Survival Support Vector Machines

Dario Comanducci

Tesina per il Corso di Health Analytics and Data-Driven Medicine

---

Master in Data Science and Statistical Learning
Università degli Studi di Firenze

20 Dicembre 2024

Survival Analysis

La Survival Analysis analizza la durata del tempo fino al verificarsi di un evento
Impiegata in vari ambiti:

- medico (aspettativa di vita)
- ingegneristico (tempo di rottura di un componente meccanico, elettronico)
- economico (es. tempo al fallimento di un'azienda)
- scienze sociali (es. tempo dal matrimonio al divorzio)

Survival Analysis

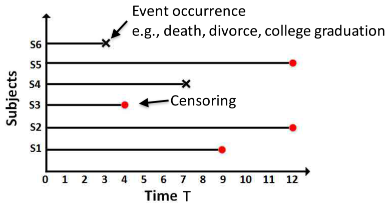
La Survival Analysis analizza la durata del tempo fino al verificarsi di un evento

Impiegata in vari ambiti:

- medico (aspettativa di vita)
- ingegneristico (tempo di rottura di un componente meccanico, elettronico)
- economico (es. tempo al fallimento di un'azienda)
- scienze sociali (es. tempo dal matrimonio al divorzio)

Occorre modellare dati di tempo all'evento:

- il risultato è il tempo y fino al verificarsi di un evento d'interesse
- tale tempo è trattato come variabile dipendente rispetto alle restanti variabili $\mathbf{x}$ dei dati
- dati caratterizzati da censure, $y_i = \min(T_i, C_i)$

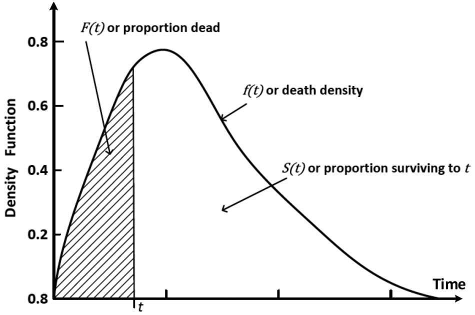


$(\mathbf{x}_i, y_i, \delta_i = \mathrm{I}(\text{evento osservato}))$

Approccio statistico
Ingredienti

- Funzione di sopravvivenza
  $S(t) = \mathrm{P}(T \geq t)$
- Incidenza cumulata
  $F(t) = \mathrm{P}(T < t) = 1 - S(t)$
- Densità della mortalità
  $f(t) = \mathrm{d}F/\mathrm{d}t$
- Hazard
  $$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = f(t)/S(t)$$
- funzione cumulata di Hazard
  $H(t) = \int_0^t h(\tau)\mathrm{d}\tau$

$$S(t) = \exp(-H(t)) \iff H(t) = -\log S(t)$$



- Funzione di sopravvivenza
  $S(t) = \mathrm{P}(T \geq t)$
- Incidenza cumulata
  $F(t) = \mathrm{P}(T < t) = 1 - S(t)$
- Densità della mortalità
  $f(t) = \mathrm{d}F/\mathrm{d}t$
- Hazard
  $$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = f(t)/S(t)$$
- funzione cumulata di Hazard
  $H(t) = \int_0^t h(\tau)\mathrm{d}\tau$

$$S(t) = \exp(-H(t)) \iff H(t) = -\log S(t)$$

Parametrizzazione θ per $f(t)$ o $h(t)$

- Esponenziale, $\theta = \lambda > 0$
  $h(t) = \exp(\lambda t)$
- Weibull, $\theta = (\lambda, \phi > 0)$
  $h(t) = \phi t^{\phi - 1} \exp(\lambda t)$
- Gompertz, $\theta = (\lambda, \phi > 0)$
  $h(t) = \exp(\lambda t) \exp(\phi t)$

θ è funzione di $\{(\mathbf{x}_i, y_i, \delta_i)\}$, legame modellabile anche con tecniche di ML

Bias nel modello in caso di distribuzione non adeguata ai dati

Approccio statistico

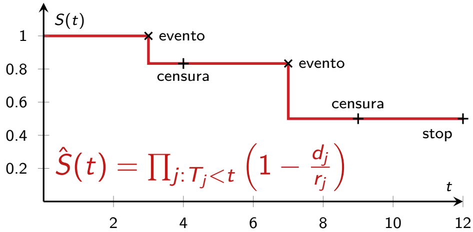
Metodi non parametrici (Kaplan-Meier)

- **Funzione di sopravvivenza**
  $S(t) = \mathrm{P}(T \geq t)$
- **Incidenza cumulata**
  $F(t) = \mathrm{P}(T < t) = 1 - S(t)$
- **Densità della mortalità**
  $f(t) = \mathrm{d}F/\mathrm{d}t$
- **Hazard**
  $$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = f(t)/S(t)$$
- **funzione cumulata di Hazard**
  $H(t) = \int_0^t h(\tau)\mathrm{d}\tau$

$$S(t) = \exp(-H(t)) \iff H(t) = -\log S(t)$$

K eventi su N istanze
$\mathcal{T} = \{T_1 \geq \ldots \geq T_K\}$, $N - K$ censure
Per $T_j \in \mathcal{T}$, siano:

- $d_j \geq 1$ # eventi osservati fino T_j
- $c_{j-1} = \#$ di censure in (T_{j-1}, T_j)
- soggetti a rischio:
  $r_j = r_{j-1} - d_j - c_{j-1}$



- Funzione di sopravvivenza
  $S(t) = \mathrm{P}(T \geq t)$
- Incidenza cumulata
  $F(t) = \mathrm{P}(T < t) = 1 - S(t)$
- Densità della mortalità
  $f(t) = \mathrm{d}F/\mathrm{d}t$
- Hazard
  $$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = f(t)/S(t)$$
- funzione cumulata di Hazard
  $H(t) = \int_0^t h(\tau)\mathrm{d}\tau$

$$S(t) = \exp(-H(t)) \iff H(t) = -\log S(t)$$

Ipotesi di hazard proporzionali
$h(t|\mathbf{x}) = h_0(t) \exp(\beta^\top \mathbf{x})$

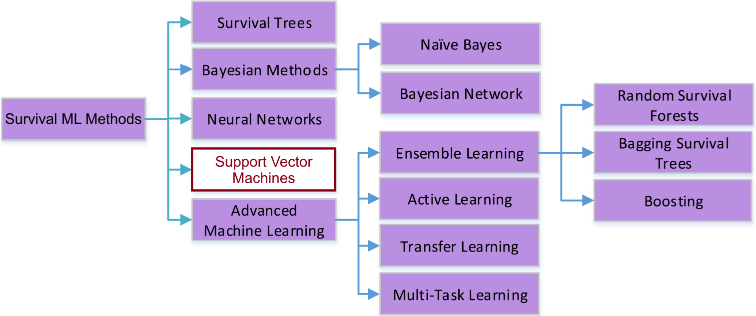
- $h_0(t) \geq 0$: parte non parametrica (hazard di base)
- $\exp(\beta^\top \mathbf{x})$: parte parametrica

$$\hat{\beta} = \arg\max_{\beta} \overbrace{\prod_i \left(\frac{\exp(\beta^\top \mathbf{x}_i)}{\sum_{j \in \mathcal{R}_i} \exp(\beta^\top \mathbf{x}_j)} \right)}^{\text{partial likelihood}}$$

$\mathcal{R}_i$: soggetti a rischio al tempo T_i

$$S(t) = \exp\left(-H_0(t)\exp\left(\beta^\top \mathbf{x}\right)\right) = \exp(-H_0(t))^{\exp\left(\beta^\top \mathbf{x}\right)}$$

Approccio tramite Machine Learning

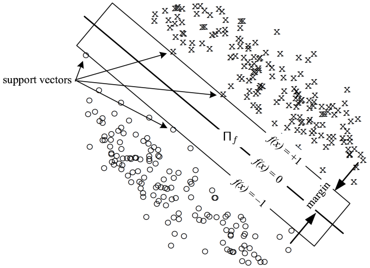


Survival Support Vector Machines

Specializzano

- Regression SVM
- Ranking SVM

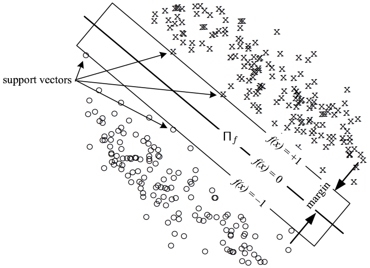
- $(\mathbf{x}_i, z_i) \quad \mathbf{x}_i \in \mathbb{R}^d, z_i \in \{\pm 1\} \quad (i = 1 \dots n)$
- $f(\mathbf{x}|\mathbf{w}, b) = \mathbf{w}^\top \mathbf{x} + b$



- $(\mathbf{x}_i, z_i) \quad \mathbf{x}_i \in \mathbb{R}^d, z_i \in \{\pm 1\} \quad (i = 1 \dots n)$
- vincoli di separabilità: $\forall(\mathbf{x}_i, y_i)$ deve valere

$$z_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \equiv \begin{cases} f(\mathbf{x}) \geq +1 & (z_i = +1) \\ f(\mathbf{x}) \leq -1 & (z_i = -1) \end{cases}$$

- $f(\mathbf{x}|\mathbf{w}, b) = \mathbf{w}^\top \mathbf{x} + b$



Support Vector Machines

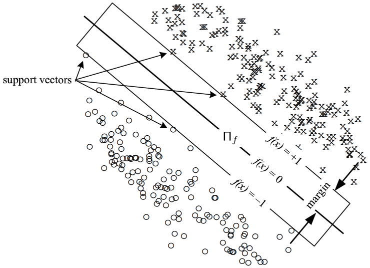
Dati linearmente separabili

- $(\mathbf{x}_i, z_i) \quad \mathbf{x}_i \in \mathbb{R}^d, z_i \in \{\pm 1\} \quad (i = 1 \ldots n)$
- vincoli di separabilità: $\forall(\mathbf{x}_i, y_i)$ deve valere

$$z_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \equiv \begin{cases} f(\mathbf{x}) \geq +1 & (z_i = +1) \\ f(\mathbf{x}) \leq -1 & (z_i = -1) \end{cases}$$

- iperpiano $\Pi_f = \{\mathbf{x} : f(\mathbf{x}) = 0\}$ con il massimo *margine* di separazione tra le due classi

- $f(\mathbf{x}|\mathbf{w}, b) = \mathbf{w}^\top \mathbf{x} + b$

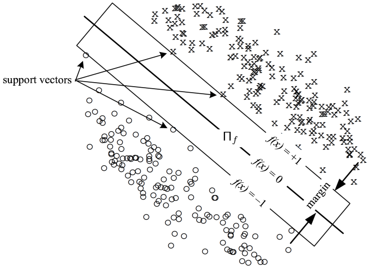


- $(\mathbf{x}_i, z_i) \quad \mathbf{x}_i \in \mathbb{R}^d, z_i \in \{\pm 1\} \quad (i = 1 \ldots n)$
- vincoli di separabilità: $\forall(\mathbf{x}_i, y_i)$ deve valere

$$z_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \equiv \begin{cases} f(\mathbf{x}) \geq +1 & (z_i = +1) \\ f(\mathbf{x}) \leq -1 & (z_i = -1) \end{cases}$$

- iperpiano $\Pi_f = \{\mathbf{x} : f(\mathbf{x}) = 0\}$ con il massimo *margine* di separazione tra le due classi

- $f(\mathbf{x}|\mathbf{w}, b) = \mathbf{w}^\top \mathbf{x} + b$
- $\Pi_{+1}, \Pi_{-1} \parallel \Pi_f$ t. c.
  $\Pi_{+1} : f(\mathbf{x}) = +1$
  $\Pi_{-1} : f(\mathbf{x}) = -1$
- $d_\perp(\Pi_{+1}, \Pi_{-1}) = \frac{2}{\|\mathbf{w}\|}$



- $(\mathbf{x}_i, z_i) \quad \mathbf{x}_i \in \mathbb{R}^d, z_i \in \{\pm 1\} \quad (i = 1 \dots n)$
- vincoli di separabilità: $\forall (\mathbf{x}_i, y_i)$ deve valere

$$z_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \equiv \begin{cases} f(\mathbf{x}) \geq +1 & (z_i = +1) \\ f(\mathbf{x}) \leq -1 & (z_i = -1) \end{cases}$$

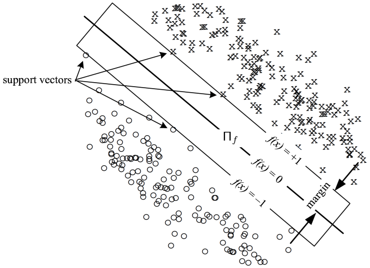
- iperpiano $\Pi_f = \{\mathbf{x} : f(\mathbf{x}) = 0\}$ con il massimo *margine* di separazione tra le due classi

Problema primale

$$\hat{\Pi}_f = \arg\max_{\mathbf{w}} \frac{2}{\|\mathbf{w}\|} = \arg\min_{\mathbf{w}} \overbrace{\frac{1}{2}\|\mathbf{w}\|^2}^{c(\mathbf{w})}$$

$$\text{t. c. } \underbrace{1 - z_i f(\mathbf{x}_i)}_{g_i(\mathbf{w}, b)} \leq 0 \quad \forall i$$

- $f(\mathbf{x}|\mathbf{w}, b) = \mathbf{w}^\top \mathbf{x} + b$
- $\Pi_{+1}, \Pi_{-1} \parallel \Pi_f$ t. c.
  $\Pi_{+1} : f(\mathbf{x}) = +1$
  $\Pi_{-1} : f(\mathbf{x}) = -1$
- $d_\perp(\Pi_{+1}, \Pi_{-1}) = \frac{2}{\|\mathbf{w}\|}$



- $(\mathbf{x}_i, z_i) \quad \mathbf{x}_i \in \mathbb{R}^d, z_i \in \{\pm 1\} \quad (i = 1 \dots n)$
- vincoli di separabilità: $\forall(\mathbf{x}_i, y_i)$ deve valere

$$z_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \equiv \begin{cases} f(\mathbf{x}) \geq +1 & (z_i = +1) \\ f(\mathbf{x}) \leq -1 & (z_i = -1) \end{cases}$$

- iperpiano $\Pi_f = \{\mathbf{x} : f(\mathbf{x}) = 0\}$ con il massimo *margine* di separazione tra le due classi

- $f(\mathbf{x}|\mathbf{w}, b) = \mathbf{w}^\top \mathbf{x} + b$
- $\Pi_{+1}, \Pi_{-1} \parallel \Pi_f$ t. c.
  $\Pi_{+1} : f(\mathbf{x}) = +1$
  $\Pi_{-1} : f(\mathbf{x}) = -1$
- $d_\perp(\Pi_{+1}, \Pi_{-1}) = \frac{2}{\|\mathbf{w}\|}$

Problema duale

$$\arg\max_{\boldsymbol{\lambda}} \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2} \boldsymbol{\lambda}^\top \mathrm{diag}(\mathbf{z})\, \mathtt{G}\, \mathrm{diag}(\mathbf{z})\, \boldsymbol{\lambda}$$

$$\text{t. c. } \lambda_i \geq 0, \quad \boldsymbol{\lambda}^\top \mathbf{z} = 0$$

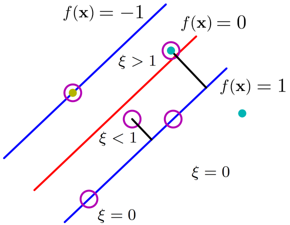
$$\hat{\mathbf{w}} = \sum_i \hat{\lambda}_i z_i \mathbf{x}_i \qquad \hat{b} = \frac{\sum_i \hat{\lambda}_i (z_i - \hat{\mathbf{w}}^\top \mathbf{x}_i)}{\sum_i \hat{\lambda}_i}$$

- λ_i: moltiplicatori di Lagrange
- $\boldsymbol{\lambda} = [\lambda_1 \dots \lambda_n]^\top \; \mathbf{z} = [z_1 \dots z_n]^\top$
- $\mathtt{G}$: matrice di Gram $G_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$

Rilassamento del vincolo di separabilità:

variabili di slack ξ_i

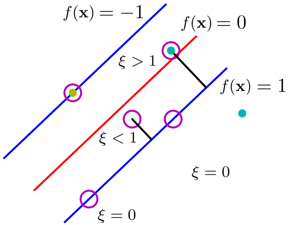
- $\xi_i = 0$ per le classificazioni "sicure"
- altrimenti, $\xi_i = |z_i - f(\mathbf{x}_i)|$ per cui
  - se $f(\mathbf{x}_i) = 0$ allora $\xi_i = 1$
  - se $\xi_i > 1$, il punto è nel lato sbagliato
  - se $0 < \xi_i < 1$, il punto è nel lato corretto ma all'interno del margine



Rilassamento del vincolo di separabilità:

variabili di slack ξ_i

- $\xi_i = 0$ per le classificazioni "sicure"
- altrimenti, $\xi_i = |z_i - f(\mathbf{x}_i)|$ per cui
  - se $f(\mathbf{x}_i) = 0$ allora $\xi_i = 1$
  - se $\xi_i > 1$, il punto è nel lato sbagliato
  - se $0 < \xi_i < 1$, il punto è nel lato corretto ma all'interno del margine



Problema primale (lin. sep)

$$\hat{\Pi}_f = \arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{t. c. } z_i f(\mathbf{x}_i) \geq 1$$

Problema duale (lin. sep)

$$\arg\max_{\boldsymbol{\lambda}} \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2}\boldsymbol{\lambda}^\top \mathrm{diag}(\mathbf{z})\, \mathbb{G}\, \mathrm{diag}(\mathbf{z})\, \boldsymbol{\lambda}$$

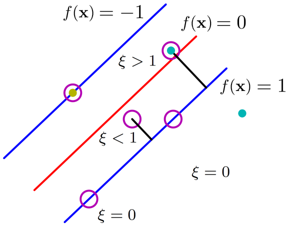
$$\text{t. c. } 0 \leq \lambda_i \qquad \forall\, i$$

$$\boldsymbol{\lambda}^\top \mathbf{z} = 0$$

Rilassamento del vincolo di separabilità:

variabili di slack ξ_i

- $\xi_i = 0$ per le classificazioni "sicure"
- altrimenti, $\xi_i = |z_i - f(\mathbf{x}_i)|$ per cui
  - se $f(\mathbf{x}_i) = 0$ allora $\xi_i = 1$
  - se $\xi_i > 1$, il punto è nel lato sbagliato
  - se $0 < \xi_i < 1$, il punto è nel lato corretto ma all'interno del margine



Problema primale

$$\hat{\Pi}_f = \arg\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_i \xi_i$$

$$\text{t. c. } z_i f(\mathbf{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

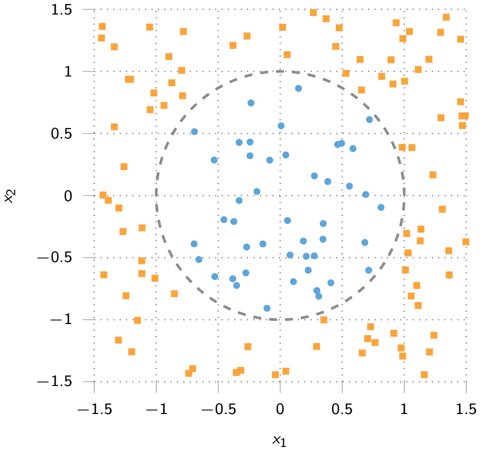
γ : parametro di regolarizzazione

Problema duale

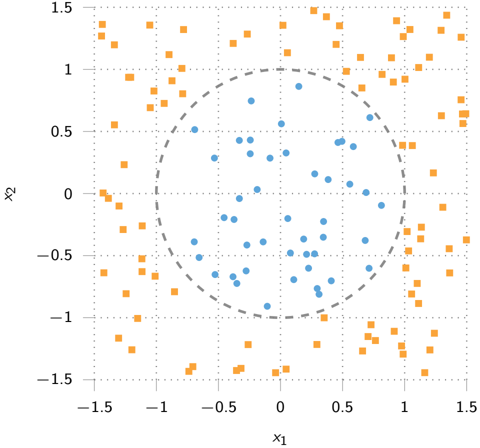
$$\arg\max_{\boldsymbol{\lambda}} \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2} \boldsymbol{\lambda}^\top \mathrm{diag}(\mathbf{z})\, \mathbb{G}\, \mathrm{diag}(\mathbf{z})\, \boldsymbol{\lambda}$$

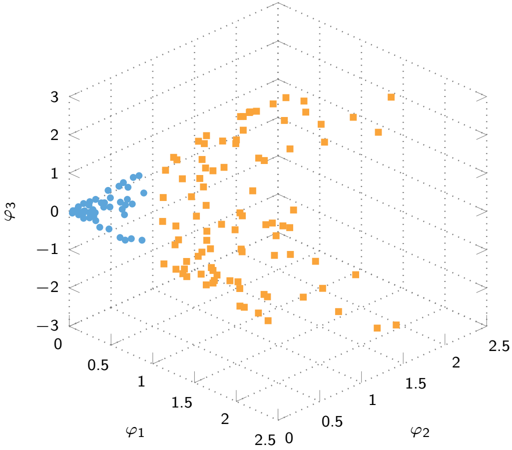
$$\text{t. c. } 0 \leq \lambda_i \leq \gamma \quad \forall i$$

$$\boldsymbol{\lambda}^\top \mathbf{z} = 0$$



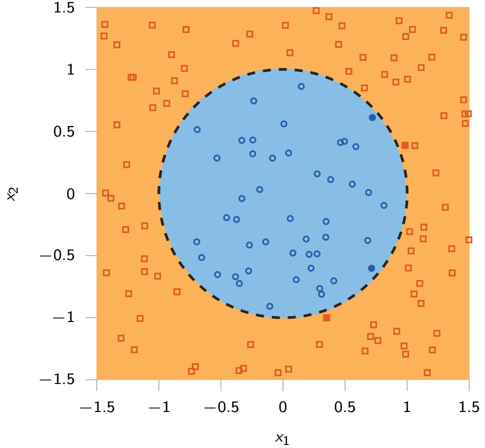
bordo di separazione: $x_1^2 + x_2^2 = 1$



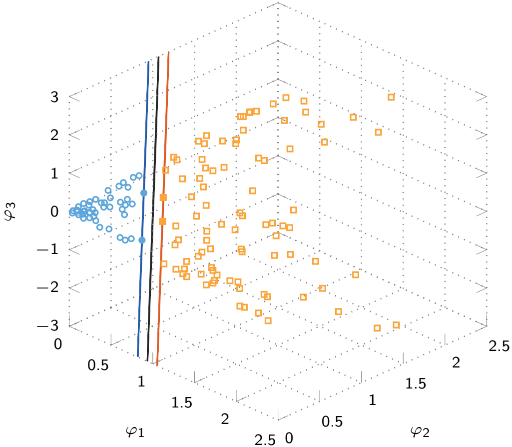


bordo di separazione: $x_1^2 + x_2^2 = 1$

$$\phi(\mathbf{x}) = [x_1^2 \; x_2^2 \; \sqrt{2}x_1x_2]^\top$$



classificazione SVM ($\gamma = 1000$)



$\hat{\Pi}_f \subset \mathbb{R}^3$

Problema primale

$$\hat{\Pi}_f = \arg\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_i \xi_i$$

$$\text{t. c. } z_i f(\phi(\mathbf{x}_i)) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

γ : parametro di regolarizzazione

Problema duale

$$\arg\max_{\boldsymbol{\lambda}} \mathbf{1}^\top \boldsymbol{\lambda} - \frac{1}{2}\boldsymbol{\lambda}^\top \mathrm{diag}(\mathbf{z})\, \mathbb{K}\, \mathrm{diag}(\mathbf{z})\, \boldsymbol{\lambda}$$

$$\text{t. c. } 0 \leq \lambda_i \leq \gamma, \qquad \boldsymbol{\lambda}^\top \mathbf{z} = 0$$

$$\mathbb{K} : \quad K_{ij} = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Non è necessario esplicitare $\phi(\mathbf{x})$: basta un kernel $\kappa(\mathbf{u},\mathbf{v})$ in $K_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j)$ per cui $\mathbb{K}$ sia *definita positiva*

$$f(\mathbf{x}) = \sum_i (\hat{\lambda}_i z_i \kappa(\mathbf{x}_i, \mathbf{x})) + \hat{b}$$

$$\hat{b} = \frac{\sum_i \lambda_i(\gamma - \lambda_i) z_i - \hat{\lambda}_i \sum_j z_j \kappa(\mathbf{x}_j, \mathbf{x}_i)}{\sum_i \lambda_i(\gamma - \lambda_i)}$$

- $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$ lineare
- $\kappa(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j)^d$ potenza
- $\kappa(\mathbf{x}_i,\mathbf{x}_j) = (\alpha \mathbf{x}_i^\top \mathbf{x}_j + \beta)^d$ polinomiale
- $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\sigma^2})$ RBF gauss.

Kernel per ambiti medicali

Dataset medicali: eterogeineità a causa di variabili continue, ordinali e categoriche

$$\kappa(\mathbf{u}, \mathbf{v}) = \frac{1}{d} \sum_{p=1}^{d} \kappa_p(u_p, v_p)$$

- u_p, v_p categoriche o binarie

$$\kappa_p(u_p, v_p) = \mathrm{I}(u_p = v_p) = \begin{cases} 1 & (u_p = v_p) \\ 0 & \text{altrimenti} \end{cases}$$

- u_p, v_p continue o ordinali

$$\kappa_p(u_p, v_p) = \frac{\Delta_p - |u_p - v_p|}{\Delta_p}$$

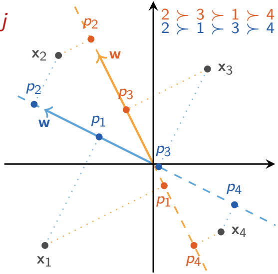
Δ_p: escursione massima per la p^a variabile

Ranking Support Vector Machines

training set $\mathcal{D} = \{(\mathbf{x}_k, z_k)\}$ *ordinato*

- z_k è il ranking di $\mathbf{x}_k$: $z_i < z_j \equiv \mathbf{x}_i$ è preferibile a $\mathbf{x}_j$ $(\mathbf{x}_i \succ \mathbf{x}_j)$
- $\mathcal{P} = \{(i,j) : \mathbf{x}_i \succ \mathbf{x}_j\}$ *insieme delle preferenze* delle coppie confrontabili
- $f(\mathbf{x})$ *funzione di ranking*: restituisce un punteggio per ciascuna istanza $\mathbf{x}$ t. c. $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ per qualsiasi $(i,j) \in \mathcal{P}$
- $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$: si cerca $\mathbf{w}$ lungo cui *la proiezione di* $\mathbf{x}_i$ *è maggiore di quella per* $\mathbf{x}_j$

$$\mathbf{w}^\top \mathbf{x}_i > \mathbf{w}^\top \mathbf{x}_j \iff \mathbf{w}^\top \underbrace{(\mathbf{x}_i - \mathbf{x}_j)}_{\Delta \mathbf{x}_{ij}} > 0 \quad \text{se } \mathbf{x}_i \succ \mathbf{x}_j$$

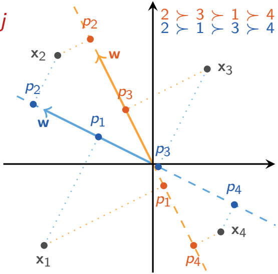


Ranking Support Vector Machines

training set $\mathcal{D} = \{(\mathbf{x}_k, z_k)\}$ *ordinato*

- z_k è il ranking di $\mathbf{x}_k$: $z_i < z_j \equiv \mathbf{x}_i$ è preferibile a $\mathbf{x}_j$ $(\mathbf{x}_i \succ \mathbf{x}_j)$
- $\mathcal{P} = \{(i,j) : \mathbf{x}_i \succ \mathbf{x}_j\}$ *insieme delle preferenze* delle coppie confrontabili
- $f(\mathbf{x})$ *funzione di ranking*: restituisce un punteggio per ciascuna istanza $\mathbf{x}$ t. c. $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ per qualsiasi $(i,j) \in \mathcal{P}$
- $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$: si cerca $\mathbf{w}$ lungo cui *la proiezione di* $\mathbf{x}_i$ *è maggiore di quella per* $\mathbf{x}_j$

$$\mathbf{w}^\top \mathbf{x}_i > \mathbf{w}^\top \mathbf{x}_j \iff \mathbf{w}^\top \underbrace{(\mathbf{x}_i - \mathbf{x}_j)}_{\Delta \mathbf{x}_{ij}} > 0 \quad \text{se } \mathbf{x}_i \succ \mathbf{x}_j$$



Problema Primale:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{(i,j)\in\mathcal{P}} \xi_{ij}$$

$$\text{t. c. } \mathbf{w}^\top \Delta \mathbf{x}_{ij} \geq 1 - \xi_{ij} \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{P}$$

Ranking Support Vector Machines

training set $\mathcal{D} = \{(\mathbf{x}_k, z_k)\}$ *ordinato*

- z_k è il ranking di $\mathbf{x}_k$: $z_i < z_j \equiv \mathbf{x}_i$ è preferibile a $\mathbf{x}_j$ $(\mathbf{x}_i \succ \mathbf{x}_j)$
- $\mathcal{P} = \{(i,j) : \mathbf{x}_i \succ \mathbf{x}_j\}$ *insieme delle preferenze* delle coppie confrontabili
- $f(\mathbf{x})$ *funzione di ranking*: restituisce un punteggio per ciascuna istanza $\mathbf{x}$ t. c. $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ per qualsiasi $(i,j) \in \mathcal{P}$
- $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$: si cerca $\mathbf{w}$ lungo cui *la proiezione di* $\mathbf{x}_i$ *è maggiore di quella per* $\mathbf{x}_j$
  $$\mathbf{w}^\top \mathbf{x}_i > \mathbf{w}^\top \mathbf{x}_j \iff \mathbf{w}^\top \underbrace{(\mathbf{x}_i - \mathbf{x}_j)}_{\Delta \mathbf{x}_{ij}} > 0 \quad \text{se } \mathbf{x}_i \succ \mathbf{x}_j$$

Problema Primale:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}, \xi_{ij}} \frac{1}{2} \|\mathbf{w}\|^2 + \gamma \sum_{(i,j) \in \mathcal{P}} \xi_{ij}$$

$$\text{t. c. } \mathbf{w}^\top \Delta \mathbf{x}_{ij} \geq 1 - \xi_{ij} \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{P}$$

Spiegazione (intuitiva):

$\mathbf{w}$ definito a meno di un fattore di scala
$\Rightarrow$ posso scalare $\mathbf{w}$ in modo che

$$\mathbf{w}^\top (\mathbf{x}_i - \mathbf{x}_j) \geq 1 \quad \text{se } \mathbf{x}_i \succ \mathbf{x}_j$$

Una coppia (i, j) con dati $(\mathbf{x}_i, y_i, \delta_i)$ e $(\mathbf{x}_j, yj, \delta_j)$ è confrontabile solo se

- si osserva l'evento sia per i che per j $(\delta_i, \delta_j = 1)$
- la coppia (i, j) mostra un evento ed una censura, con l'evento che si manifesta prima della censura (es. $T_j < C_i$)

Un confronto tra le istanze i e j è valido solo se l'istanza con il tempo minore non è censurata; altrimenti la coppia non è confrontabile

$$\mathcal{P} = \{(i, j) : (y_i > y_j) \wedge (\delta_j = 1), \ \ i, j = 1 \dots N\}$$

Una coppia (i, j) con dati $(\mathbf{x}_i, y_i, \delta_i)$ e $(\mathbf{x}_j, yj, \delta_j)$ è confrontabile solo se

- si osserva l'evento sia per i che per j $(\delta_i, \delta_j = 1)$
- la coppia (i, j) mostra un evento ed una censura, con l'evento che si manifesta prima della censura (es. $T_j < C_i$)

Un confronto tra le istanze i e j è valido solo se l'istanza con il tempo minore non è censurata; altrimenti la coppia non è confrontabile

$$\mathcal{P} = \{(i, j) : (y_i > y_j) \wedge (\delta_j = 1),\ \ i, j = 1 \ldots N\}$$

C-index (o di Harrell)

Indice di concordanza: *rapporto tra il numero di coppie concordanti e il numero di coppie comparabili*

$$c = \frac{1}{|\mathcal{P}|} \sum_{i:\delta_i=0} \sum_{j=T_i<T_j} \mathrm{I}(\hat{T}_i < \hat{T}_j)$$

Veteran's Administration Lung Cancer Trial

- x_1: Age in years
- x_2: Celltype squamous, smallcell, adeno, large
- x_3: Karnofsky score
- x_4: Months from Diagnosis
- x_5: Prior therapy: no/yes
- x_6: Treatment: standard, test
- y: Survival in days
- δ: dead (1)/censored (0)
- Dataset size 137; Censored 9
- Treated 68; Not Treated 69
- Censored & Treated 4

Veteran's Administration Lung Cancer Trial

- x_1: Age in years
- x_2: Celltype squamous, smallcell, adeno, large
- x_3: Karnofsky score
- x_4: Months from Diagnosis
- x_5: Prior therapy: no/yes
- x_6: Treatment: standard, test
- y: Survival in days
- δ: dead (1)/censored (0)
- Dataset size 137; Censored 9
- Treated 68; Not Treated 69
- Censored & Treated 4

Traning e test set

Training set: 91 individui (67%)

- Censored 5
- Treated 46; Not Treated 45

Test set: 46 individui (33%)

- Censored 4
- Treated 22; Not Treated 24

Veteran's Administration Lung Cancer Trial

- x_1: Age in years
- x_2: Celltype squamous, smallcell, adeno, large
- x_3: Karnofsky score
- x_4: Months from Diagnosis
- x_5: Prior therapy: no/yes
- x_6: Treatment: standard, test
- y: Survival in days
- δ: dead (1)/censored (0)
- Dataset size 137; Censored 9
- Treated 68; Not Treated 69
- Censored & Treated 4

Traning e test set

Training set: 91 individui (67%)

- Censored 5
- Treated 46; Not Treated 45

Test set: 46 individui (33%)

- Censored 4
- Treated 22; Not Treated 24

Modelli a confronto

- modello di Cox
- Ranking SVM (lineare)
- Ranking SVM (RBF)

http://localhost:8889/lab

- A. Daemen & B. De Moor. Development of a kernel function for clinical data. In *2009 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2009.
- S. Pölsterl, N. Navab, & A. Katouzian. Fast training of support vector machines for survival analysis. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, 2015.
- S. Pölsterl, N. Navab, & A. Katouzian. An Efficient Training Algorithm for Kernel Survival Support Vector Machines. In *3rd Workshop on Machine Learning in Life Sciences*, 2016.
- D. Sontag. Lecture 6: Physiological time-series. In *6.S897/HST.956 Machine Learning for Healthcare*. MIT, Cambridge MA, 2019.
- V. Van Belle, K. Pelckmans, S. Van Huffel, & J. A. K. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.
- P. Wang, Y. Li, & C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6), 2019.

- C. M. Bishop. Pattern Recognition and Machine Learning. *Springer*, 2007 (1st edition).
- O. Chapelle & S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13:201–215, 06 2010.
- R. Herbrich, T. Graepel, & K. Obermayer. Large margin rank boundaries for ordinal regression. In *P. J. Bartlett, B. Schölkopf, D. Schuurmans, & A. J. Smola, editors, Advances in Large Margin Classifiers*, MIT Press, 2000.
- T. Joachims. Optimizing search engines using clickthrough data. In *Proc. of the Eighth ACM Int. Conf. on Knowledge Discovery and Data Mining*, NY USA, 2002. ACM.
- T.-M. Kuo, C.-P. Lee, & C.-J. Lin. Large-scale kernel rankSVM. In *Proc. of SIAM International Conference on Data Mining*, 2014.
- C.-P. Lee & C.-J. Lin. Large-scale linear rankSVM. *Neural Computation*, 26(4):781–817, 2014.
- W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes: The Art of Scientific Computing. *Cambridge University Press*, 2007 (3rd edition).
- S. J. Russell & P. Norvig. Artificial Intelligence: A Modern Approach. *Prentice Hall*, 2002 (2nd edition).
- Hwanjo Yu & Sungchul Kim. Svm tutorial – classification, regression and ranking. In *G. Rozenberg, T. Bäck, and J. N. Kok, editors, Handbook of Natural Computing*. Springer Berlin Heidelberg, 2012.

Ranking Support Vector Machines

Formulazione efficiente

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{(i,j)\in\mathcal{P}} \xi_{ij}$$

$$\text{t. c. } \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij} \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{P}$$

Ranking Support Vector Machines

Formulazione efficiente

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{(i,j)\in\mathcal{P}} \xi_{ij}$$

$$\text{t. c. } \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij} \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq \overbrace{\max(0, 1 - \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j))}^{\text{ReLU}(\mathbf{w}^\top\Delta\mathbf{x}_{ij})} \geq 0$$

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\sum_{(i,j)\in\mathcal{P}} \text{ReLU}(\mathbf{w}^\top\Delta\mathbf{x}_{ij})^2$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{(i,j)\in\mathcal{P}} \xi_{ij}$$

$$\text{t. c. } \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij} \quad \forall(i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq 0 \quad \forall(i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq \overbrace{\max(0, 1 - \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j))}^{\mathrm{ReLU}(\mathbf{w}^\top\Delta\mathbf{x}_{ij})} \geq 0$$

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\sum_{(i,j)\in\mathcal{P}} \mathrm{ReLU}(\mathbf{w}^\top\Delta\mathbf{x}_{ij})^2$$

$$\frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{\gamma}{2}(\mathbf{1} - \mathtt{A}\mathtt{X}\mathbf{w})^\top\mathtt{D}_\mathbf{w}(\mathbf{1} - \mathtt{A}\mathtt{X}\mathbf{w})$$

- $\mathtt{X} = [\mathbf{x}_1 \ldots \mathbf{x}_N]^\top$; $\mathbf{1} = [1 \ldots 1]^\top \in \mathbb{R}^p$ $(p = |\mathcal{P}|)$
- $\mathtt{A} \in \mathbb{R}^{p\times N}$ matrice *sparsa* : la k^a riga corrisponde alla k^a coppia $(i,j)\in\mathcal{P}$, con $A_{ki} = 1$ e $A_{kj} = -1$ (0 altrimenti)
- $\mathtt{D}_\mathbf{w}$ matrice diagonale $p \times p$ t. c. in corrispondenza della k^a coppia $(i,j)\in\mathcal{P}$, $(\mathtt{D}_\mathbf{w})_{kk} = \mathrm{I}(\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) < 1)$

Ranking Support Vector Machines

Formulazione efficiente

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{(i,j)\in\mathcal{P}} \xi_{ij}$$

$$\text{t. c. } \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij} \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{P}$$

$$\xi_{ij} \geq \overbrace{\max(0, 1 - \mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j))}^{\mathrm{ReLU}(\mathbf{w}^\top \Delta\mathbf{x}_{ij})} \geq 0$$

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\sum_{(i,j)\in\mathcal{P}} \mathrm{ReLU}(\mathbf{w}^\top \Delta\mathbf{x}_{ij})^2$$

$$\frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{\gamma}{2}(\mathbf{1} - \mathtt{A}\mathtt{X}\mathbf{w})^\top \mathtt{D}_\mathbf{w}(\mathbf{1} - \mathtt{A}\mathtt{X}\mathbf{w})$$

- $\mathtt{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^\top$; $\mathbf{1} = [1 \dots 1]^\top \in \mathbb{R}^p$ $(p = |\mathcal{P}|)$
- $\mathtt{A} \in \mathbb{R}^{p\times N}$ matrice *sparsa* : la k^a riga corrisponde alla k^a coppia $(i,j)\in\mathcal{P}$, con $A_{ki} = 1$ e $A_{kj} = -1$ (0 altrimenti)
- $\mathtt{D}_\mathbf{w}$ matrice diagonale $p \times p$ t. c. in corrispondenza della k^a coppia $(i,j)\in\mathcal{P}$, $(\mathtt{D}_\mathbf{w})_{kk} = \mathrm{I}(\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) < 1)$

Con kernel:

$$\frac{1}{2}\boldsymbol{\lambda}^\top \mathtt{K}\boldsymbol{\lambda} + \frac{\gamma}{2}(\mathbf{1} - \mathtt{A}\mathtt{K}\boldsymbol{\lambda})^\top \mathtt{D}_\lambda(\mathbf{1} - \mathtt{A}\mathtt{K}\boldsymbol{\lambda})$$