Support Vector Machine

Indice

Motivazione gestionale

Hard margin: la geometria

Soft margin

La formulazione duale e i support vector

Kernel: frontiere non lineari da un QP convesso

Classi sbilanciate e costi asimmetrici

Support Vector Regression

Esercizi

Obiettivi (1/2)

Alla fine di questo capitolo sarai in grado di:

- formulare la classificazione a margine rigido e morbido come QP;
- ricavare e risolvere il duale, riconoscendo i vettori di supporto;

Obiettivi (2/2)

Alla fine di questo capitolo sarai in grado di:

- ottenere frontiere non lineari con i kernel restando nel mondo convesso;
- gestire classi sbilanciate con costi asimmetrici e valutare con le metriche giuste.

Prerequisiti

- Da questo modulo, Capitolo 2 «Richiami: convessità, programmazione quadratica e condizioni KKT»: i programmi quadratici convessi, le condizioni KKT e i moltiplicatori come prezzi ombra
- Dal modulo *Ottimizzazione convessa*, Capitolo 6 «Problemi di ottimizzazione convessi e concavi»: i problemi di ottimizzazione convessi e le loro condizioni di ottimalità
- Da questo modulo, Capitolo 3 «Il solver per i modelli non lineari»: gli obiettivi quadratici in `gurobipy`

Motivazione gestionale

Introduzione: Motivazione gestionale

- La Support Vector Machine collega direttamente l'ottimizzazione convessa al machine learning
- Il classificatore si ottiene risolvendo un **QP**, con una lettura geometrica chiara
- Applicazioni gestionali immediate: rischio di credito, churn, manutenzione predittiva, controllo qualità, frodi

Nessuna libreria di machine learning

- In questo capitolo *non* usiamo librerie di ML: ogni modello è un QP scritto e risolto con Gurobi
- L'obiettivo è capire *che cosa* ottimizza un classificatore
- **Nota modellistica**: le etichette $y_i \in \{-1, +1\}$ sono *dati osservati*, non variabili
- Le variabili (coefficienti, intercetta, scarti, moltiplicatori) sono tutte continue

Il capitolo chiude il cerchio

- Gli stessi strumenti usati per produzione, logistica e finanza — QP convessi, dualità, moltiplicatori — sono il motore di uno dei classificatori più eleganti del machine learning
- Capire la SVM come problema di ottimizzazione, e non come scatola nera di libreria, dà il controllo completo su ciò che il modello sta davvero facendo

Riepilogo

- La SVM è un classificatore ottenuto risolvendo un QP convesso
- Le etichette y_i sono dati; tutte le variabili sono continue
- Ogni modello del capitolo è scritto e risolto con Gurobi, senza librerie di ML

Hard margin: la geometria

Introduzione: Hard margin

- Partiamo dal caso più semplice: due classi separabili da un iperpiano
- Tra tutti gli iperpiani separatori cerchiamo quello con il **margine** più ampio
- Vedremo che massimizzare il margine è un QP convesso con vincoli lineari

Il problema a parole

- *Decidiamo* i coefficienti di un iperpiano che separi due classi di osservazioni
- *L'obiettivo* è massimizzare il margine, cioè la distanza tra l'iperpiano e i punti più vicini delle due classi (un classificatore con margine ampio generalizza meglio)
- *I vincoli* chiedono che ogni osservazione stia dal lato giusto dell'iperpiano — nel soft margin, a meno di uno scarto pagato in obiettivo

Dati (input del modello) (1/2)

Simbolo	Tipo	Significato
n	$\in \mathbb{Z}_{\geq 1}$	numero di osservazioni del training set, indicizzate da $i \in \{1, 2, \ldots, n\}$
p	$\in \mathbb{Z}_{\geq 1}$	numero di caratteristiche, indicizzate da $j \in \{1, 2, \ldots, p\}$
x_{ij}	$\in \mathbb{Q}$	valore della caratteristica j per l'osservazione i; $\boldsymbol{x}_i \in \mathbb{Q}^p$ è il vettore dell'osservazione i

Dati (input del modello) (2/2)

Simbolo	Tipo	Significato
y_i	$\in \{-1, +1\}$	etichetta *osservata* della classe di i (un dato, non una variabile)
C	$\in \mathbb{Q}_{>0}$	peso delle violazioni del margine (solo soft margin)

- C è maiuscolo per aderenza alla letteratura di machine learning: un'eccezione alla convenzione che riserva le maiuscole a insiemi e matrici

Variabili decisionali

Introduciamo le seguenti $p+1$ variabili libere:

$$\begin{cases} w_j = \text{coefficiente } j \text{ dell'iperpiano separatore} \\ b = \text{intercetta dell'iperpiano} \end{cases} \qquad \forall j \in \{1, 2, \ldots, p\}$$

Esse definiscono l'iperpiano

$$\sum_{j=1}^{p} w_j \, x_j + b = 0$$

Modello QP hard margin

Modello QP hard margin

$$\min \quad \frac{1}{2}\sum_{j=1}^{p} w_j^2 \tag{12.1a}$$

$$\text{soggetto a} \quad y_i\Big(\sum_{j=1}^{p} w_j\, x_{ij} + b\Big) \geq 1, \quad \forall i \in \{1,2,\ldots,n\}, \tag{12.1b}$$

$$w_j \gtreqless 0, \quad \forall j \in \{1,2,\ldots,p\}, \tag{12.1c}$$

$$b \gtreqless 0. \tag{12.1d}$$

Descrizione del modello QP hard margin

- La funzione obiettivo quadratica convessa minimizza $\frac{1}{2}\|\boldsymbol{w}\|^2$: poiché il margine geometrico tra gli iperpiani di supporto è $2/\|\boldsymbol{w}\|$, minimizzarla equivale a *massimizzare il margine*
- I vincoli lineari di classificazione impongono che ogni osservazione stia dal lato giusto dell'iperpiano, a distanza almeno unitaria in scala $\boldsymbol{w}$ (n vincoli lineari)
- Gli ultimi vincoli definiscono le variabili del modello, tutte libere

Esempio: Due punti, tutto a mano (1/3)

- $\boldsymbol{x}_1 = (0, 0)$ con $y_1 = -1$ e $\boldsymbol{x}_2 = (2, 2)$ con $y_2 = +1$

Passo 1 — vincoli

$$-(b) \geq 1 \qquad \text{e} \qquad 2w_1 + 2w_2 + b \geq 1$$

cioè

$$b \leq -1 \qquad \text{e} \qquad 2(w_1 + w_2) \geq 1 - b \geq 2$$

Esempio: Due punti, tutto a mano (2/3)

Passo 2 — simmetria e ottimo

- Per minimizzare $w_1^2 + w_2^2$ con $w_1 + w_2 \geq 1$ conviene $w_1 = w_2 = t$ e prendere le disuguaglianze attive:

$$4t + b = 1, \quad b = -1 \quad \Rightarrow \quad t = 1/2$$

- Quindi $\boldsymbol{w} = (0{,}5;\ 0{,}5)$ e $b = -1$
- L'iperpiano $0{,}5x_1 + 0{,}5x_2 = 1$ è l'asse del segmento tra i due punti

Esempio: Due punti, tutto a mano (3/3)

Passo 3 — margine

$$\frac{2}{\|\boldsymbol{w}\|} = \frac{2}{\sqrt{0{,}5}} = 2\sqrt{2} = 2{,}83$$

- È esattamente la distanza tra i due punti
- Entrambi i vincoli sono attivi: entrambi i punti sono *support vector*
- Con dati reali sovrapposti, però, l'hard margin è inammissibile: serve il soft margin

Riepilogo

- L'hard margin cerca l'iperpiano separatore di margine massimo
- Margine $= 2/\|\boldsymbol{w}\|$: massimizzarlo equivale a minimizzare $\frac{1}{2}\|\boldsymbol{w}\|^2$
- È un QP convesso con n vincoli lineari e $p + 1$ variabili libere
- I punti con vincolo attivo sono i *support vector*
- Con classi sovrapposte il modello è inammissibile

Soft margin

Introduzione: Soft margin

- Con dati reali sovrapposti l'hard margin è inammissibile
- Permettiamo ad alcune osservazioni di violare il margine, pagando la violazione in obiettivo
- Il parametro C governa il compromesso tra margine ampio e aderenza ai dati

Variabili di scarto

Si aggiungono n variabili non negative di scarto:

$\xi_i =$ violazione del margine dell'osservazione i, $\quad \forall i \in \{1, 2, \ldots, n\}$

- $\xi_i = 0$: l'osservazione rispetta il margine
- $\xi_i > 1$: l'osservazione è classificata in modo errato

Modello QP soft margin (primale)

Modello QP soft margin (primale)

$$\min \quad \frac{1}{2}\sum_{j=1}^{p} w_j^2 + C\sum_{i=1}^{n} \xi_i \tag{12.2a}$$

$$\text{soggetto a} \quad y_i\Big(\sum_{j=1}^{p} w_j\, x_{ij} + b\Big) \geq 1 - \xi_i, \quad \forall i \in \{1, 2, \ldots, n\}, \tag{12.2b}$$

$$w_j \gtreqless 0, \qquad \forall j \in \{1, 2, \ldots, p\}, \tag{12.2c}$$

$$b \gtreqless 0, \tag{12.2d}$$

$$\xi_i \geq 0, \qquad \forall i \in \{1, 2, \ldots, n\}. \tag{12.2e}$$

Descrizione del modello QP soft margin

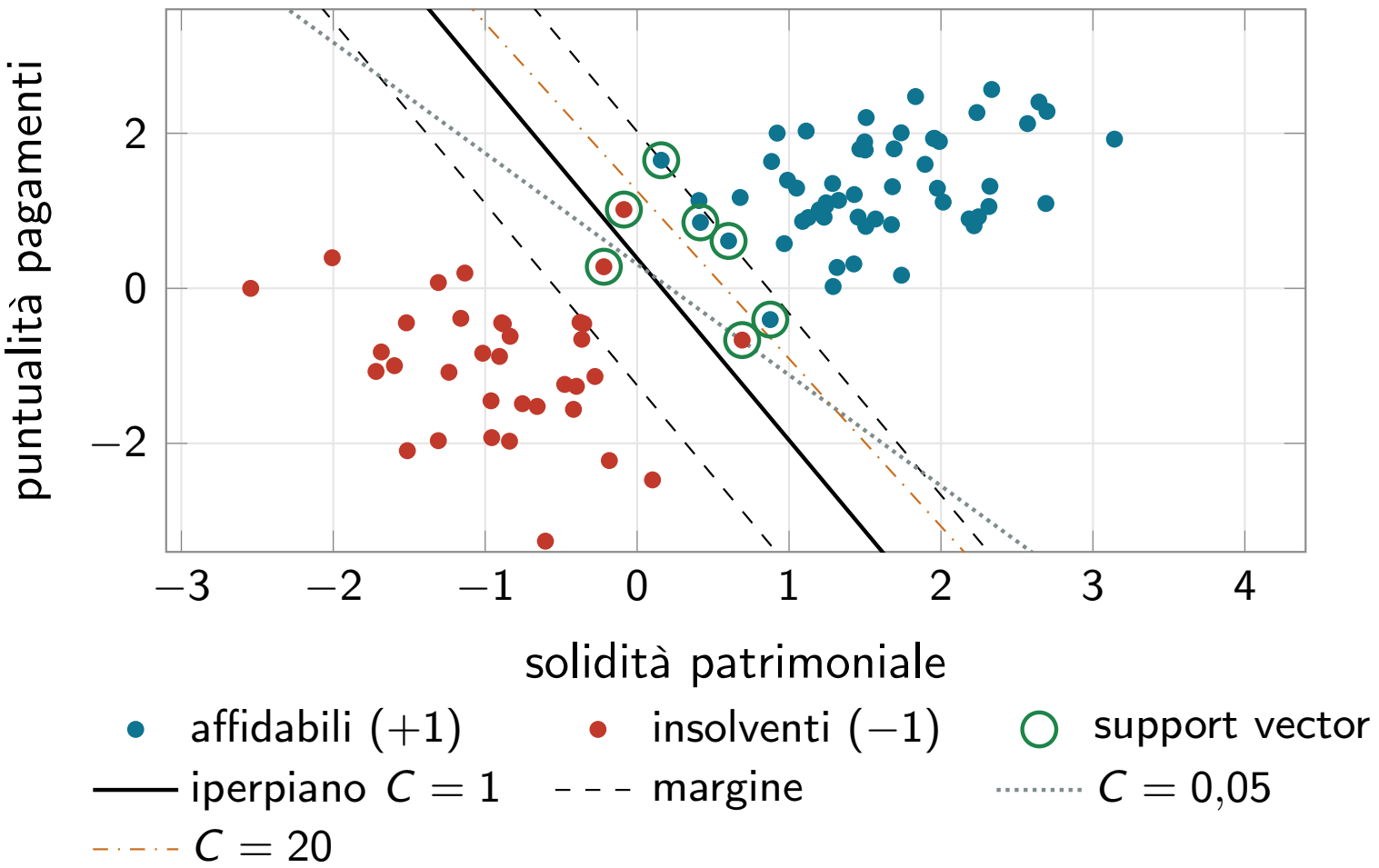
- La funzione obiettivo quadratica convessa bilancia due termini:
  - la semplicità del modello ($\frac{1}{2}\|\boldsymbol{w}\|^2$, margine ampio)
  - il costo delle violazioni sul training set ($C \sum_i \xi_i$)
- C è il prezzo assegnato alle violazioni
- I vincoli lineari richiedono la classificazione corretta *a meno dello scarto* ξ_i, pagato in obiettivo (n vincoli lineari)
- Variabili: w_j e b libere, ξ_i non negative

Caso di studio: rischio di credito

- 90 clienti, due indicatori standardizzati: solidità patrimoniale e puntualità nei pagamenti
- 55 affidabili e 35 insolventi, parzialmente sovrapposti

C	$\boldsymbol{w}$	b	margine	errori	nel margine
0,05	(0,678; 0,474)	−0,148	2,418	2	22
1,00	(1,433; 0,611)	−0,236	1,284	2	7
20,00	(5,058; 2,338)	−2,936	0,359	0	3

Soft margin: iperpiano, margine e support vector ($C = 1$)



Come leggere il grafico

- I 90 clienti nel piano dei due indicatori: affidabili (teal), insolventi (rosso), support vector cerchiati in verde
- La retta nera è l'iperpiano ottimo per $C = 1$ con le sue rette di margine (tratteggiate)
- Le rette grigia e arancione sono gli iperpiani per $C = 0{,}05$ (margine ampio) e $C = 20$ (margine stretto)

Il compromesso regolarizzazione–aderenza

- C piccolo compra un margine ampio accettando 22 punti dentro il margine
- C grande insegue ogni punto del training set (zero errori, margine strettissimo)
- È il compromesso **regolarizzazione–aderenza**: C va scelto con dati di validazione, mai guardando il test set
- Le tre rette sono simili ma non uguali: la pendenza cambia perché i punti "pagati" cambiano

Riepilogo

- Il soft margin aggiunge gli scarti $\xi_i \geq 0$, pagati C ciascuno in obiettivo
- Resta un QP convesso, sempre ammissibile
- C piccolo: margine ampio, più violazioni; C grande: margine stretto, aderenza al training set
- C si sceglie con dati di validazione, mai con il test set

La formulazione duale e i support vector

Introduzione: La formulazione duale

- Associamo un moltiplicatore a ogni vincolo di classificazione e scriviamo il duale del soft margin
- Il duale rivela quali osservazioni determinano davvero la frontiera: i **support vector**
- Il duale dipende dai dati solo tramite prodotti scalari: è la porta verso i kernel

Moltiplicatori

Associando un moltiplicatore α_i a ciascun vincolo di classificazione del soft margin, si ottengono n nuove variabili:

$\alpha_i =$ moltiplicatore del vincolo di classificazione dell'osservazione i

$$\forall i \in \{1, 2, \ldots, n\}$$

Modello QP duale del soft margin

Modello QP duale del soft margin

$$\max \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, \boldsymbol{x}_i' \boldsymbol{x}_j \tag{12.3a}$$

$$\text{soggetto a} \quad \sum_{i=1}^{n} \alpha_i \, y_i = 0, \tag{12.3b}$$

$$\alpha_i \leq C, \quad \forall i \in \{1, 2, \ldots, n\}, \tag{12.3c}$$

$$\alpha_i \geq 0, \quad \forall i \in \{1, 2, \ldots, n\}. \tag{12.3d}$$

Descrizione del modello duale

- La funzione obiettivo quadratica concava dipende dai dati *solo* tramite i prodotti scalari $\boldsymbol{x}_i'\boldsymbol{x}_j$: è l'osservazione che apre la porta al kernel
- Il vincolo lineare $\sum_i \alpha_i y_i = 0$ deriva dalla stazionarietà rispetto all'intercetta b (un vincolo lineare)
- I vincoli $\alpha_i \leq C$ limitano ogni moltiplicatore al prezzo C delle violazioni (n vincoli lineari)
- I vincoli $\alpha_i \geq 0$ definiscono le variabili del modello

Relazione primale-duale

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \, y_i \, \boldsymbol{x}_i \qquad\qquad f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}'\boldsymbol{x} + b)$$

Sul caso del rischio di credito ($C = 1$):

- $\boldsymbol{w}$ dal duale $= (1{,}433;\ 0{,}611)$, $b = -0{,}236$: identici al primale
- valore duale 4,7868 = valore primale (dualità forte)

moltiplicatore	punti	significato
$\alpha_i = 0$	83	NON influenzano la frontiera
$0 < \alpha_i < C$	2	esattamente sul margine
$\alpha_i = C$	5	dentro il margine o mal classificati

Perché si chiamano support vector (1/2)

- La soluzione dipende *solo* dai 7 punti con $\alpha_i > 0$: sono loro a "sostenere" l'iperpiano
- Gli altri 83 potrebbero sparire senza cambiare nulla
- Il duale, inoltre, dipende dai dati solo tramite i prodotti scalari $\boldsymbol{x}_i' \boldsymbol{x}_j$: questa osservazione apre la porta al kernel

Perché si chiamano support vector (2/2)

La complementarietà KKT organizza tutto:

- $\alpha_i = 0 \Rightarrow$ punto oltre il margine
- $0 < \alpha_i < C \Rightarrow$ esattamente sul margine (da questi si ricava b)
- $\alpha_i = C \Rightarrow$ il vincolo "paga" scarto $\xi_i > 0$

Riepilogo

- Il duale ha una variabile $\alpha_i \in [0, C]$ per ogni osservazione
- $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$: la frontiera dipende solo dai punti con $\alpha_i > 0$, i support vector
- Dualità forte: valore duale = valore primale
- I dati entrano nel duale solo tramite i prodotti scalari $\boldsymbol{x}_i' \boldsymbol{x}_j$

Kernel: frontiere non lineari da un QP convesso

Introduzione: Kernel

- Non tutte le classi si separano con una retta
- Sfruttando il duale, sostituiamo il prodotto scalare con un **kernel**
- Otteniamo frontiere curve restando in un QP convesso

Il kernel RBF

- Sostituiamo $\boldsymbol{x}_i' \boldsymbol{x}_j$ con un kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, ad esempio il kernel RBF (*Radial Basis Function*) gaussiano

$$K(\boldsymbol{x}, \boldsymbol{z}) = e^{-\gamma \|\boldsymbol{x} - \boldsymbol{z}\|^2}$$

- Si classifica in uno spazio trasformato senza mai costruirlo:

$$f(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right)$$

- La frontiera può essere arbitrariamente curva, ma il training *resta un QP convesso* se il kernel è valido (matrice semidefinita positiva)

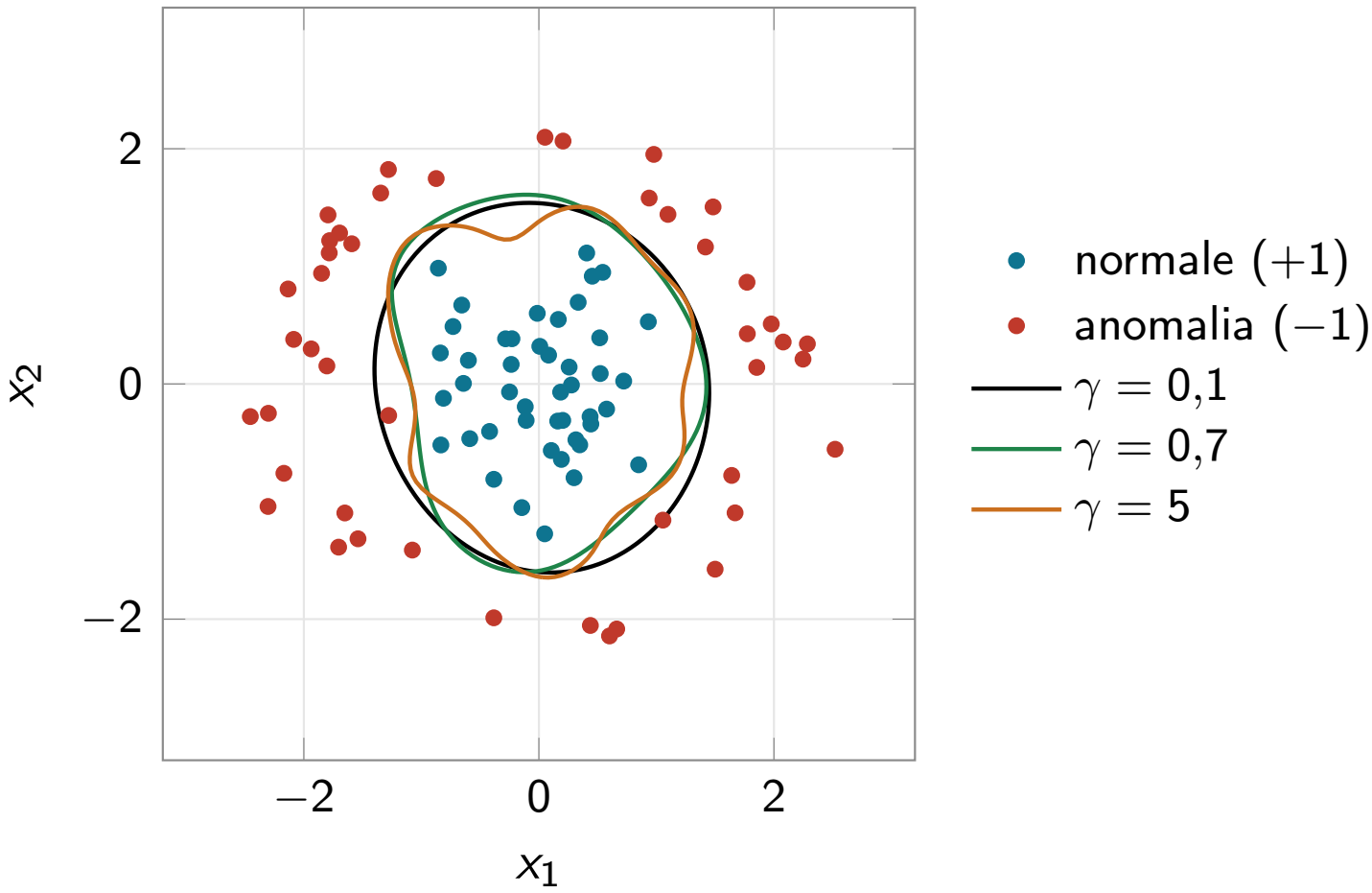
Caso di studio: anomalie "a corona"

- 90 osservazioni di processo
- Il funzionamento normale al centro, le anomalie disposte ad anello
- Nessuna retta può separarle

γ	errori di training	support vector
0,1	1	20
0,7	0	15
5,0	0	61

Con $\gamma = 5$ il modello "memorizza" i punti.

Kernel RBF: la frontiera di livello 0



Come leggere il grafico

- Il dataset "a corona": funzionamento normale al centro, anomalie ad anello
- Le curve sono le frontiere di decisione del kernel RBF per tre valori di γ
- Con $\gamma = 0{,}1$ la frontiera è un cerchio regolare
- Con $\gamma = 5$ si frammenta intorno ai singoli punti (overfitting)

Overfitting

Osservazione 1: Convessità non vuol dire generalizzazione

Con $\gamma = 5$ la frontiera si frammenta in isole intorno ai singoli punti e i support vector salgono a 61 su 90: il modello sta memorizzando il training set (overfitting), pur restando un QP convesso.

Convessità del training e capacità di generalizzare sono proprietà diverse. C e γ vanno scelti insieme per cross-validation.

Riepilogo

- Il kernel sostituisce il prodotto scalare nel duale: frontiere non lineari senza costruire lo spazio trasformato
- Con un kernel valido il training resta un QP convesso
- γ grande: frontiera frammentata, molti support vector, overfitting
- C e γ si scelgono insieme per cross-validation

Classi sbilanciate e costi asimmetrici

Introduzione: Classi sbilanciate

- Nelle applicazioni gestionali la classe importante è rara
- L'errore sui rari costa di più (un insolvente non individuato, una frode non scoperta)
- Vediamo come incorporare costi asimmetrici nel modello

Pesare gli scarti

Basta pesare gli scarti in modo diverso per le due classi:

$$C_- \sum_{i:\, y_i=-1} \xi_i \; + \; C_+ \sum_{i:\, y_i=+1} \xi_i$$

	precision	recall insolventi
costi uguali ($C = 1$)	1,00	0,94
costo 10$\times$ sugli insolventi	0,97	1,00

Costi decisionali, non frequenze

- Con il costo 10× la frontiera arretra verso la classe "sana"
- Nessun insolvente sfugge (recall 100%) al prezzo di qualche falso allarme in più (precision dal 100% al 97%)
- I pesi devono riflettere i *costi decisionali*: quanto costa un fido concesso a chi non ripagherà vs. un cliente buono respinto
- Non le frequenze delle classi

Riepilogo

- Con classi sbilanciate si usano due prezzi C_- e C_+ per gli scarti
- Il modello resta un QP convesso
- Si valuta con precision e recall, non con la sola accuratezza
- I pesi riflettono i costi decisionali, non le frequenze

Support Vector Regression

Introduzione: Support Vector Regression

- La stessa idea si applica alla stima di una funzione continua
- La SVR ignora gli errori entro un tubo di ampiezza ε
- Solo i punti fuori dal tubo determinano la stima

Il modello SVR

$$\min \ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

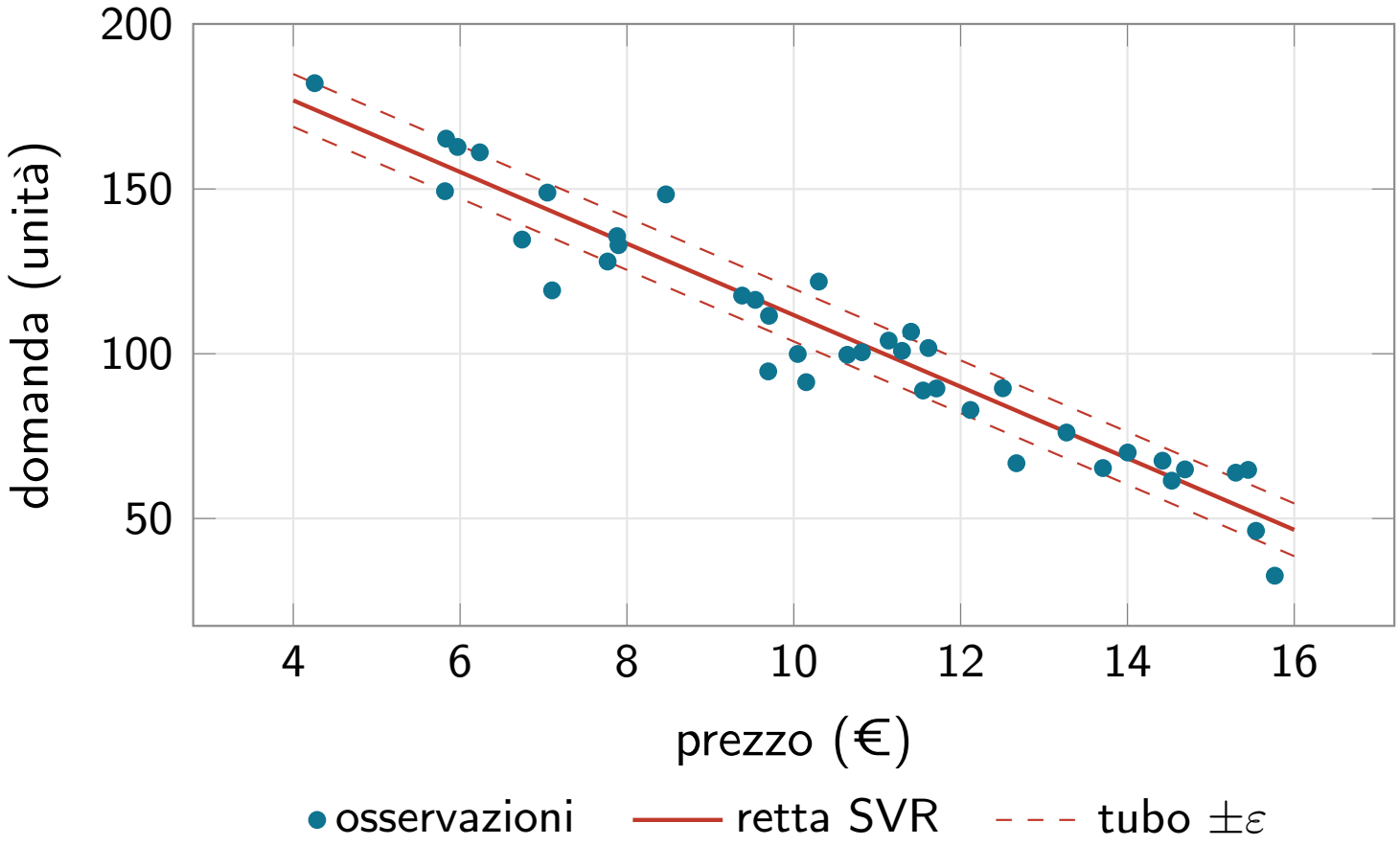
con

$$|y_i - (\boldsymbol{w}'\boldsymbol{x}_i + b)| \leq \varepsilon + \xi_i^{(*)} \qquad \forall i$$

Sulla relazione prezzo–domanda (40 osservazioni):

- retta SVR: domanda $= -10{,}86 \cdot$ prezzo $+ 220{,}32$ (vera: $-11p + 220$)
- tubo $\varepsilon = 8$: 14 punti su 40 fuori dal tubo $\rightarrow$ solo loro determinano la retta

SVR: solo i punti fuori dal tubo pagano penalità



Come leggere il grafico

- Le 40 osservazioni della relazione prezzo–domanda
- La retta stimata e il tubo di tolleranza $\pm\varepsilon = 8$
- Solo i 14 punti fuori dal tubo pagano penalità e determinano la retta

Riepilogo

- La SVR stima una funzione continua con un QP convesso
- Gli errori entro il tubo $\pm\varepsilon$ non si pagano
- Solo i punti fuori dal tubo determinano la retta: sono i support vector della regressione

Esercizi

Introduzione: Esercizi

- Gli esercizi mettono alla prova i concetti del capitolo
- Si risolvono a mano o con lo script del laboratorio e Gurobi
- Toccano hard margin, complementarietà, scelta di C, soglia decisionale e One-Class SVM

Esercizio: verifica

- Risolvere a mano l'hard margin con

$$\boldsymbol{x}_1 = (0,0),\ y_1 = -1 \qquad \text{e} \qquad \boldsymbol{x}_2 = (4,0),\ y_2 = +1$$

- Verificare con Gurobi:

$$\boldsymbol{w} = (0{,}5;\ 0), \qquad b = -1, \qquad \text{margine } 4$$

Esercizio: complementarietà

- Per $C = 1$, elencare dallo script i 7 punti con $\alpha_i > 0$
- Verificare numericamente le tre condizioni di complementarietà:

$$\alpha_i = 0 \Rightarrow \xi_i = 0 \text{ e margine } \geq 1; \quad \text{ecc.}$$

Esercizio: curva di validazione

- Dividere i 90 clienti in training (60) e validazione (30)
- Tracciare l'errore di validazione per $C \in \{0{,}01;\ 0{,}1;\ 1;\ 10;\ 100\}$ e scegliere C
- L'errore di training è monotono in C? E quello di validazione?

Esercizio: soglia decisionale

- A parità di modello ($C = 1$), spostare la soglia: classificare insolvente se

$$\boldsymbol{w}'\boldsymbol{x} + b < \theta, \qquad \theta \in [-1, 1]$$

- Tracciare precision e recall in funzione di θ
- Dove sta il punto che massimizza il profitto se un insolvente non individuato costa 10 e un falso allarme 1?

Esercizio: one-class

- Implementare la One-Class SVM sul dataset a corona usando solo i punti normali:

$$\min \ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \quad \text{soggetto a} \quad \boldsymbol{w}'\phi(\boldsymbol{x}_i) \geq \rho - \xi_i$$

- Kernel RBF, $\gamma = 0{,}7$
- Quale frazione di anomalie individua per $\nu \in \{0{,}05;\ 0{,}1;\ 0{,}2\}$?

Riepilogo finale (1/2)

Modelli:

- Hard margin: QP convesso, massimizza il margine $2/\|\boldsymbol{w}\|$
- Soft margin: scarti ξ_i pagati C, sempre ammissibile
- Duale: $\alpha_i \in [0, C]$, $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$

Support vector:

- Solo i punti con $\alpha_i > 0$ determinano la frontiera
- La complementarietà KKT dice dove sta ciascun punto

Riepilogo finale (2/2)

Estensioni:

- Kernel: frontiere non lineari da un QP convesso
- Costi asimmetrici C_-, C_+ per classi sbilanciate
- SVR: regressione con un tubo di tolleranza $\pm\varepsilon$

Scelte del modellista:

- C e γ con dati di validazione, mai con il test set
- I pesi riflettono i costi decisionali, non le frequenze