman-Morrison-Woodbury formula

$$(M + UV^\top)^{-1} = M^{-1} - M^{-1}U(I + V^\top M^{-1}U)^{-1}V^\top M^{-1}$$

$$\text{Let } Q_{ij} = y_i y_j \phi(x_i)^\top \phi(x_j) \Rightarrow Q = VV^\top$$

k = the number of columns in V is the dimension of feature space - ϕ(x)

$$(D + VV^\top)^{-1} = \underset{O(n)}{D^{-1}} - \underset{O(nk)}{D^{-1}V}\overset{O(k^3)}{(I + \underset{O(k^2n)}{V^\top D^{-1}V})^{-1}}\underset{O(nk)}{V^\top D^{-1}}$$

Per iteration complexity is O(nk^2) and storage is O(nk)

Return to the linear formulation

$$\min_{\xi,w,\beta} \quad \frac{1}{2}w^\top w + c\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad -y_i(w^\top x_i + \beta) \le -1 + \xi_i, \quad i = 1, \ldots, n$$

$$\xi \ge 0, \qquad i = 1, \ldots, n.$$

$$L(w, b, \alpha, \xi) = \frac{1}{2}w^\top w + c\sum_{i=1}^{n}\xi_i - \sum_i \alpha_i(y_i(w^\top x_i + \beta) - 1 + \xi_i) - \nu^\top \xi$$

$$\nabla_w L = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$

$$\nabla_\xi L = c - \alpha - \nu = 0$$

$$\nabla_\beta L = y^\top \alpha = 0$$

$$\alpha \ge 0, \quad \nu \ge 0$$

Optimality Conditions

Perturbed KKT conditions

$$
\begin{aligned}
&\alpha_i s_i = \mu, \quad i = 1, \ldots, n, \\
&(c - \alpha_i)\xi_i = \mu, \quad i = 1, \ldots, n, \\
&y^\top \alpha = 0, \\
&w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, \\
&-y_i(w^\top x_i + \beta) + s_i - \xi_i = -1, \quad i = 1, \ldots, n, \\
&0 < \alpha < c, \; s > 0, \; \xi < 0.
\end{aligned}
$$

Or in vector matrix terms…

Perturbed KKT conditions

$$
\begin{aligned}
&\mathcal{A}s = \mu e, \\
&(C - \mathcal{A})\xi = \mu e \\
&y^\top \alpha = 0, \\
&w - (YX)^\top \alpha = 0, \\
&-YXw - y\beta + s - \xi = -e
\end{aligned}
$$

The Newton system

$$\mathcal{A}\Delta s + S\Delta\alpha = \mu e - \mathcal{A}s$$
$$(C-\mathcal{A})\Delta\xi - \Xi\Delta\alpha = \mu e - (C-\mathcal{A})\xi$$
$$y^\top \Delta\alpha = -y^\top \alpha,$$
$$\Delta w - (YX)^\top \Delta\alpha = -w + (YX)^\top \alpha,$$
$$-YX\Delta w - y\Delta\beta + \Delta s - \Delta\xi = -e - YXw - y\beta - s + \xi$$

$$\Downarrow$$

$$y^\top \Delta\alpha = -y^\top \alpha,$$
$$\Delta w - (YX)^\top \Delta\alpha = -w + (YX)^\top \alpha,$$
$$-YX\Delta w - y\Delta\beta - (\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi)\Delta\alpha =$$
$$-e + YXw + y\beta - s + \xi - \mathcal{A}^{-1}\mu e + s + (C-\mathcal{A})^{-1}\mu e - \xi$$

The Newton system

$$\begin{bmatrix} (\mathcal{A}^{-1}S + (C - \mathcal{A})^{-1}\Xi) & YX & y \\ (YX)^\top & -I & 0 \\ y^\top & 0 & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta w \\ \Delta\beta \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

The Newton system

$$\begin{bmatrix} -(\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi) & YX & y \\ (YX)^\top & -I & 0 \\ y^\top & 0 & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta w \\ \Delta\beta \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

$$I + (YX)^\top (\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi)^{-1} YX \Delta w - y\Delta\beta = r$$

or

$$(YX(YX)^\top + (\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi)^{-1})\Delta\alpha - (YX)y\Delta\beta = r$$

The Newton system

$$\begin{bmatrix} (\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi) & YX & y \\ (YX)^\top & -I & 0 \\ y^\top & 0 & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta w \\ \Delta\beta \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

$O(nk^2)$

$$I + (YX)^\top(\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi)^{-1}YX\Delta w - y\Delta\beta = r$$

or

$O(n^3)$

$$(YX(YX)^\top + (\mathcal{A}^{-1}S + (C-\mathcal{A})^{-1}\Xi)^{-1})\Delta\alpha - (YX)y\Delta\beta = r$$

Complexity

Interior point method for nonlinear SVMs:

- Need to solve $(Q+D)p=r$. Q is completely dense, with rank k_d.
- If $k_d \sim n$, then each IPM iteration is $O(n^3)$ operations and $O(n^2)$ memory.
- If $k_d << n$, then each IPM iteration is $O(nk_d^2)$ operations and $O(nk_d)$ memory.

Interior point method for linear SVMs:

- Can solve $(VV^\top + D)p = r$ or $(I + VD^{-1}V^\top)p = r$.
- $V = XY$ and is as sparse as the data. In large scale cases V can be sparse and the complexity per step is similar to linear programming.

Optimization methods for convex problems

- Interior Point methods
  - Best iteration complexity O(log(1/ϵ)), in practice <50.
  - Worst per-iteration complexity (sometimes prohibitive)
- Active set methods
  - Exponential complexity in theory, often linear in practice.
  - Better per iteration complexity.
- Gradient based methods
  - $O(1/\sqrt{\epsilon})$or O(1/ϵ) iterations
  - Matrix/vector multiplication per iteration
- Nonsmooth gradient based methods
  - O(1/ϵ) or O(1/ϵ^2) iterations
  - Matrix/vector multiplication per iteration
- Block coordinate descent
  - Iteration complexity ranges from unknown to similar to FOMs.
  - Per iteration complexity can be constant.