IN550 Machine Learning

Classificazione discriminativa

Vincenzo Bonifaci

Classificazione generativa vs. discriminativa

Approccio generativo

- Stima $\Pr(x, y)$ per poi dedurre $\Pr(y|x)$
- Confronta le $\Pr(y|x)$ per trovare la classe più verosimile

Esempi: QDA, LDA, Naive Bayes

Approccio discriminativo

- Stima $\Pr(y|x)$
- Confronta le $\Pr(y|x)$ per trovare la classe più verosimile

Oppure, evitando completamente le probabilità,

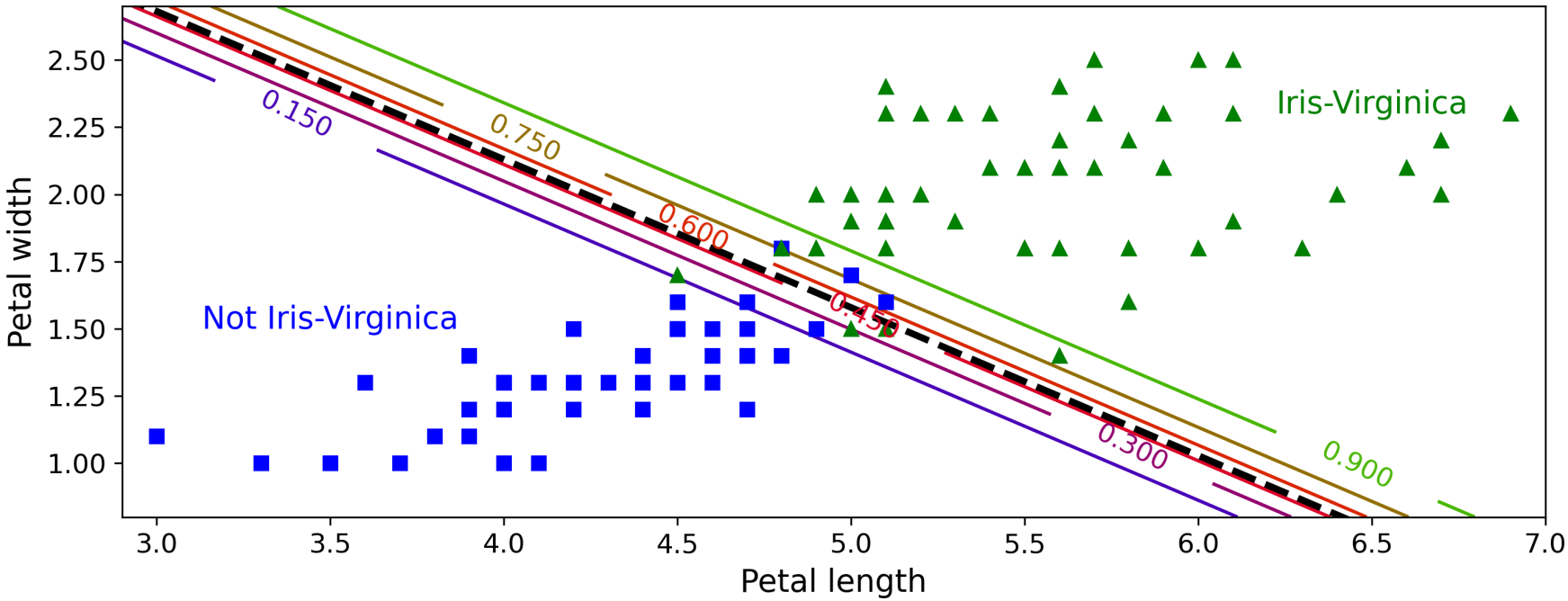
- Costruisci direttamente una funzione da $\mathcal{X}$ a $\mathcal{Y}$

Esempi: Regressione logistica, Percettrone, Support Vector Machines

Stima di probabilità per etichette binarie

Stima della probabilità condizionata (etichette binarie)

Dato: un insieme di esempi (x, y) con $x \in \mathbb{R}^{d+1}$ e $y \in \{0, 1\}$
Trova: una funzione $h : \mathcal{X} \to [0, 1]$ con $h(x) \approx \Pr(y = 1|x)$



Un modello lineare per la stima di probabilità?

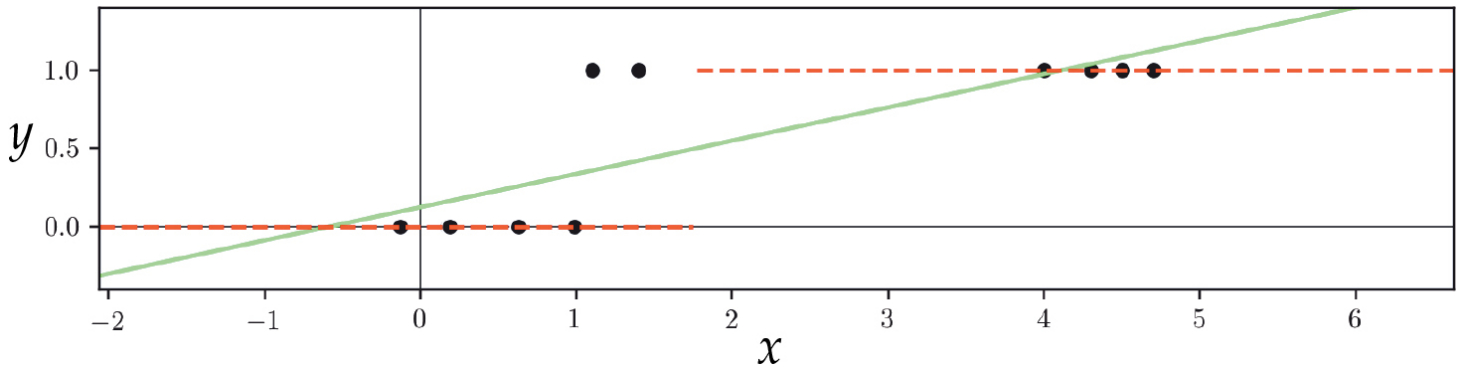
Dato x, vogliamo stimare $\Pr(y = 1|x)$ attraverso una funzione lineare

$$w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d = w^\top x$$

Vorremmo che $\Pr(y = 1|x)$:

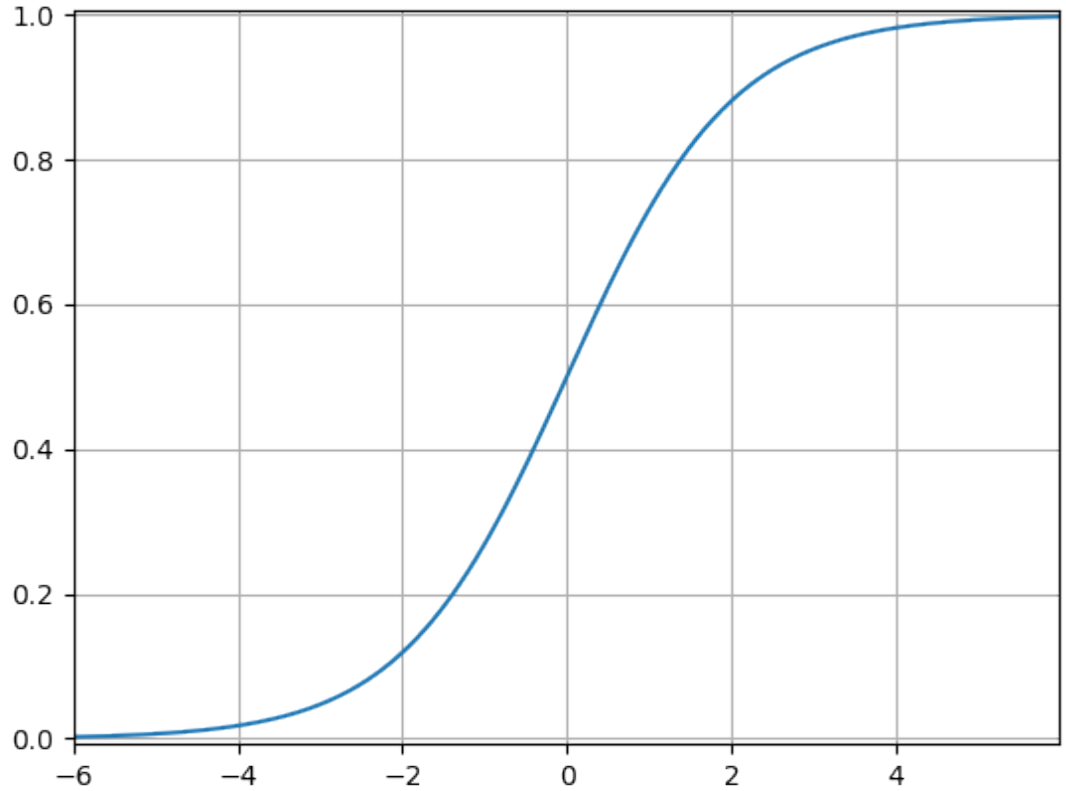
- aumenti quando la funzione lineare aumenta
- sia 50% quando la funzione lineare vale zero

Come convertire $w^\top x$ in una probabilità?



La funzione sigmoide (sigmoide logistica)

$$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad \in [0, 1]$$



Alcune proprietà della sigmoide

Se $\sigma(z) = (1 + \exp(-z))^{-1}$, allora per ogni $z \in \mathbb{R}$,

$$1 - \sigma(z) = \sigma(-z)$$

Se $\sigma(z) = (1 + \exp(-z))^{-1}$, allora per ogni $z \in \mathbb{R}$,

$$\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$$

Regressione logistica binaria (etichette 0/1)

Assumiamo che:

$$\Pr(y = 1|x) = \sigma(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}$$

Ne consegue che:

$$\Pr(y = 0|x) = 1 - \sigma(w^\top x) = \sigma(-w^\top x) = \frac{1}{1 + \exp(w^\top x)}$$

Qualunque sia $y \in \{0, 1\}$,

$$\Pr(y|x) = (h(x))^y \, (1 - h(x))^{1-y}$$

dove $h(x) = \sigma(w^\top x)$

La classe di ipotesi della regressione logistica binaria

Nella *regressione logistica*, l'insieme delle ipotesi è l'insieme $\mathcal{H}_{sig}$ delle funzioni ottenute componendo la sigmoide con una funzione lineare da $\mathcal{X} \subseteq \mathbb{R}^{d+1}$ a $\mathbb{R}$:

$$h \in \mathcal{H}_{sig} \quad \Leftrightarrow \quad h(x) = \sigma(w^\top x) \text{ per qualche } w \in \mathbb{R}^{d+1}$$

MLE nella regressione logistica (etichette 0/1)

Principio di massima verosimiglianza

Dati gli esempi $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, scegli $w \in \mathbb{R}^{d+1}$ che massimizza la funzione di *verosimiglianza* [*likelihood*]:

$$\mathcal{L}(w) = \prod_{i=1}^{m} \Pr(y^{(i)} | x^{(i)}; w)$$

Massimizzare $\mathcal{L}(w)$ equivale a minimizzare la *cross-entropia binaria*:

$$\sum_{i=1}^{m} \left[-y^{(i)} \log h(x^{(i)}) - (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right]$$

Funzione costo nella regressione logistica (etichette 0/1)

In altre parole stiamo assumendo la seguente funzione costo:

Funzione costo *cross-entropia binaria* (etichette 0/1)

$$\ell(h(x), y) = \begin{cases} -\log h(x) & \text{se } y = 1 \\ -\log(1 - h(x)) & \text{se } y = 0 \end{cases}$$

È una funzione convessa nel vettore w dei parametri

Rischio empirico nella regressione logistica (etichette 0/1)

Il principio MLE in questo caso è quindi equivalente al principio Empirical Risk Minimization con la funzione obiettivo cross-entropia

ERM nella regressione logistica (etichette 0/1)

Dati gli esempi $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, scegli $w \in \mathbb{R}^{d+1}$ che minimizza

$$\begin{aligned} \mathrm{RE}_S(w) &= -\frac{1}{m} \sum_{i=1}^{m} \log \Pr(y^{(i)} | x^{(i)}; w) \\ &= \frac{1}{m} \sum_{i=1}^{m} \left[-y^{(i)} \log h(x^{(i)}) - (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right] \end{aligned}$$

SGD per la regressione logistica (etichette 0/1)

Possiamo minimizzare il costo con i metodi gradiente

Per esempio con SGD: $w \leftarrow w - \eta \nabla \ell(w)$

Calcolando $\nabla \ell(w)$ e sfruttando $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ otteniamo

$$\nabla \ell(w) = (h(x) - y)x$$

Regola SGD per la regressione logistica (0/1)

$$w \leftarrow w - \eta \cdot (h(x) - y) \cdot x$$

NB. La regola ha la stessa struttura della regola Least Mean Squares (LMS), ma il significato di $h(x)$ è differente

Regressione logistica binaria (etichette ± 1)

La nostra assunzione diviene:

$$\Pr(y = +1|x) = \sigma(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}$$

Ne consegue che:

$$\Pr(y = -1|x) = \sigma(-w^\top x) = \frac{1}{1 + \exp(w^\top x)}$$

Qualunque sia $y \in \{-1, +1\}$,

$$\Pr(y|x) = \sigma(y \cdot w^\top x) = \frac{1}{1 + \exp(-y \cdot w^\top x)}$$

Funzione costo nella regressione logistica (etichette ± 1)

La funzione di costo si può ora esprimere come:

Funzione costo *log-loss* (etichette ± 1)

$$\begin{aligned}\ell &= -\log \Pr(y|x) \\ &= \log(1 + \exp(-y \cdot w^\top x)) \\ &= \log\left[\exp(0) + \exp(-y \cdot w^\top x)\right] \\ &= \mathrm{lse}(0, -y \cdot w^\top x)\end{aligned}$$

dove la funzione lse (*Log-Sum-Exp*) è

$$\mathrm{lse}(a, b) \stackrel{\text{def}}{=} \log\left[\exp(a) + \exp(b)\right]$$

Funzione Log-Sum-Exp

Funzione *lse*

Per $a_1, a_2, \ldots, a_K \in \mathbb{R}$,

$$\text{lse}(a_1, a_2, \ldots, a_K) \stackrel{\text{def}}{=} \log\left[\exp(a_1) + \exp(a_2) + \ldots + \exp(a_K)\right]$$

Ha analogie con la funzione $\max(a_1, \ldots, a_K)$:

- Quando $a_i \gg a_j$ per ogni $j \neq i$, $\text{lse}(a_1, \ldots, a_K) \approx a_i$
- Si ha sempre
  $\max(a_1, \ldots, a_K) \leq \text{lse}(a_1, \ldots, a_K) \leq (\log K) + \max(a_1, \ldots, a_K)$
- È una funzione convessa del vettore a
- È differenziabile ovunque

MLE nella regressione logistica

Principio di massima verosimiglianza

Dati gli esempi $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, scegli $w \in \mathbb{R}^{d+1}$ che massimizza la funzione di *verosimiglianza* [*likelihood*]:

$$\prod_{i=1}^{m} \Pr(y^{(i)} | x^{(i)}; w)$$

Rischio empirico nella regressione logistica (etichette ± 1)

Equivalentemente, passando al logaritmo, vogliamo *minimizzare* il rischio empirico, in linea col principio ERM:

ERM nella regressione logistica (etichette ± 1)

Dati gli esempi $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, scegli $w \in \mathbb{R}^{d+1}$ che minimizza

$$\begin{aligned} \mathrm{RE}_S(w) &= -\frac{1}{m} \sum_{i=1}^{m} \log \Pr(y^{(i)} | x^{(i)}; w) \\ &= \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y^{(i)} \cdot w^\top x^{(i)})) \end{aligned}$$

Minimizzazione del rischio empirico nella regressione logistica

Rischio empirico nella regressione logistica (etichette ± 1)

$$\mathrm{RE}_S(w) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y^{(i)} \cdot w^\top x^{(i)}))$$

- Il minimizzante w^* non è esprimibile in forma chiusa
- $\mathrm{RE}_S(w)$ è una funzione convessa in w

$\Rightarrow$ Il problema di ottimizzazione corrispondente è convesso

$\Rightarrow$ Possiamo trovare w^* attraverso Gradient Descent o le sue varianti

Si possono usare anche metodi del secondo ordine (la funzione è sia convessa che ovunque differenziabile)

SGD per la regressione logistica (etichette ± 1)

Per derivare la regola di aggiornamento di SGD ricapitoliamo le assunzioni:

- Ipotesi in $\mathcal{H}_{sig}$: $h(x) = \sigma(w^\top x)$
- Costo log-loss: $\ell = \log(1 + \exp(-y \cdot w^\top x))$

Prendendo le derivate parziali di ℓ, otteniamo

$$\begin{aligned}
\frac{\partial \ell}{\partial w_j} &= \frac{1}{1 + \exp(-yw^\top x)} \frac{\partial}{\partial w_j} \left[1 + \exp(-yw^\top x)\right] \\
&= \frac{\exp(-yw^\top x)}{1 + \exp(-yw^\top x)} \frac{\partial}{\partial w_j} \left[-yw^\top x\right] \\
&= -\frac{1}{1 + \exp(+yw^\top x)} \cdot y \cdot x_j \\
&= -\Pr(-y|x) \cdot y \cdot x_j
\end{aligned}$$

SGD per la regressione logistica (etichette ± 1)

In forma vettoriale,

$$\nabla \ell(w) = -\Pr(-y|x; w) \cdot y \cdot x$$

La regola di aggiornamento SGD (con etichette ± 1) è quindi

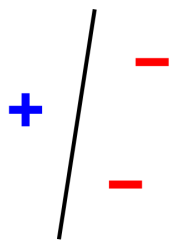
$$w \leftarrow w + \eta \cdot \Pr(-y|x;\, w) \cdot y \cdot x$$

NB. Gli aggiornamenti del vettore w sono equivalenti a quelli visti per le etichette 0/1.

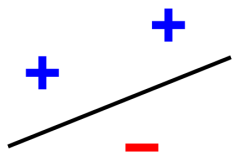
Separabilità lineare

Un insieme di esempi (x, y) con etichette di due tipi ($+$ e $-$) è *linearmente separabile* se esiste $w \in \mathbb{R}^{d+1}$ tale che:

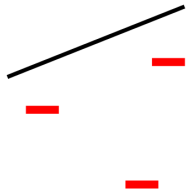
- $w^\top x > 0$ ogniqualvolta x è un esempio di tipo $+$
- $w^\top x < 0$ ogniqualvolta x è un esempio di tipo $-$



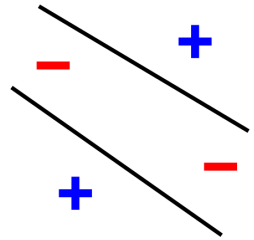
separabile



separabile



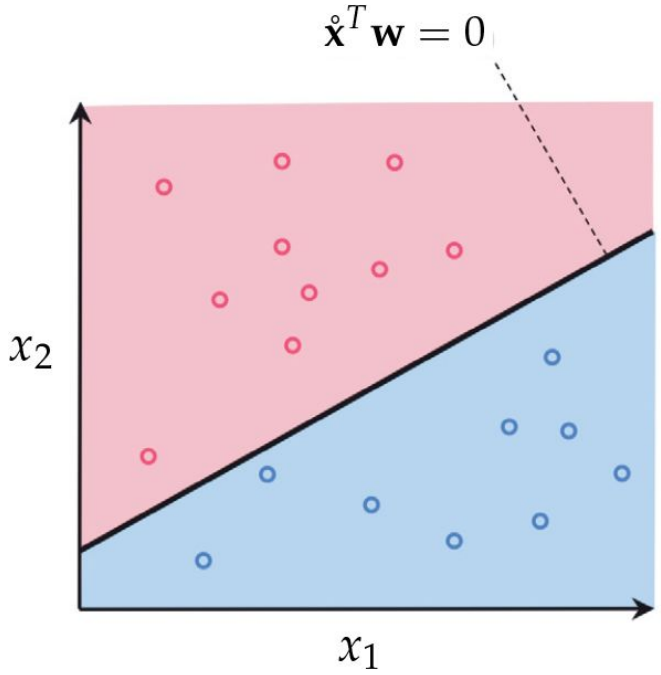
separabile



non separabile

Valori $y \cdot w^\top x$ e separabilità lineare

Supponiamo che il vettore w separi perfettamente gli esempi positivi e negativi:



Questo significa che per ogni (x, y),

- $y = +1$ e $w^\top x > 0$, oppure
- $y = -1$ e $w^\top x < 0$

In altre parole, $-y \cdot w^\top x$ è sempre < 0

Costo log-loss e separabilità lineare

Se w separa perfettamente gli esempi positivi e negativi:

$$-y \cdot w^\top x < 0 \qquad \text{per ogni } (x, y)$$

allora qualunque sia il costo log-loss

$$\log \left[1 + \exp(-y \cdot w^\top x)\right] > 0$$

possiamo diminuirlo semplicemente usando $2w$ invece di w:

$$\log \left[1 + \exp(-y \cdot 2w^\top x)\right] < \log \left[1 + \exp(-y \cdot w^\top x)\right]$$

Questo significa che il minimo della funzione si ha per $\|w\| \to \infty$!

La frontiera di decisione definita da w e $2w$ è la stessa, ma il divergere di w può causare instabilità numerica negli algoritmi

Regressione logistica regolarizzata

Per evitare il divergere di w, si può considerare il problema di ottimizzazione vincolata

$$\begin{aligned} \underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \ & \frac{1}{m} \sum_{i=1}^{m} \log \left[1 + \exp(-y^{(i)} \cdot w^\top x^{(i)}) \right] \\ \text{subject to} \ & \|\omega\|_2^2 \leq 1 \end{aligned}$$

che (per qualche $\lambda > 0$) equivale alla seguente formulazione

Regressione logistica con regolarizzazione ℓ_2

$$\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \ \frac{1}{m} \sum_{i=1}^{m} \log \left[1 + \exp(-y^{(i)} \cdot w^\top x^{(i)}) \right] + \lambda \|\omega\|_2^2$$

Regressione logistica in `scikit-learn`

Regressione logistica	Iperparametri	Interfaccia `scikit-learn`
Non regolarizzata	–	`LogisticRegression(penalty='none')`
Regolarizzata	$C\ (= \frac{1}{2\lambda})$	`LogisticRegression(penalty='l2', C)`

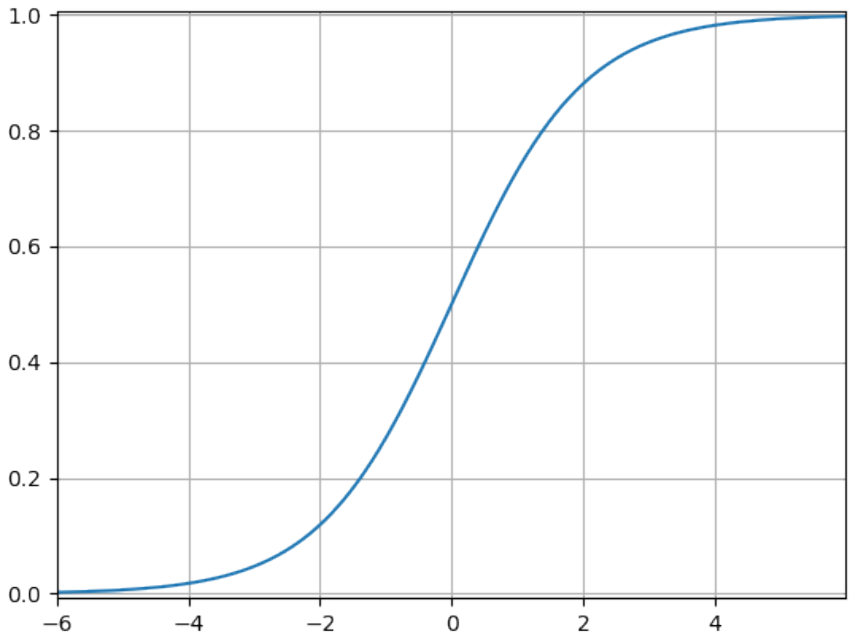
Interfaccia Stochastic Gradient Descent con funzione costo *log-loss*:

Regressione logistica (± 1)	Iperparametri	Interfaccia `scikit-learn`
Regolarizzata	$\alpha\ (= \lambda), \eta, T$	`SGDClassifier(loss='log_loss', alpha, learning_rate, max_iter)`

Percettroni e Support Vector Machines

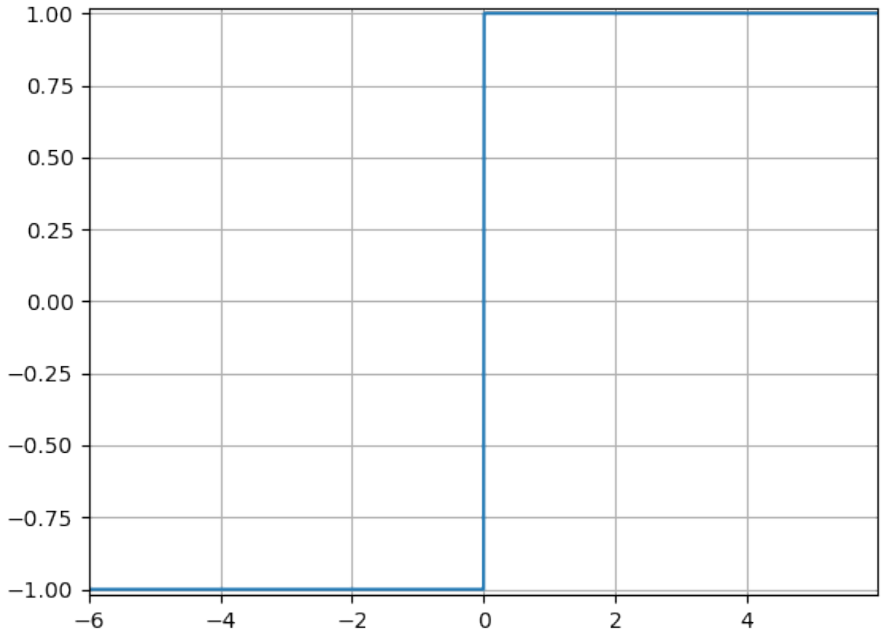
Il regressore logistico

- $x = (1, x_1, \ldots, x_d)$, $p \in [0, 1]$ (probabilità di $y = 1|x$)
- $w = (w_0, \ldots, w_d)$
- $\hat{p} = h(x) = \sigma(w^\top x) = \frac{1}{1+\exp(-w^\top x)}$



Il Percettrone

- $x = (1, x_1, \ldots, x_d)$, $y \in \{-1, +1\}$ (etichetta)
- $w = (w_0, \ldots, w_d)$
- $\hat{y} = h(x) = \text{sign}(w^\top x)$



Se la vera etichetta è y, la predizione è corretta $\Leftrightarrow y \cdot w^\top x > 0$

La classe di ipotesi del Percettrone

Nel *Percettrone*, l'insieme delle ipotesi è l'insieme $\mathcal{H}_P$ delle funzioni ottenute componendo la funzione sign con una funzione lineare da $\mathcal{X}$ a $\mathbb{R}$:

$$h \in \mathcal{H}_P \Leftrightarrow h(x) = \mathrm{sign}(w^\top x) \text{ per qualche } w \in \mathbb{R}^{d+1}$$

Una funzione di costo per il Percettrone

La funzione più naturale (costo 0/1) purtroppo è ardua da ottimizzare! (si ha un problema NP-arduo)

Cerchiamo di costruire un *surrogato* della funzione costo:

- Se $y \cdot w^\top x > 0$, poniamo costo $= 0$ (la predizione è corretta)
- Se $y \cdot w^\top x \leq 0$, poniamo costo $= -y \cdot w^\top x$

Perceptron Loss (per etichette ± 1)

$$\ell \stackrel{\text{def}}{=} \max(-y \cdot w^\top x, 0)$$

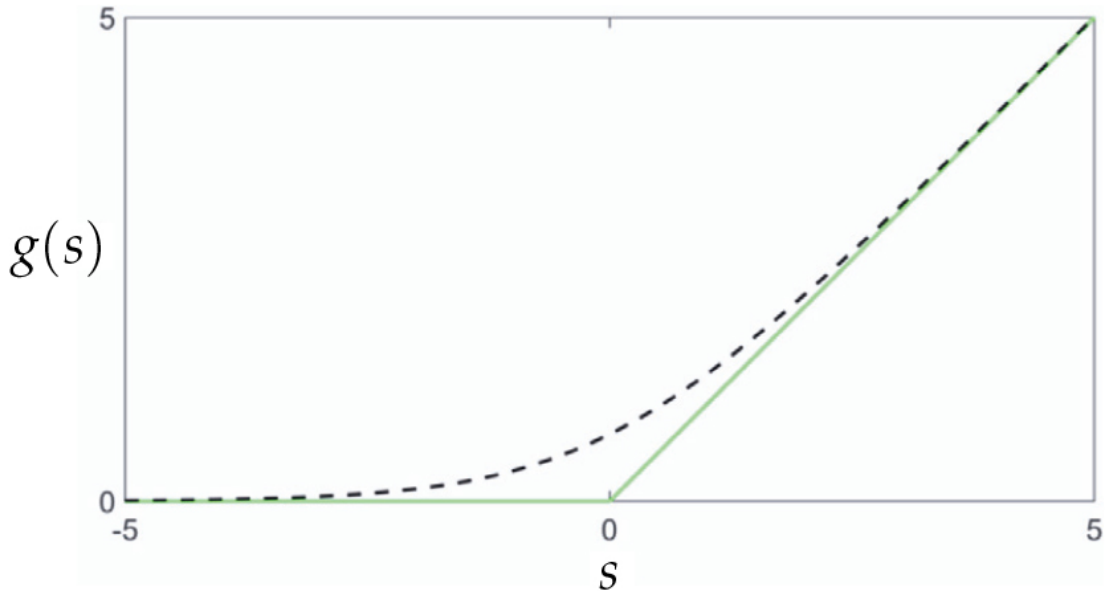
Per ogni (x, y), questa funzione costo è convessa in w, quindi possiamo usarla con i metodi gradiente!

Notare la similarità formale con la funzione costo log-loss: $\ell = \mathsf{lse}(-y \cdot w^\top x, 0)$

Funzione costo: Percettrone vs. regressione logistica

$$\ell_{percettrone} = \max(-y \cdot w^\top x, 0)$$

$$\ell_{logistica} = \mathrm{lse}(-y \cdot w^\top x, 0)$$



$g(s) = \max(s, 0)$ è chiamata *ReLU* (Rectified Linear Unit)

- È convessa
- È differenziabile ovunque tranne in $s = 0$

Minimizzazione del rischio empirico nel percettrone

Rischio empirico nel percettrone

$$\mathrm{RE}_S(w) = \frac{1}{m}\sum_{i=1}^{m} \max(-y^{(i)} \cdot w^\top x^{(i)}, 0)$$

- $\mathrm{RE}_S(w)$ è una funzione convessa in w

$\Rightarrow$ Il problema di ottimizzazione corrispondente è convesso

$\Rightarrow$ Possiamo trovare w^* attraverso Gradient Descent o le sue varianti

Non è ottimizzabile con metodi del secondo ordine

SGD per il Percettrone

- Se $y \cdot w^\top x > 0$, allora $\ell = 0$ (la predizione è corretta)
- Se $y \cdot w^\top x \leq 0$, allora $\ell = -y \cdot w^\top x$

Calcolando il gradiente del costo, otteniamo:

- Nel primo caso, $\nabla \ell = 0$
- Nel secondo caso, $\nabla \ell = -y \cdot x$

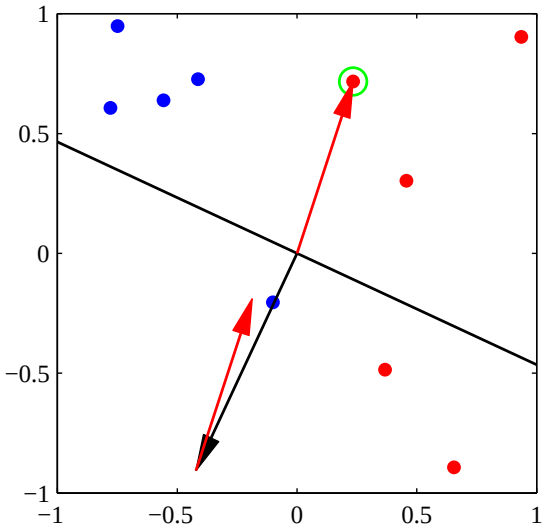
Regola di aggiornamento SGD per il Percettrone

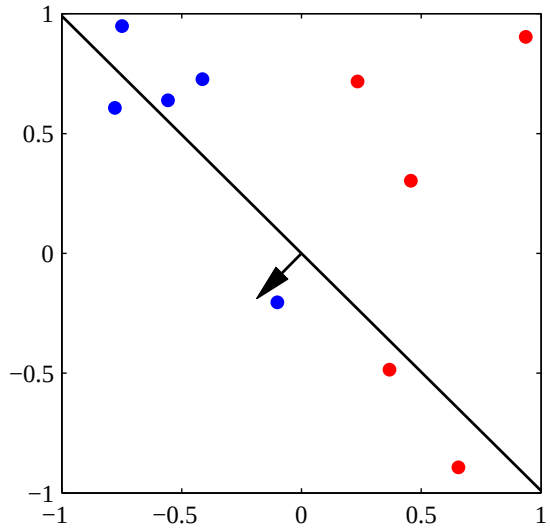
- Se $y \cdot w^\top x > 0$, poni $w \leftarrow w - \eta \cdot 0$
- Se $y \cdot w^\top x \leq 0$, poni $w \leftarrow w + \eta \cdot y \cdot x$

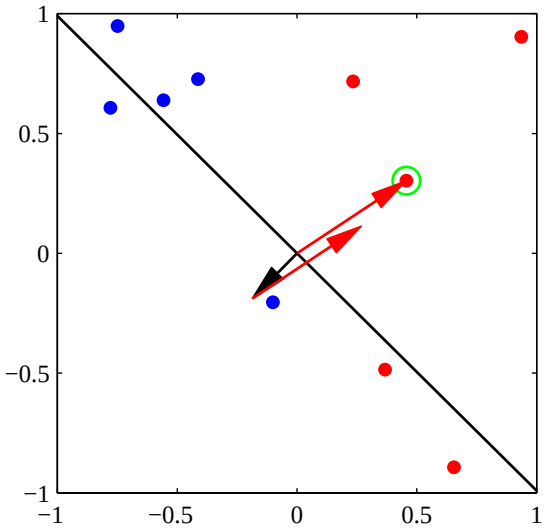
Algoritmo del Percettrone

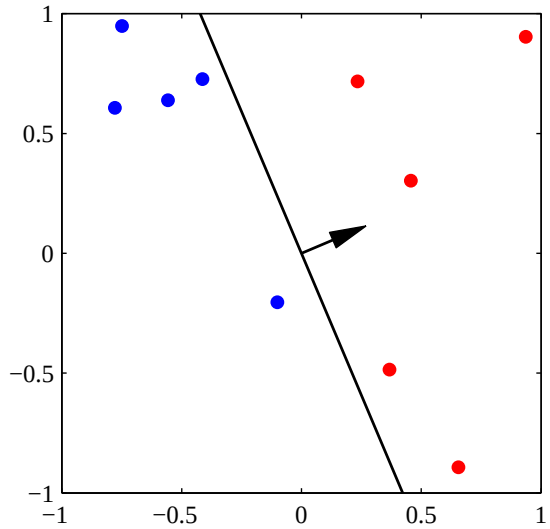
1. Inizializza $w^{(1)} = 0_{(d+1)\times 1}$
2. Per $t = 1, 2, \ldots,$, considera ciclicamente ogni esempio (x, y):
   - Se $y \cdot w^{(t)\top} x \leq 0$, aggiorna $w^{(t+1)} \leftarrow w^{(t)} + \eta\, y \cdot x$

Algoritmo del Percettrone









Convergenza del Percettrone

Teorema (Convergenza del Percettrone)

Se il training set è linearmente separabile:

- L'algoritmo del Percettrone trova un'ipotesi con rischio empirico pari a zero (= separa perfettamente gli esempi di training)
- L'algoritmo converge in un numero finito di passi

Forma duale del Percettrone

Algoritmo del Percettrone

1. Inizializza $w = 0_{(d+1)\times 1}$
2. Per $t = 1, 2, \ldots,$ considera ciclicamente ogni esempio (x, y):
   - Se (x, y) è misclassificato, aggiorna $w \leftarrow w + \eta\, y \cdot x$

Quindi l'output dell'algoritmo avrà la forma

$$w = \eta \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

dove α_i è il # di volte che l'esempio i-esimo ha causato un aggiornamento

$\Rightarrow$ Possiamo rappresentare w tramite $\alpha = (\alpha_1, \ldots, \alpha_m)$ (variabili *duali*)

Forma duale del Percettrone

Algoritmo del Percettrone (forma primale)

1. Inizializza $w = 0_{(d+1)\times 1}$
2. Per $t = 1, 2, \ldots,$ considera ciclicamente ogni esempio $(x^{(i)}, y^{(i)})$:
   - Se $(x^{(i)}, y^{(i)})$ è misclassificato, aggiorna $w \leftarrow w + \eta\, y^{(i)} \cdot x^{(i)}$

Algoritmo del Percettrone (forma duale)

1. Inizializza $\alpha = 0_{m\times 1}$
2. Per $t = 1, 2, \ldots,$ considera ciclicamente ogni esempio $(x^{(i)}, y^{(i)})$:
   - Se $(x^{(i)}, y^{(i)})$ è misclassificato, poni $\alpha_i \leftarrow \alpha_i + 1$
3. Restituisci $w = \eta \sum_i \alpha_i y^{(i)} x^{(i)}$

Forma primale vs. forma duale: confronto computazionale

Sia k il costo del calcolo di un prodotto scalare tra elementi di $\mathcal{X}$

Forma primale

- 1 predizione: calcola $w^\top x$ in tempo $O(d)$
- 1 aggiornamento: aggiorna $w \in \mathbb{R}^{d+1}$ in tempo $O(d)$

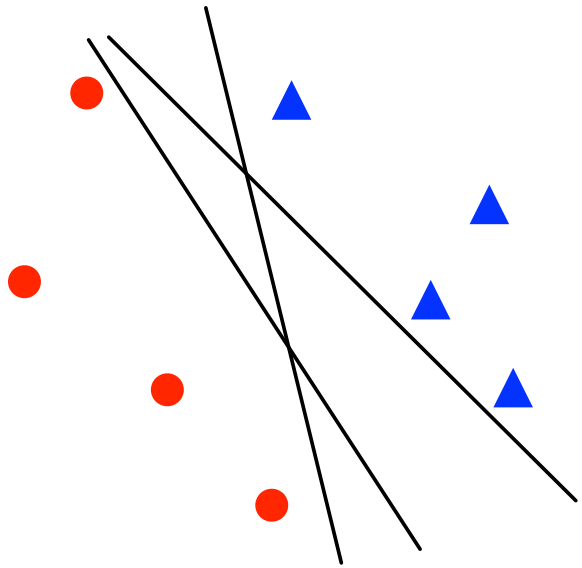
Forma duale

- 1 predizione: calcola $w^\top x = \eta \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle$ in tempo $O(mk)$
- 1 aggiornamento: aggiorna $\alpha_i \in \mathbb{N}$ in tempo $O(1)$

La forma duale può risultare conveniente se $mk \ll d$

La forma duale è anche chiamata *kernel perceptron*

Oltre il Percettrone



Domanda: Possiamo selezionare il separatore più "robusto"?

Verso una separazione robusta

Dati: m esempi $(x, y) \in \mathbb{R}^{d+1} \times \{-1, +1\}$
Trova: $w \in \mathbb{R}^{d+1}$ tale che

$$y \cdot w^\top x > 0 \qquad \text{per ogni esempio } (x, y)$$

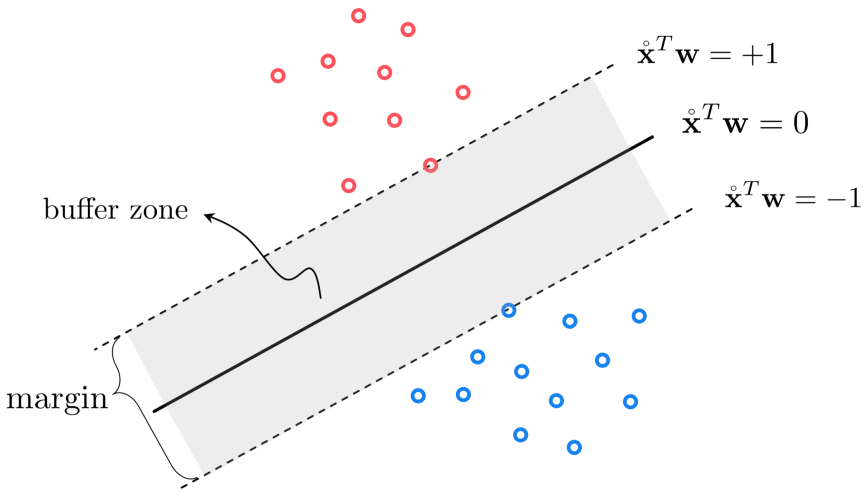
Scalando w, è equivalente a richiedere

$$y \cdot w^\top x \geq 1 \qquad \text{per ogni esempio } (x, y)$$

Margine di un separatore lineare

Margine di un separatore

Il *margine* di $w \in \mathbb{R}^{d+1}$ è $2/\left\|(w_1, \ldots, w_d)\right\| = 2/\left\|\omega\right\|$



massimizzare il margine $\equiv$ minimizzare $w_1^2 + w_2^2 + \ldots + w_d^2$

Separazione a massimo margine

Hard-margin Support Vector Machine (Hard SVM)

$$\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}}\ w_1^2 + w_2^2 + \ldots + w_d^2$$

$$\text{subject to } y \cdot w^\top x \geq 1 \qquad \text{per ogni esempio } (x, y)$$

È un problema di minimizzazione convessa vincolata:

- funzione obiettivo convessa
- vincoli lineari

$\Rightarrow$ è risolubile in modo efficiente

Importante: ha soluzione solo se gli esempi sono linearmente separabili

Hard SVM: formulazione alternativa

$$\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \ \|\omega\|_2^2$$

$$\text{subject to } y \cdot w^\top x \geq 1 \qquad \text{per ogni esempio } (x, y)$$

è equivalente a

$$\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \ \|\omega\|_2^2$$

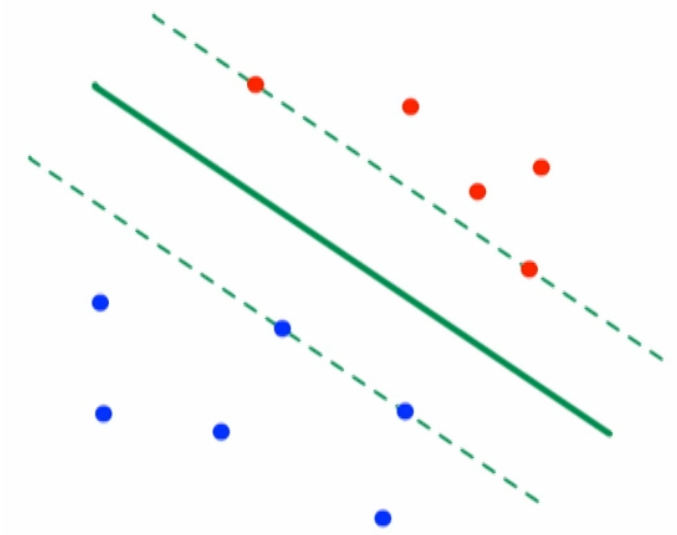
$$\text{subject to } \max(0, 1 - y \cdot w^\top x) = 0 \text{ per ogni esempio } (x, y)$$

Vettori di supporto

Si può dimostrare che la soluzione ottima ha la forma

$$w^* = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

con $\alpha_i \neq 0$ solo per gli $x^{(i)}$ giacenti sul margine (*vettori di supporto*)



Ma cosa fare se il dataset non è linearmente separabile?

Il caso non separabile

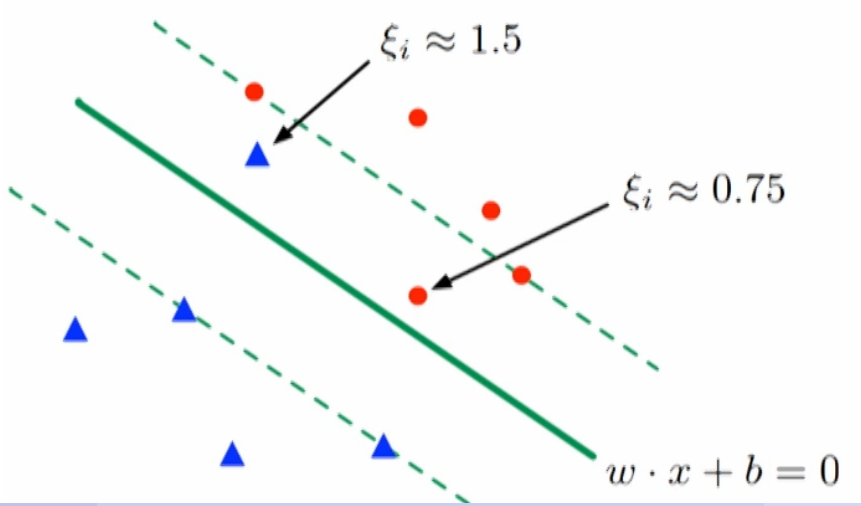
Introduciamo delle variabili di *slack* ("allentamento"):

Soft-Margin Support Vector Machine (Soft SVM)

$$\underset{w}{\text{minimize}} \ \|\omega\|_2^2 + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to } y^{(i)} \cdot w^\top x^{(i)} \geq 1 - \xi_i \text{ per } i = 1, 2, \ldots, m$$

$$\xi \geq 0$$



Compromesso tra margine e slack

Che ruolo gioca l'iperparametro C?

Soft-Margin Support Vector Machine

$$\begin{aligned} &\underset{w}{\text{minimize}} \ \|\omega\|_2^2 + C \sum_{i=1}^{m} \xi_i \\ &\text{subject to } y^{(i)} \cdot w^\top x^{(i)} \geq 1 - \xi_i \text{ per } i = 1, 2, \ldots, m \\ &\qquad \xi \geq 0 \end{aligned}$$

$C = 0$: restituisce $w^* = 0$ (gli allentamenti non sono penalizzati)

$C \to \infty$: equivalente a Hard-margin SVM

Soft-SVM: formulazione alternativa

$$\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \ \|\omega\|_2^2 + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to } y^{(i)} \cdot w^\top x^{(i)} \geq 1 - \xi_i \qquad \text{per ogni esempio } (x^{(i)}, y^{(i)})$$

$$\xi_i \geq 0$$

è equivalente a

$$\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \ \|\omega\|_2^2 + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to } \max(0, 1 - y^{(i)} \cdot w^\top x^{(i)}) = \xi_i \text{ per ogni esempio } (x^{(i)}, y^{(i)})$$

Percettrone e SVM in `scikit-learn`

Approccio	Iperparametri	Interfaccia `scikit-learn`
Percettrone	T	`Perceptron(max_iter)`
Soft SVM (hinge loss)	C	`LinearSVC(C)`

Interfacce Stochastic Gradient Descent:

Approccio	Iperparametri	Interfaccia `scikit-learn`
Percettrone	T, η	`SGDClassifier(loss='perceptron', penalty=None, max_iter, learning_rate)`
Soft SVM (hinge loss)	$\alpha(=\frac{1}{C}), T, \eta$	`SGDClassifier(loss='hinge', penalty='l2', alpha, max_iter, learning_rate)`

Classificazione binaria: metriche di qualità

Duplice ruolo delle funzioni di costo

La funzione di costo per misurare la qualità delle predizioni nei problemi di classificazione è in genere la funzione costo 0-1:

$$\ell(\hat{y}, y) = \begin{cases} 0 & \text{se } \hat{y} = y \\ 1 & \text{se } \hat{y} \neq y \end{cases}$$

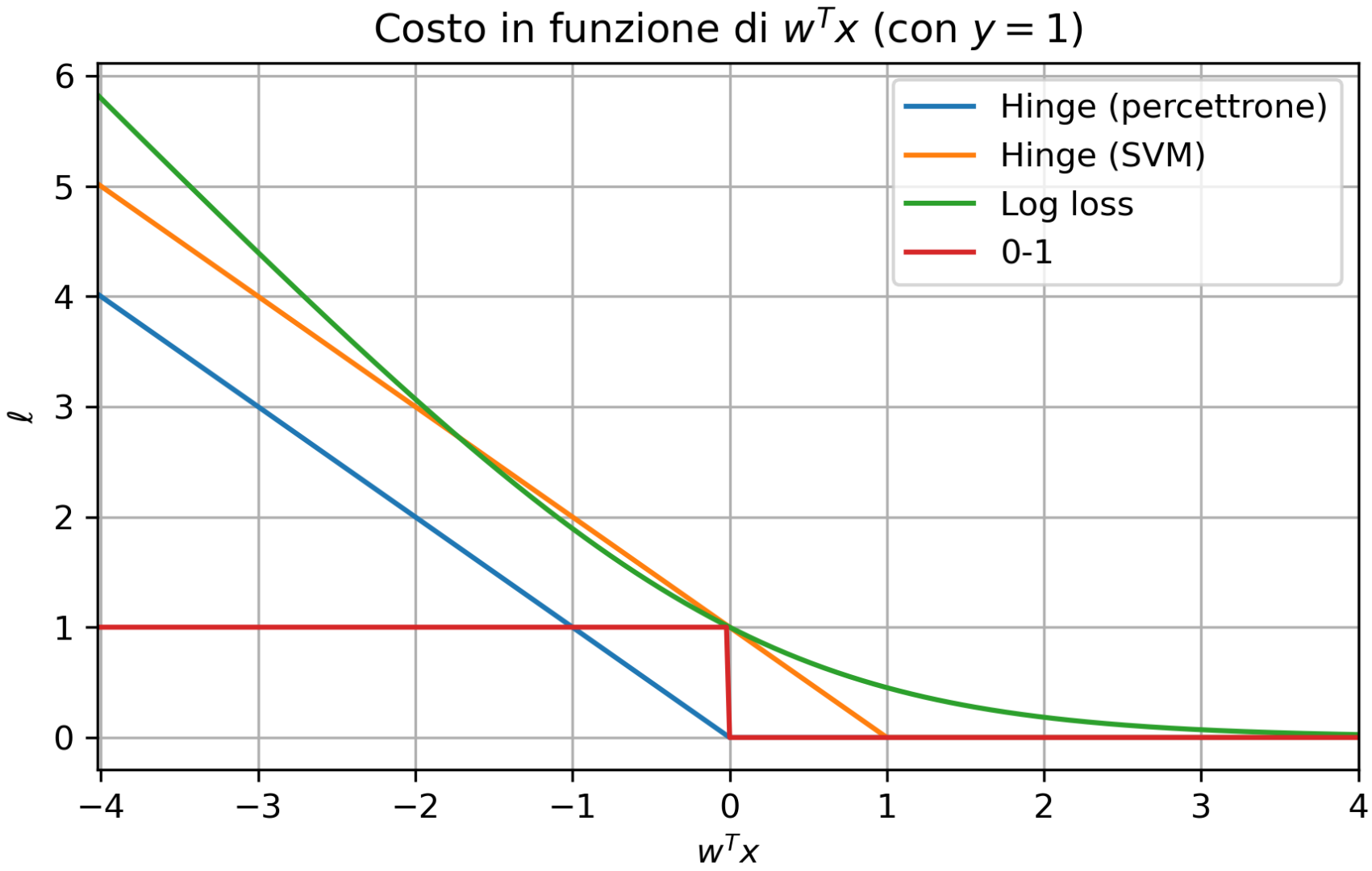
$\Rightarrow$ *Accuratezza*: frazione di esempi correttamente classificati

La funzione di costo per apprendere il modello è un suo *surrogato*:

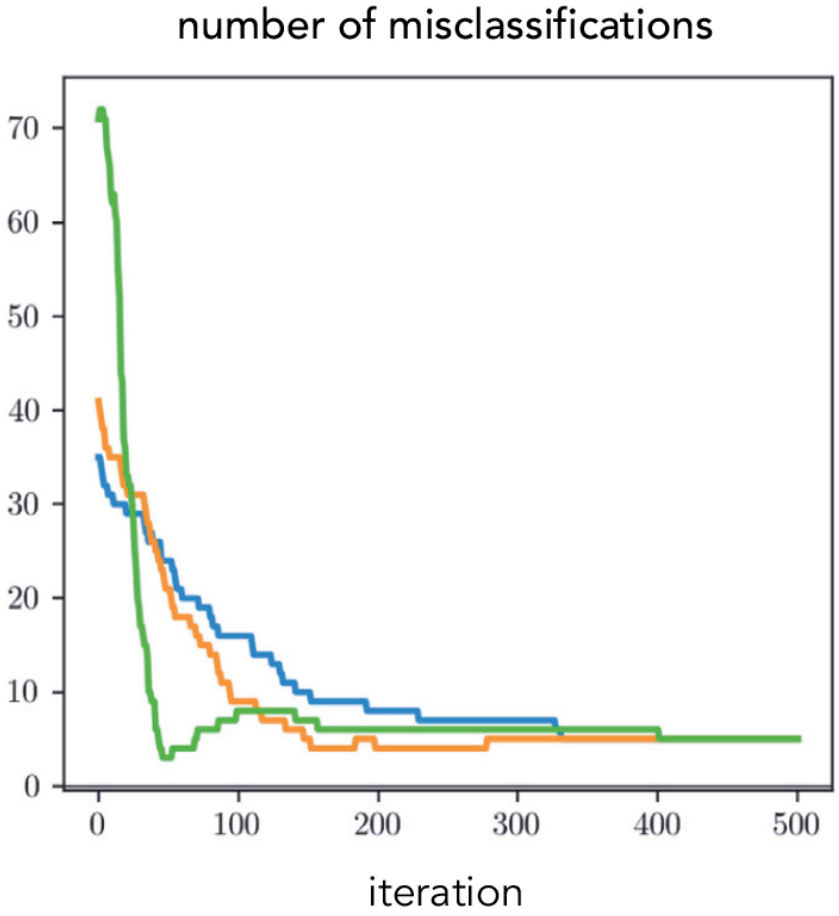
- Log loss/cross-entropia (regressione logistica)
- Perceptron loss (percettrone)
- Hinge loss (SVM)

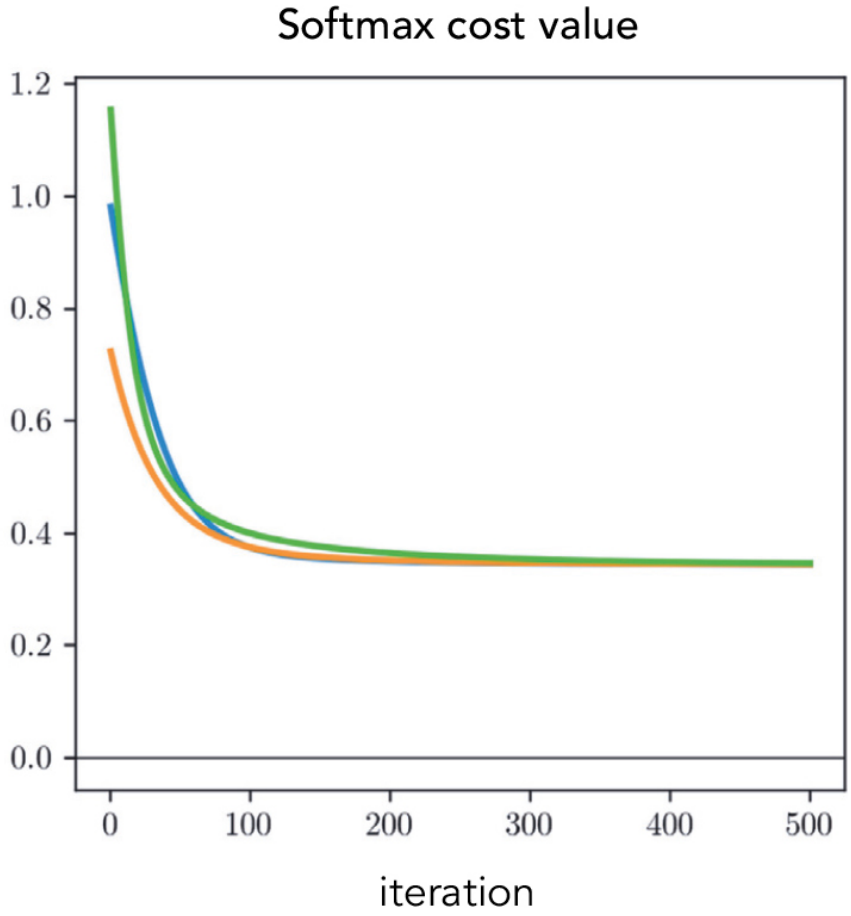
Il modello va validato sulla funzione originale, non sul surrogato
$\Rightarrow$ si usa l'(in)accuratezza come metrica durante validazione e test

Costi surrogati vs. costo 0-1



Esempio


number of misclassifications
70
60
50
40
30
20
10
0
0
100
200
300
400
500
iteration


Softmax cost value
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
iteration

Inconvenienti dell'accuratezza come metrica

L'accuratezza a volte può presentare inconvenienti

- se diversi tipi di misclassificazione hanno costo molto diverso
- se c'è uno sbilanciamento tra le classi, in cui i casi positivi (o quelli negativi) sono rari

Esempio: diagnosi di una malattia rara ($<$1 per mille della popolazione). Un classificatore che restituisce sempre `NO` ha accuratezza 99.9%, ma naturalmente è del tutto inutile

Matrice di confusione

Per problemi con forte sbilanciamento, è utile separare i tipi di errore attraverso una *matrice di confusione*

	$y = 1$	$y = 0$
$\hat{y} = 1$	*Veri Positivi* Abbiamo urlato al lupo! Abbiamo salvato il villaggio.	*Falsi Positivi* Errore: il lupo non c'era. Abbiamo innervosito tutti.
$\hat{y} = 0$	*Falsi Negativi* C'era un lupo, ma non lo abbiamo stanato. Ha mangiato tutto il pollame.	*Veri Negativi* Nessun lupo, nessun allarme. Tutto tranquillo.

Accuratezza e matrice di confusione

Accuratezza [accuracy]

$$\mathcal{A} = \frac{VP + VN}{VP + VN + FP + FN}$$

Sensibilità, specificità e precisione

Sensibilità [sensitivity, true positive rate, recall]

$$\text{sensibilità} \stackrel{\text{def}}{=} \frac{VP}{VP + FN} = \mathcal{A}_{y=1}$$

Specificità [specificity, true negative rate]

$$\text{specificità} \stackrel{\text{def}}{=} \frac{VN}{VN + FP} = \mathcal{A}_{y=0}$$

Precisione [precision]

$$\text{precisione} \stackrel{\text{def}}{=} \frac{VP}{VP + FP} = \mathcal{A}_{\hat{y}=1}$$

Esempio: Test SARS-CoV-2

Tipo di test	**Sensibilità**	**Specificità**
Test molecolare (rt-PCR)	91%–100%	81%–100%
Test antigenico rapido	70%–86%	95%–97%

(Fonte: Istituto Superiore di Sanità, *Test di laboratorio per SARS-CoV-2 e loro uso in sanità pubblica*)

Accuratezza bilanciata

$$\begin{aligned}\mathcal{A}_{\text{balanced}} &\overset{\text{def}}{=} \text{media aritmetica}\,(\text{sensibilità}, \text{specificità}) \\ &= \frac{1}{2}\left(\mathcal{A}_{y=1} + \mathcal{A}_{y=0}\right) \\ &= \frac{1}{2}\frac{VP}{VP+FN} + \frac{1}{2}\frac{VN}{VN+FP}\end{aligned}$$

Attenzione. La formula nel libro di testo (Watt et al. formula (6.81)) non è corretta (confonde la sensibilità con la precisione)

Metrica F_1 [F_1 score]

Metrica F_1

$$\mathcal{F}_1 \overset{\text{def}}{=} \text{media armonica (sensibilità, precisione)}$$

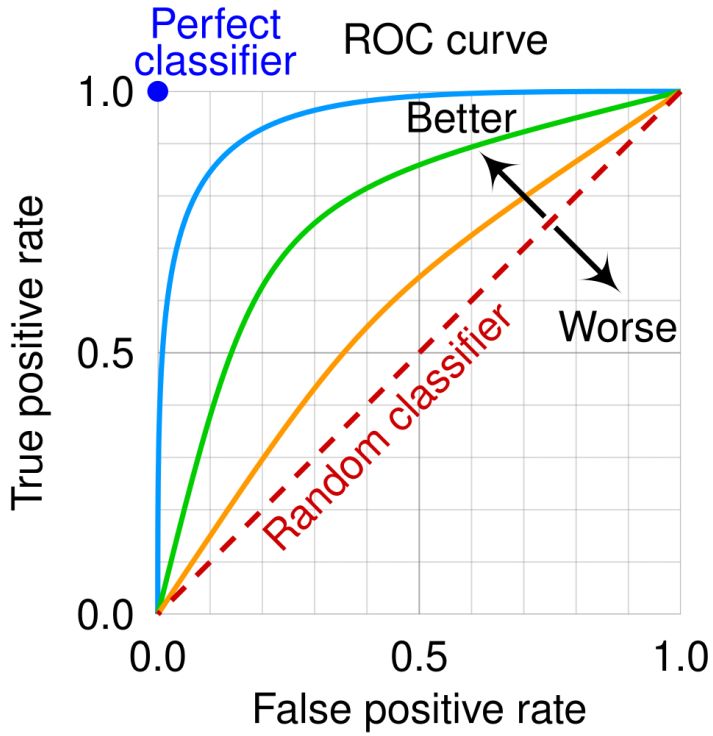
$$\frac{1}{\mathcal{F}_1} = \frac{1}{2}\left(\frac{1}{\mathcal{A}_{y=1}} + \frac{1}{\mathcal{A}_{\hat{y}=1}}\right)$$

$$= \frac{1}{2}\frac{VP + FN}{VP} + \frac{1}{2}\frac{VP + FP}{VP}$$

Può risultare utile quando si preferisce una metrica indipendente dal numero di Veri Negativi (VN)

Curva ROC e Area-Under-the-Curve (AUC)

La *curva ROC* esprime la relazione tra la sensibilità (TPR $= \mathcal{A}_{y=1}$) e il complemento della specificità (FPR $= 1 - \mathcal{A}_{y=0}$) al variare della soglia di decisione



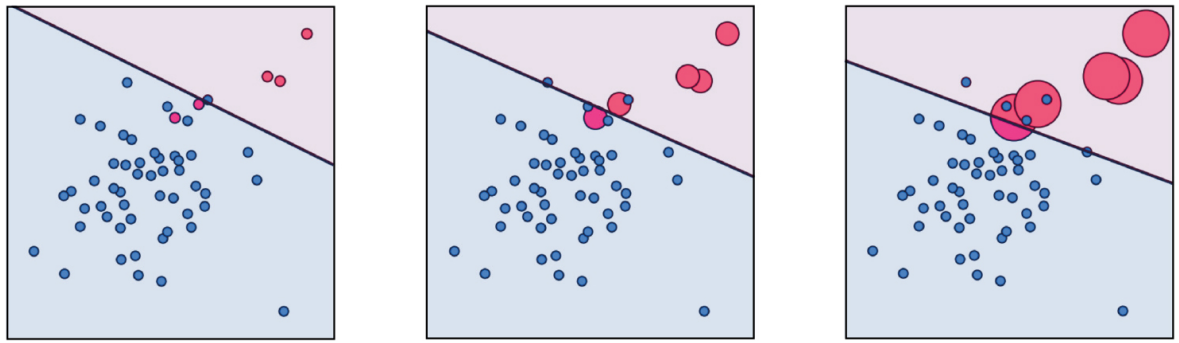
La metrica *Area-under-the-Curve (AUC)* misura l'area sotto la curva

Pesatura degli esempi

Per regolare l'influenza degli esempi di una classe sbilanciata, possiamo assegnare ad ogni esempio un *peso* β_i, come visto per la regressione

Per esempio, nella regressione logistica possiamo minimizzare

$$\mathrm{RE}_S(w) = \frac{1}{m} \sum_{i=1}^{m} \beta_i \log(1 + \exp(-y^{(i)} \cdot w^\top x^{(i)}))$$

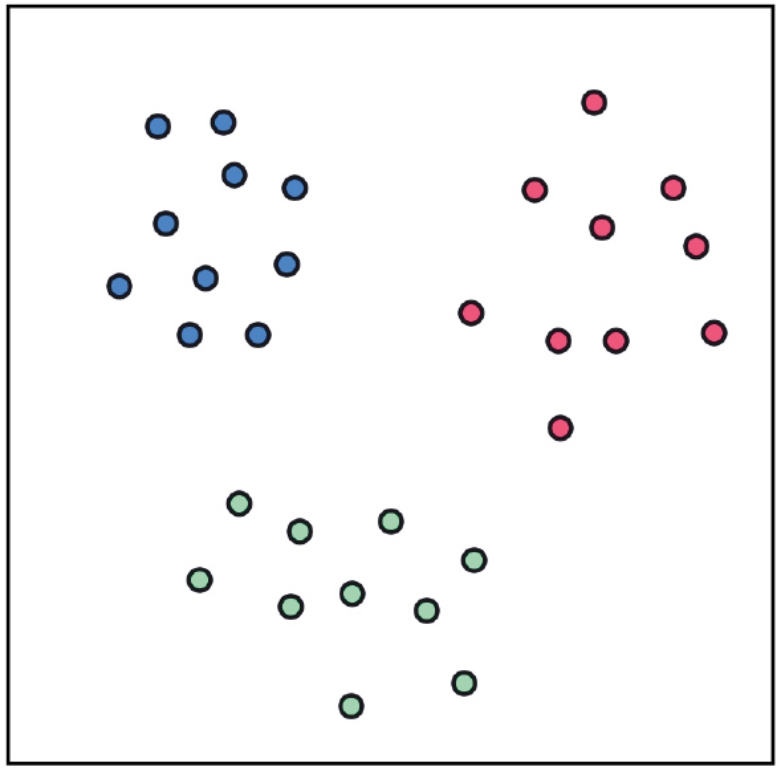


Ad esempio, si può prendere β_i inversamente proporzionale alla taglia della classe $y^{(i)}$

Classificazione multiclasse

Classificazione multiclasse

Come trattare il caso di K classi con $K > 2$?

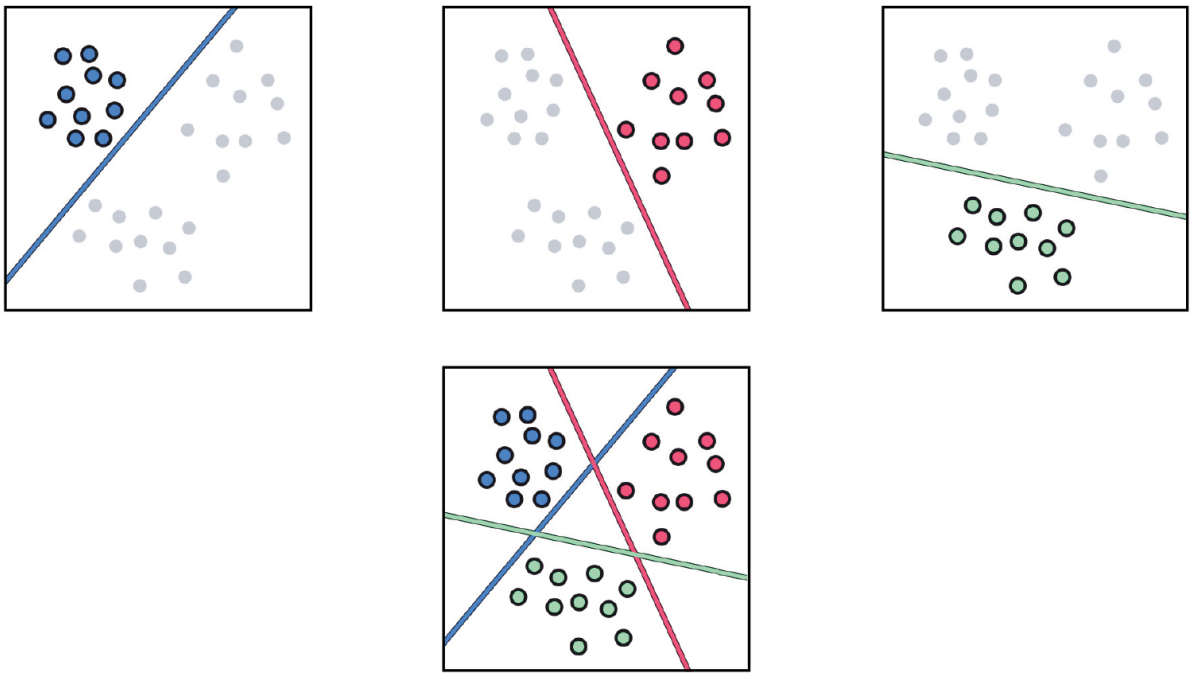


1. Approccio *one vs. rest* (applichiamo K volte un classificatore binario)
2. *Softmax multiclasse* (generalizzazione della regressione logistica)

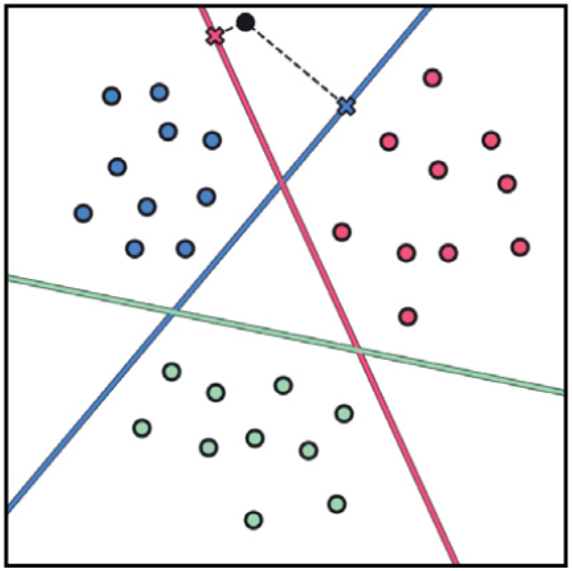
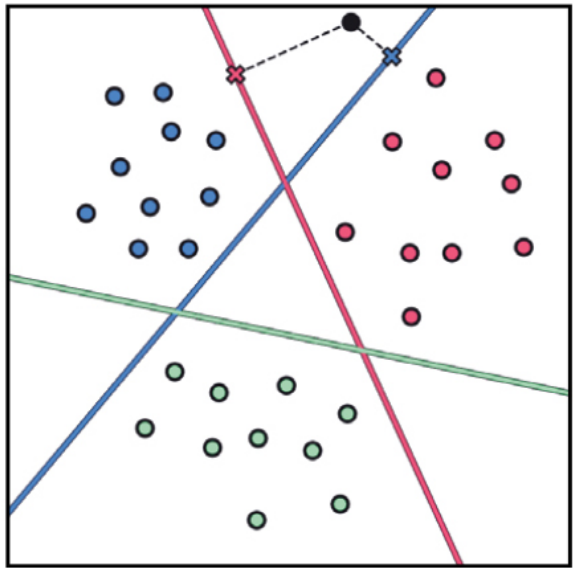
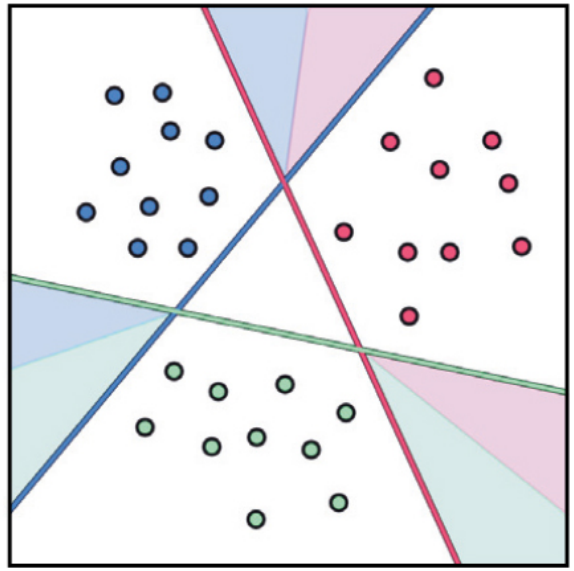
One vs. rest

Approccio *One vs. rest*

1. Apprendi K ipotesi $h^{(1)}, \ldots, h^{(K)}$, dove la j-esima ipotesi distingue la classe j dalle altre $K-1$ classi
2. Dato x, restituisci la classe j che massimizza $h^{(j)}(x)$



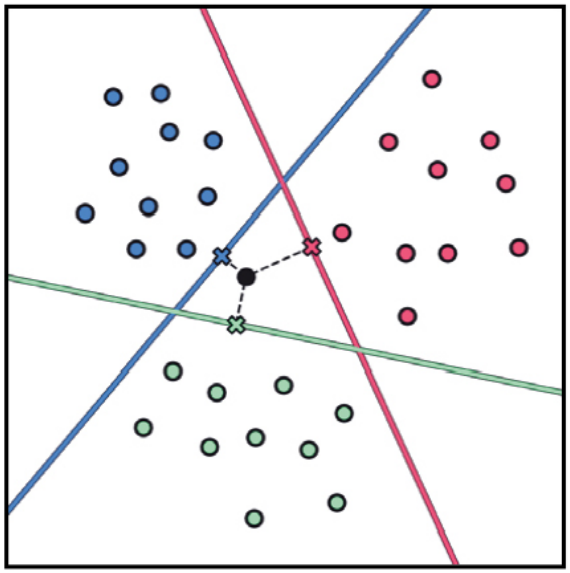
One vs. rest

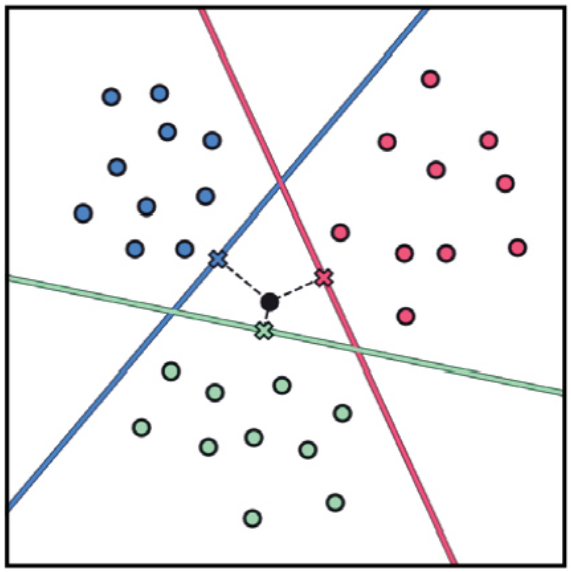




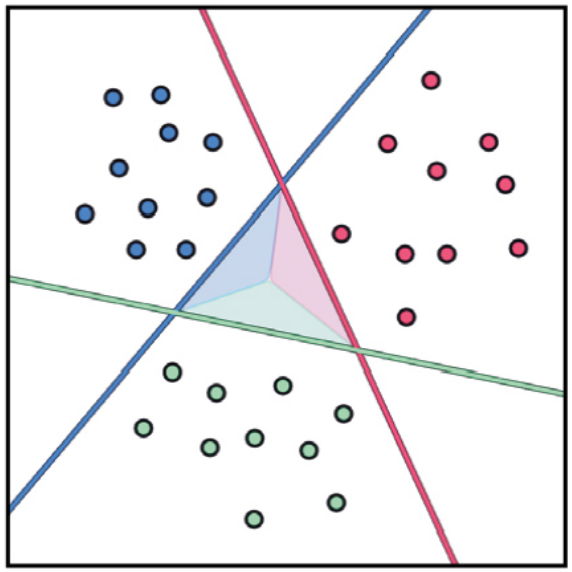
$$\underset{j=1}{\overset{K}{\operatorname{argmax}}}\, h^{(j)}(x) = \underset{j=1}{\overset{K}{\operatorname{argmax}}}\, \sigma(w^{(j)\top} x) = \underset{j=1}{\overset{K}{\operatorname{argmax}}}\, w^{(j)\top} x$$

Importante: assume che le componenti di $w^{(j)}$ siano state normalizzate, in modo che $\left\|\omega^{(j)}\right\| = 1$ per ogni j (altrimenti il confronto non è significativo)

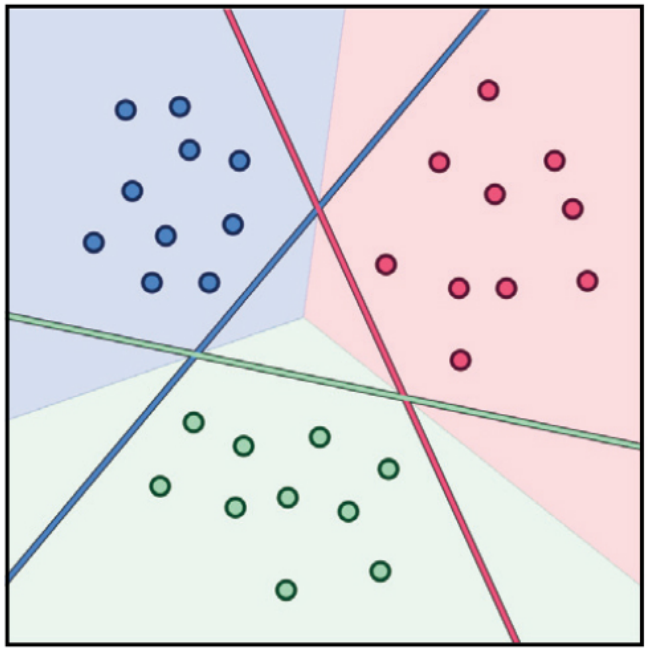
One vs. rest

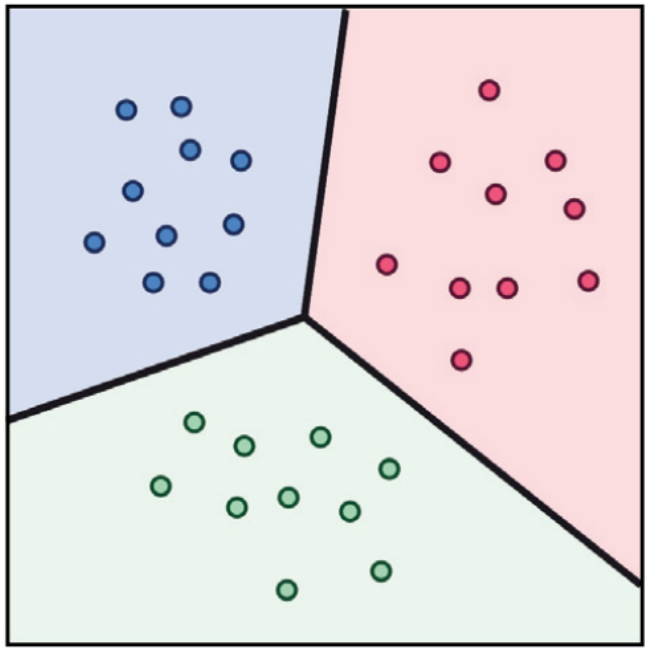




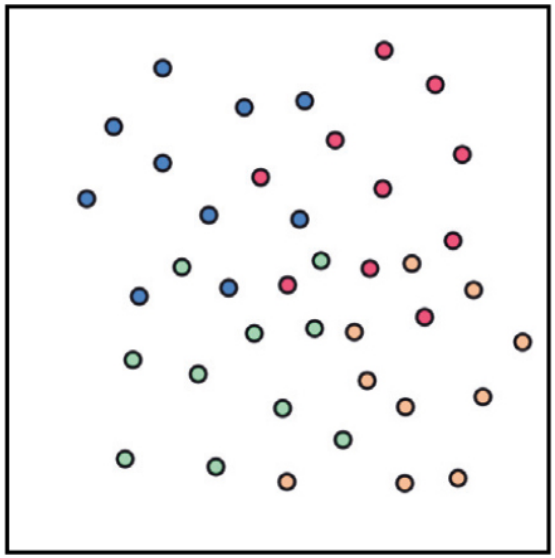


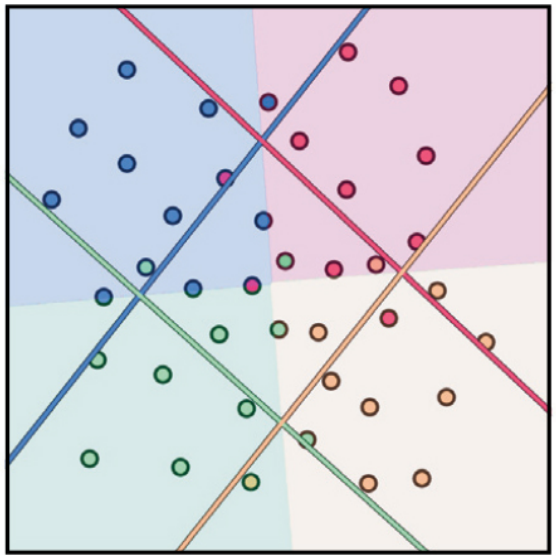
One vs. rest

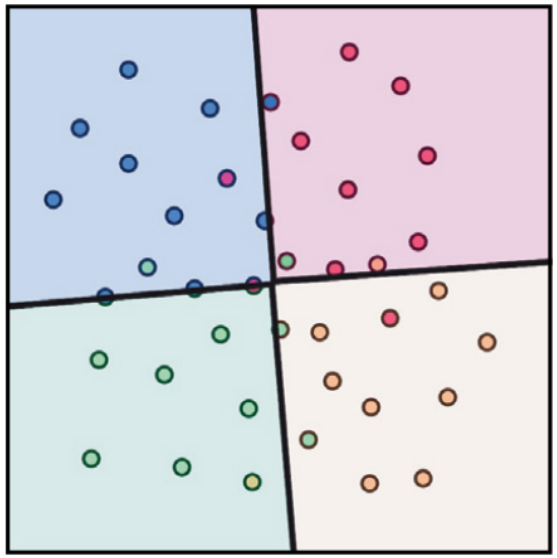




One vs. rest: un altro esempio







Softmax multiclasse

Approccio *softmax multiclasse* (o *multinomiale*)

1. Assumi
$$\Pr(y = j|x) = \frac{\exp(w^{(j)\top} x)}{\sum_{k=1}^{K} \exp(w^{(k)\top} x)}$$
2. Ottimizza i vettori $w^{(1)}, \ldots, w^{(K-1)}$ per massimizzare la verosimiglianza $\mathcal{L}(w)$ (senza perdita di generalità, $w^{(K)} = 0$)
3. Dato x, restituisci la classe j che massimizza $\Pr(y = j|x)$

È una vera generalizzazione della regressione logistica binaria:

- Quando $K = 2$, coincide con la regressione logistica binaria
- Anche per $K > 2$, la funzione da minimizzare (al punto 2) è convessa e differenziabile

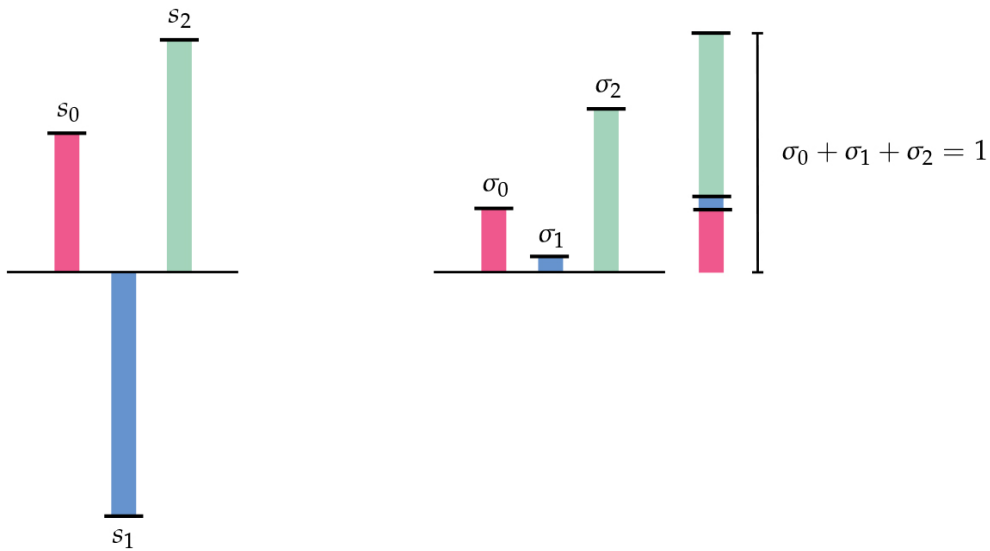
Interpretazione del softmax multiclasse

L'esponenziale normalizzato

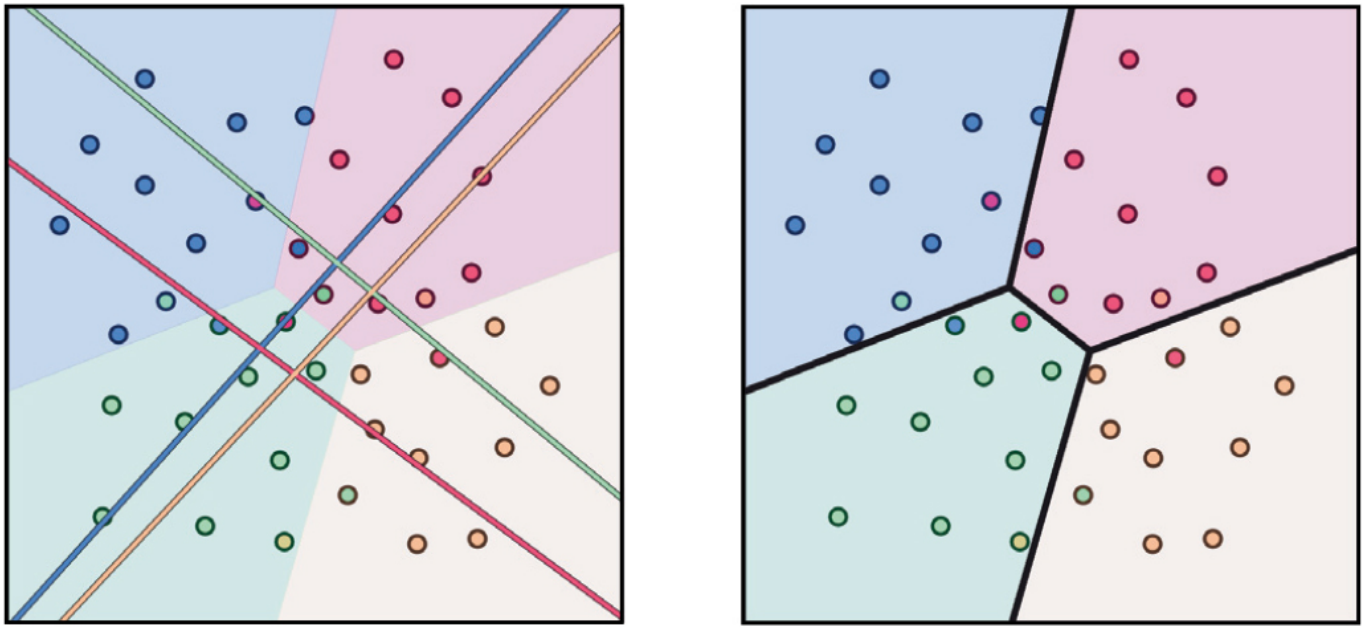
$$\frac{\exp(s_j)}{\sum_{k=1}^{K}\exp(s_k)} = \frac{\exp(s_j)}{\exp(\mathrm{lse}(s_1,\ldots,s_k))}$$

converte il vettore $(s_1,\ldots,s_K)$ in una distribuzione di probabilità: tutti gli esponenziali sono positivi, e

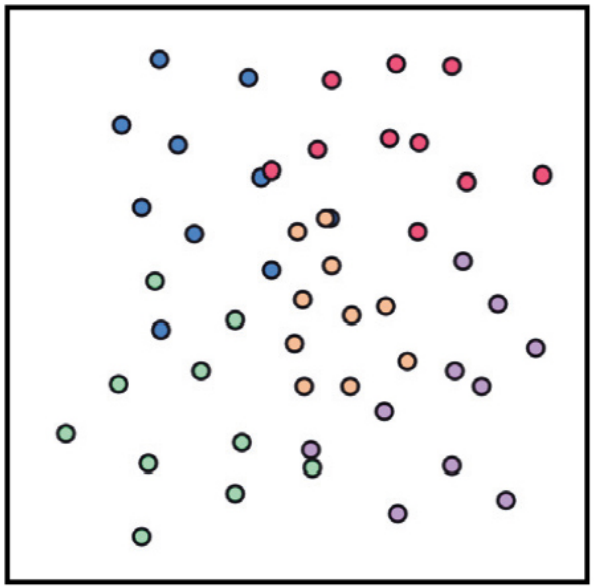
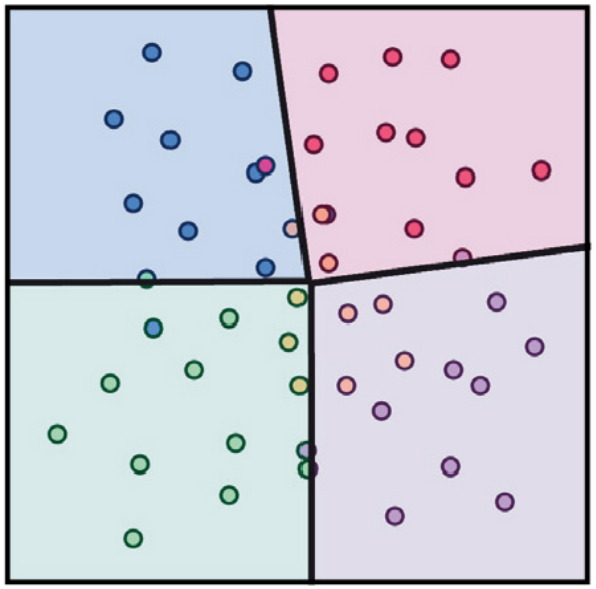
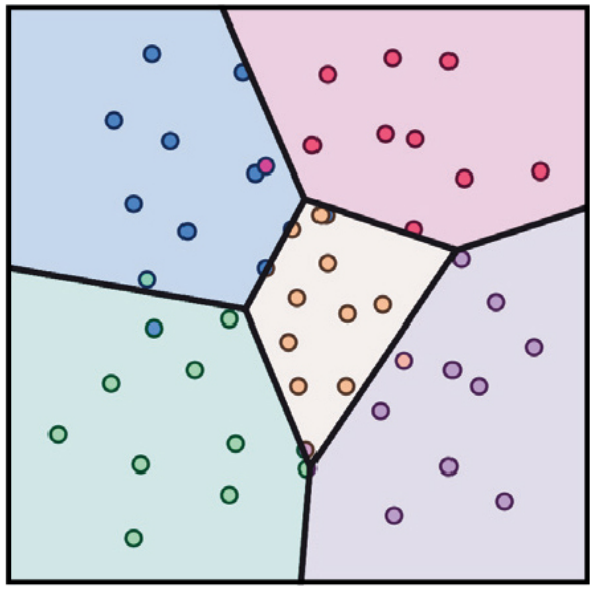
$$\sum_{j=1}^{K}\frac{\exp(s_j)}{\sum_{k=1}^{K}\exp(s_k)} = 1$$



Softmax multiclasse: esempio



Confronto tra One vs. rest e Softmax multiclasse

In `scikit-learn` con la regressione logistica:

```
LogisticRegression(multi_class='ovr')
LogisticRegression(multi_class='multinomial')
```

Matrice di confusione e accuratezza nel caso multiclasse

La matrice di confusione diventa $K \times K$:

	red	blue	green	yellow
red	8	1	1	0
blue	1	7	1	1
green	1	1	7	1
yellow	0	1	1	8

La funzione costo 0–1 è ancora applicabile
$\Rightarrow$ lo è anche la definizione dell' (in)accuratezza

Nell'esempio, l'accuratezza sarà

$$\frac{8+7+7+8}{8+1+1+0+1+7+1+1+1+1+7+1+0+1+1+8}$$