Support Vector Clustering

Asa Ben-Hur, David Horn, Hava T. Siegelmann, Vladimir Vapnik

Wittawat Jitkrittum

Gatsby tea talk

31 July 2015

Overview

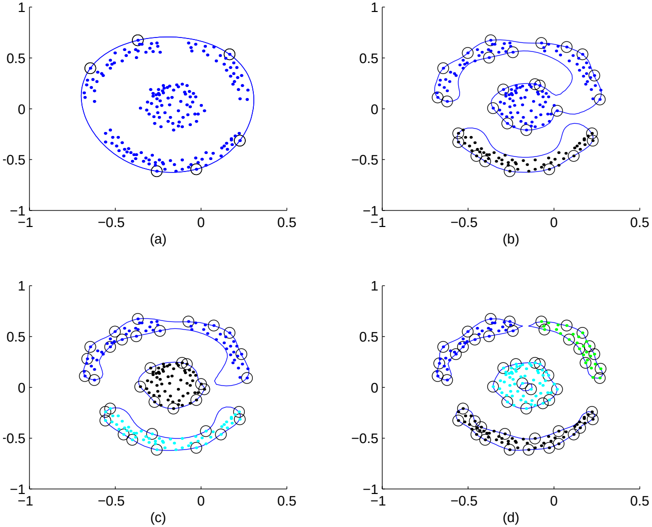
```
Support Vector Clustering
Asa Ben-Hur, David Horn, Hava T. Siegelmann, Vladimir Vapnik
Journal of Machine Learning Research, 2001.
```

- Main algorithm based on

  ```
  Support vector domain description
  David M.J Tax, Robert P.W Duin
  Pattern Recognition Letters, 1999.
  ```

- **Goal**: Divide $\{x_i\}_{i=1}^{N}$ into disjoint groups.
- **Idea**:
  1. Map x_i to $\phi(x_i)$ (RKHS).
  2. Find the minimal enclosing sphere in RKHS.
  3. Sphere in RKHS $=$ non-linear contours in the original space.
  4. Interpret the contours as the cluster boundaries.

Example



- (a),..,(d): From high to low Gaussian widths.

Support Vector Clustering (SVC)

Given $\{x_j\}_{j=1}^N$, find the smallest enclosing sphere of radius R. $a \in \mathcal{H}$ (RKHS).

$$\begin{aligned} &\min_{R,a} R^2 \\ \text{s.t.} \quad &\|\phi(x_j) - a\|_{\mathcal{H}}^2 \leq R^2. \end{aligned}$$

Soft constraints with slack variables ξ_j:

$$\begin{aligned} &\min_{R,a,\{\xi_j\}_j} R^2 + C \sum_{j=1}^N \xi_j \\ \text{s.t.} \quad &\|\phi(x_j) - a\|_{\mathcal{H}}^2 \leq R^2 + \xi_j, \\ &\xi_j \geq 0. \end{aligned}$$

- Convex problem. One optimum.

Support Vector Clustering (SVC)

Given $\{x_j\}_{j=1}^N$, find the smallest enclosing sphere of radius R. $a \in \mathcal{H}$ (RKHS).

$$\min_{R,a} R^2$$
$$\text{s.t.} \quad \|\phi(x_j) - a\|_{\mathcal{H}}^2 \leq R^2.$$

Soft constraints with slack variables ξ_j:

$$\min_{R,a,\{\xi_j\}_j} R^2 + C \sum_{j=1}^N \xi_j$$
$$\text{s.t.} \quad \|\phi(x_j) - a\|_{\mathcal{H}}^2 \leq R^2 + \xi_j,$$
$$\xi_j \geq 0.$$

- Convex problem. One optimum.

Solving SVC

With dual variables $\{\beta_j\}_j$ and $\{\mu_j\}_j$, Lagragian is

$$L = R^2 + C\sum_{j=1}^{N}\xi_j - \sum_{j=1}^{N}\underbrace{\left(R^2 + \xi_j - \|\phi(x_j) - a\|_{\mathcal{H}}^2\right)}_{\geq 0}\beta_j - \sum_{j=1}^{N}\underbrace{\xi_j}_{\geq 0}\mu_j.$$

Setting $\frac{\partial L}{\partial R} = 0, \frac{\partial L}{\partial a} = 0, \frac{\partial L}{\partial \xi_j} = 0$ leads to stationarity conditions

1. $1 = \sum_{j=1}^{N}\beta_j$
2. $a = \sum_{j=1}^{N}\beta_j\phi(x_j)$, linear combination of the mapped training points
3. $\beta_j = C - \mu_j$

KKT complementarity conditions (necessary for optimality)

1. $\left(R^2 + \xi_j - \|\phi(x_j) - a\|_{\mathcal{H}}^2\right)\beta_j = 0$
2. $\xi_j\mu_j = 0$

Solving SVC

With dual variables $\{\beta_j\}_j$ and $\{\mu_j\}_j$, Lagragian is

$$L = R^2 + C\sum_{j=1}^{N}\xi_j - \sum_{j=1}^{N}\underbrace{\left(R^2 + \xi_j - \|\phi(x_j) - a\|_{\mathcal{H}}^2\right)}_{\geq 0}\beta_j - \sum_{j=1}^{N}\underbrace{\xi_j}_{\geq 0}\mu_j.$$

Setting $\frac{\partial L}{\partial R} = 0, \frac{\partial L}{\partial a} = 0, \frac{\partial L}{\partial \xi_j} = 0$ leads to stationarity conditions

1. $1 = \sum_{j=1}^{N}\beta_j$
2. $a = \sum_{j=1}^{N}\beta_j\phi(x_j)$, linear combination of the mapped training points
3. $\beta_j = C - \mu_j$

KKT complementarity conditions (necessary for optimality)

1. $\left(R^2 + \xi_j - \|\phi(x_j) - a\|_{\mathcal{H}}^2\right)\beta_j = 0$
2. $\xi_j\mu_j = 0$

Analysis of Support Vectors ($\beta_j > 0$)

A: Constraints

1. $\|\phi(x_j) - a\|^2_{\mathcal{H}} \leq R^2 + \xi_j$
2. $\xi_j, \beta_j, \mu_j \geq 0$

B: Complementarity conditions

1. $\left(R^2 + \xi_j - \|\phi(x_j) - a\|^2_{\mathcal{H}}\right) \beta_j = 0$
2. $\xi_j \mu_j = 0$

C: Stationarity conditions

1. $1 = \sum_{j=1}^{N} \beta_j$
2. $a = \sum_{j=1}^{N} \beta_j \phi(x_j)$
3. $\beta_j = C - \mu_j$

- Consider $0 < \beta_j < C$. C3 $\Rightarrow \mu_j > 0$. B2 $\Rightarrow \xi_j = 0$. B1 $\Rightarrow$ $\|\phi(x_j) - a\|^2_{\mathcal{H}} = R^2$. $\phi(x_j)$ lies on the sphere surface. Call x_j a "support vector" (SV).
- Call x_i with $\xi_i > 0$ a "bounded support vector" (BSV). $\xi_i > 0$ means $\phi(x_i)$ lies outside the sphere by A1. B2 $\Rightarrow \mu_j = 0$. C3 $\Rightarrow \beta_j = C$.
- So, low C limits the influence of a BSV on the sphere.

Analysis of Support Vectors ($\beta_j > 0$)

A: Constraints

1. $\|\phi(x_j) - a\|^2_{\mathcal{H}} \leq R^2 + \xi_j$
2. $\xi_j, \beta_j, \mu_j \geq 0$

B: Complementarity conditions

1. $\left(R^2 + \xi_j - \|\phi(x_j) - a\|^2_{\mathcal{H}}\right) \beta_j = 0$
2. $\xi_j \mu_j = 0$

C: Stationarity conditions

1. $1 = \sum_{j=1}^N \beta_j$
2. $a = \sum_{j=1}^N \beta_j \phi(x_j)$
3. $\beta_j = C - \mu_j$

- Consider $0 < \beta_j < C$. C3 $\Rightarrow \mu_j > 0$. B2 $\Rightarrow \xi_j = 0$. B1 $\Rightarrow$ $\|\phi(x_j) - a\|^2_{\mathcal{H}} = R^2$. $\phi(x_j)$ lies on the sphere surface. Call x_j a "support vector" (SV).
- Call x_i with $\xi_i > 0$ a "bounded support vector" (BSV). $\xi_i > 0$ means $\phi(x_i)$ lies outside the sphere by A1. B2 $\Rightarrow \mu_j = 0$. C3 $\Rightarrow \beta_j = C$.
- So, low C limits the influence of a BSV on the sphere.

Analysis of Support Vectors ($\beta_j > 0$)

A: Constraints

1. $\|\phi(x_j) - a\|^2_{\mathcal{H}} \leq R^2 + \xi_j$
2. $\xi_j, \beta_j, \mu_j \geq 0$

B: Complementarity conditions

1. $\left(R^2 + \xi_j - \|\phi(x_j) - a\|^2_{\mathcal{H}}\right) \beta_j = 0$
2. $\xi_j \mu_j = 0$

C: Stationarity conditions

1. $1 = \sum_{j=1}^{N} \beta_j$
2. $a = \sum_{j=1}^{N} \beta_j \phi(x_j)$
3. $\beta_j = C - \mu_j$

- Consider $0 < \beta_j < C$. C3 $\Rightarrow \mu_j > 0$. B2 $\Rightarrow \xi_j = 0$. B1 $\Rightarrow$ $\|\phi(x_j) - a\|^2_{\mathcal{H}} = R^2$. $\phi(x_j)$ lies on the sphere surface. Call x_j a "support vector" (SV).
- Call x_i with $\xi_i > 0$ a "bounded support vector" (BSV). $\xi_i > 0$ means $\phi(x_i)$ lies outside the sphere by A1. B2 $\Rightarrow \mu_j = 0$. C3 $\Rightarrow \beta_j = C$.
- So, low C limits the influence of a BSV on the sphere.

Dual Problem

- Substituting the stationarity conditions into L gives

$$\max_{\{\beta_j\}_j} \sum_{j=1}^{N} \beta_j k(x_j, x_j) - \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j k(x_i, x_j)$$

$$\text{s.t.} \quad \sum_{j=1}^{N} \beta_j = 1, \qquad 0 \leq \beta_j \leq C$$

- μ_j dropped. $\beta_j = C - \mu_j$ replaced by $0 \leq \beta_j \leq C$.
- $\{\beta_j\}_j$ used to form $a = \sum_{j=1}^{N} \beta_j \phi(x_j)$ (sphere center).

Sphere Enclosure

- A point y is inside the sphere if

$$f(y) := \|\phi(y) - a\|_{\mathcal{H}} \leq R,$$

where radius $R := \|\phi(x_i) - a\|_{\mathcal{H}}$ and x_i is a SV i.e., $\beta_i < C$.

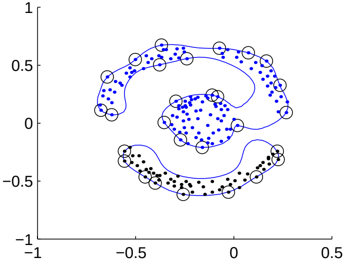
- $f(y)$ is used for cluster assignment.
- Easy to compute $f(y)$:

$$f(y) = \sqrt{k(y,y) - 2\sum_{j=1}^{N} \beta_j k(x_j, y) + \sum_{i=1}^{N}\sum_{j=1}^{N} \beta_i \beta_j k(x_i, x_j)}.$$

- Contour in data space:

$$\{y \mid \|\phi(y) - a\|_{\mathcal{H}} = R\}.$$

Cluster Assignment



(b)

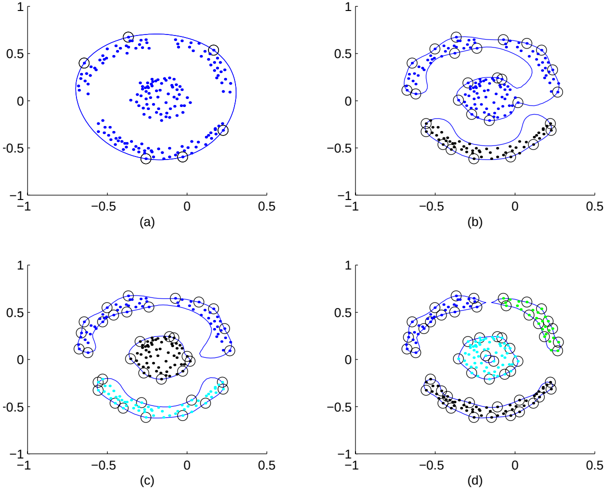
- Given two points from different clusters, any path that connects them must exit from the sphere.

- Define an adjacency matrix $A \in \{0,1\}^{N \times N}$:

$$A_{ij} = \begin{cases} 1 & \text{if for all } y \text{ on the line segment connecting } x_i, x_j,\ f(y) \leq R \\ 0 & \text{otherwise} \end{cases}$$

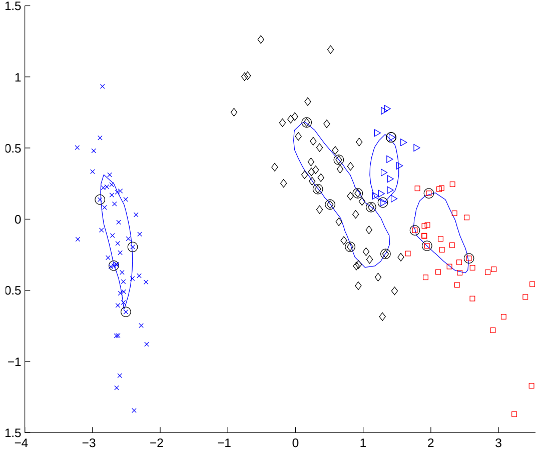
- Clusters $:=$ connected components of the graph induced by A.
- Implemented by sampling a number of points.
- BSVs can be treated as outliers, or assigned to closest cluster.

Example



- $k(x, y) = \exp(-q\|x - y\|^2)$
- (a): $q = 1$. (b): $q = 20$. (c): $q = 24$. (d): $q = 48$.
- Increasing q (decreasing width): boundary fits more tightly

Iris Data



- Iris classification data. 3 classes. 4 dimensions.
- Project to first two principal components.

References I