Una classe di problemi di "Machine Learning": la classificazione

Classificare: assegnare un individuo ad una e solo una tra più possibili classi.

Esempi: classificare un paziente come sano oppure affetto da una certa patologia; classificare un messaggio come non-spam oppure spam etc ....

Un individuo è un campione definito attraverso un certo numero di parametri numerici (features) quantificabili

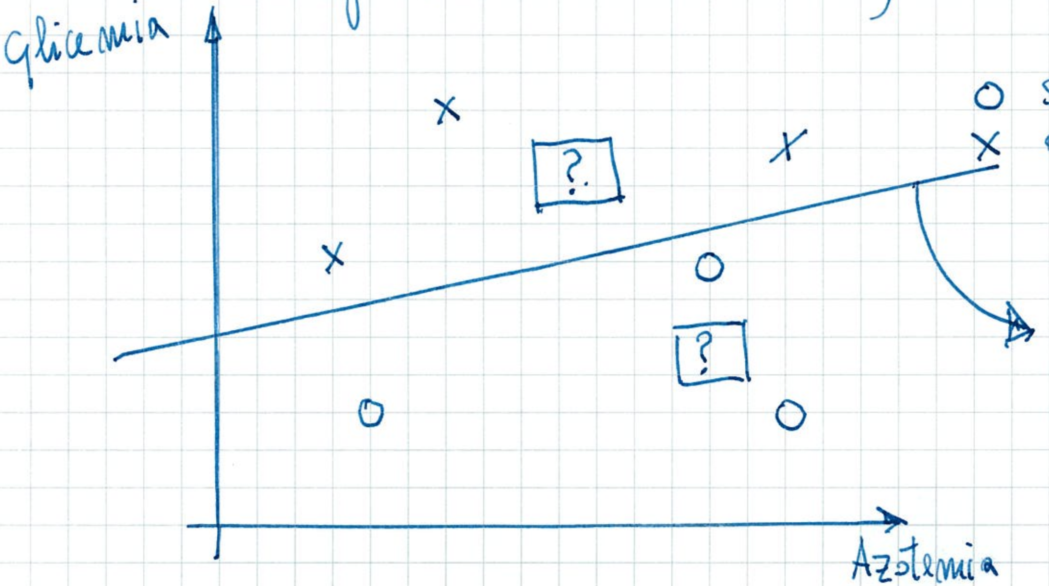
$\Rightarrow$ Un campione è rappresentato da un vettore in $\mathbb{R}^n$, dove n è il numero dei parametri.

Es. Un paziente con sospetto di diabete è caratterizzato da un unico parametro numerico, la glicemia.
È "noto" un valore soglia: 110 mg/dl
Questo numero è un classificatore: un paziente con valori di glicemia < 110 è classificato come sano in caso contrario come affetto da diabete (in prima istanza)

Un caso più complesso: una patologia associata a due parametri, per esempio (glicemia, azotemia). Come si estende il concetto di soglia? Come si costruisce un classificatore?

Si parte da una base di conoscenza, cioè un training set: un insieme di dati grazie ai quali addestrare il classificatore (terminologia mutuata dalle reti neurali)

Es. Di un certo numero di pazienti si conoscono i dati di glicemia e azotemia, ma anche l'appartenenza di classe



o sani

x affetti dalla patologia

[?] Nuovi pazienti

Questa retta fornisce un "classificatore" perché "separa" sani da malati

Un nuovo paziente di cui non è noto se è sano o ammalato viene classificato grazie a questa retta

Classificazione come problema di separazione dei campioni nello spazio dei parametri.

Nello spazio ad n dimensioni separazione mediante un iperpiano

Una retta di equazione $w_1 x_1 + w_2 x_2 = \gamma$ separa il piano in due semipiani; analogamente un iperpiano di equazione

$$w_1 x_1 + w_2 x_2 + \dots + w_n x_n = \gamma$$ separa lo spazio $\mathbb{R}^n$

in due semispazi;

$$H^- = \{ x \mid w^T x \leq \gamma \} \quad e \quad H^+ = \{ x \mid w^T x \geq \gamma \}$$

$w \in \mathbb{R}^n$ è la normale all'iperpiano $H = \{ x \mid w^T x = \gamma \}$

La classificazione come tecnica di apprendimento supervisionato (è nota la classe di appartenenza dei campioni)

Il clustering è non-supervisionato.

Sono dati $m+k$ campioni (vettori di $\mathbb{R}^n$):

$a^{(i)} \in \mathbb{R}^n$, $i = 1, \dots, m$; $b^{(l)} \in \mathbb{R}^n$, $l = 1, \dots, k$

Classificazione binaria: m soggetti di classe A e k di classe B

Problema: individuare un iperpiano $H = \{x \mid w^T x = \gamma\}$

con $w \in \mathbb{R}^n$ e $\gamma \in \mathbb{R}$ che separa i punti di A da quelli di B

Deve cioè accadere che

(1) $$a^{(i)T} w < \gamma, \quad i = 1, \dots, m \quad \text{and} \quad b_l^T w > \gamma, \quad l = 1, \dots, k$$

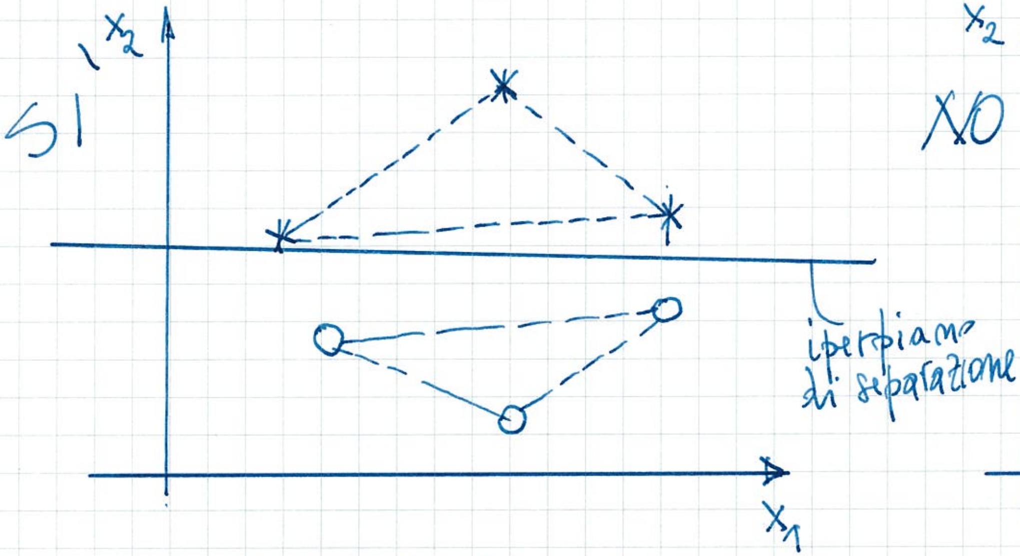
Individuare un iperpiano equivale a risolvere un sistema di $(m+k)$ disequazioni lineari in $\mathbb{R}^n$

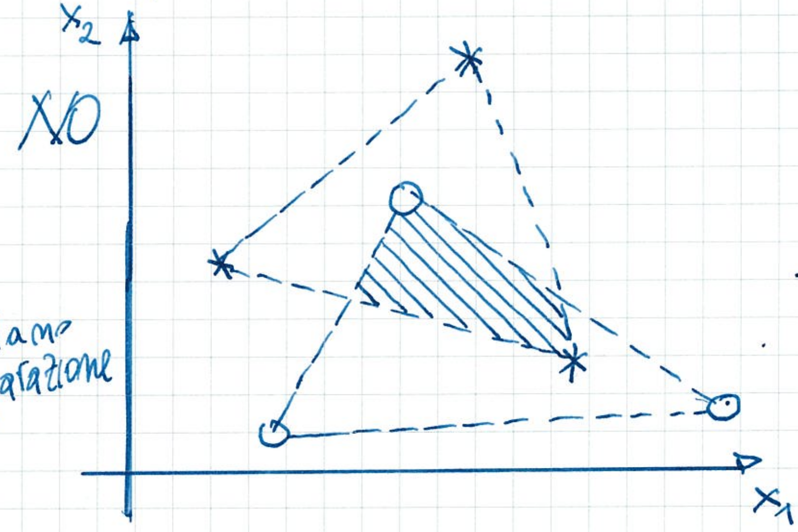
Si può dimostrare che il sistema (1) ammette soluzione se e solo se ammette soluzione il sistema

(2) $$\begin{cases} a^{(i)T} w \leq \gamma - 1, & i = 1, \dots, m \\ b^{(l)T} w \geq \gamma + 1, & l = 1, \dots, k \end{cases}$$

DIMOSTRAZIONE?

Il sistema (2) ammette sempre soluzione?





* Punti di classe A

o " " " B

Il sistema (2) ammette soluzione (esiste un iperpiano di separazione)

$$\Longleftrightarrow \quad \text{conv } A \cap \text{conv } B = \emptyset$$

Con Conv (X) si intende il più piccolo insieme convesso che contiene X. E' indicato come "copertura convessa di X"

Assegnato un data set $A \cup B$, non è facile valutare a priori se la condizione $conv(A) \cap conv(B) = \emptyset$ sia verificata

Problema: costruire un modello matematico che
  a) restituisca un iperpiano di separazione se i due insiemi sono separabili;
  b) restituisca un iperpiano che "quasi separa" i due insiemi nell'altro caso.

Costruiamo una "funzione d'errore"

Un iperpiano definito dalla coppia $(w \in \mathbb{R}^n, \gamma \in \mathbb{R})$ classifica correttamente un punto:

a) $a^{(i)}$ se $a^{(i)T} w - \gamma + 1 \leq 0 \Rightarrow \underbrace{e_i(w,\gamma) = \max\left(0, a^{(i)T} w - \gamma + 1\right)}$

errore di classificazione per il punto $a^{(i)}$ associato all'iper=piano $H(w,\gamma)$

b) $b^{(\ell)}$ se $-b^{(\ell)T} w + \gamma + 1 \leq 0 \Rightarrow \underbrace{e_\ell(w,\gamma) = \max\left(0, -b^{(\ell)T} w + \gamma + 1\right)}$

errore di classificazione per il punto $b^{(\ell)}$ associato all'iper=piano $H(w,\gamma)$

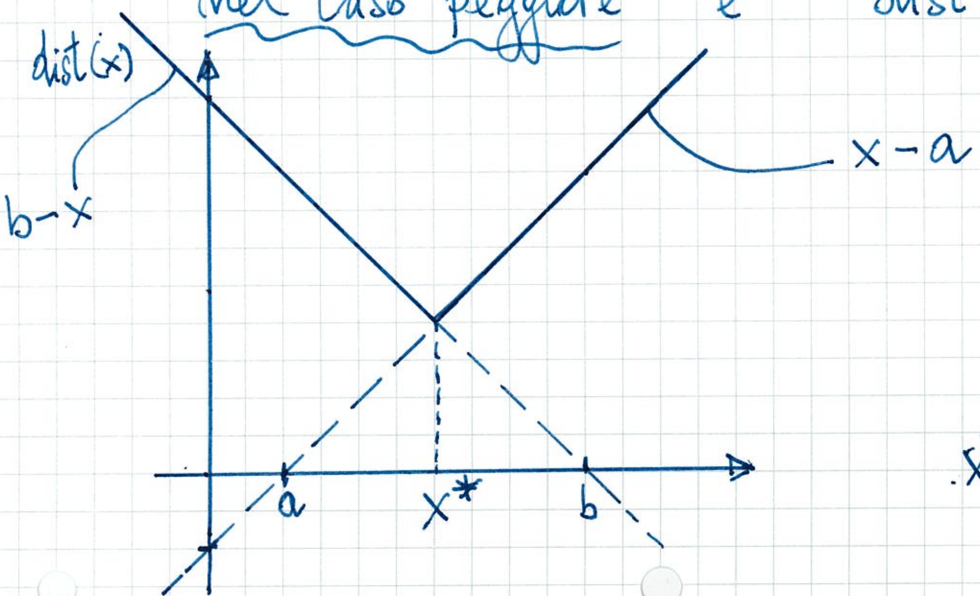
Problemi di min max (max min)

Una persona attende al binario di una stazione un amico (che ha un pesante bagaglio). Vuole aiutarlo, ma non ha alcuna informazione sulla carrozza su cui viaggia l'amico.
La persona sceglie il punto medio del marciapiede.

PERCHE'?

Il segmento $[a, b]$ rappresenta il marciapiede. Bisogna scegliere la posizione $x \in [a, b]$

Se si sceglie la posizione x, la distanza da percorrere nel caso peggiore è $\text{dist}(x) = \max(x-a,\ b-x)$



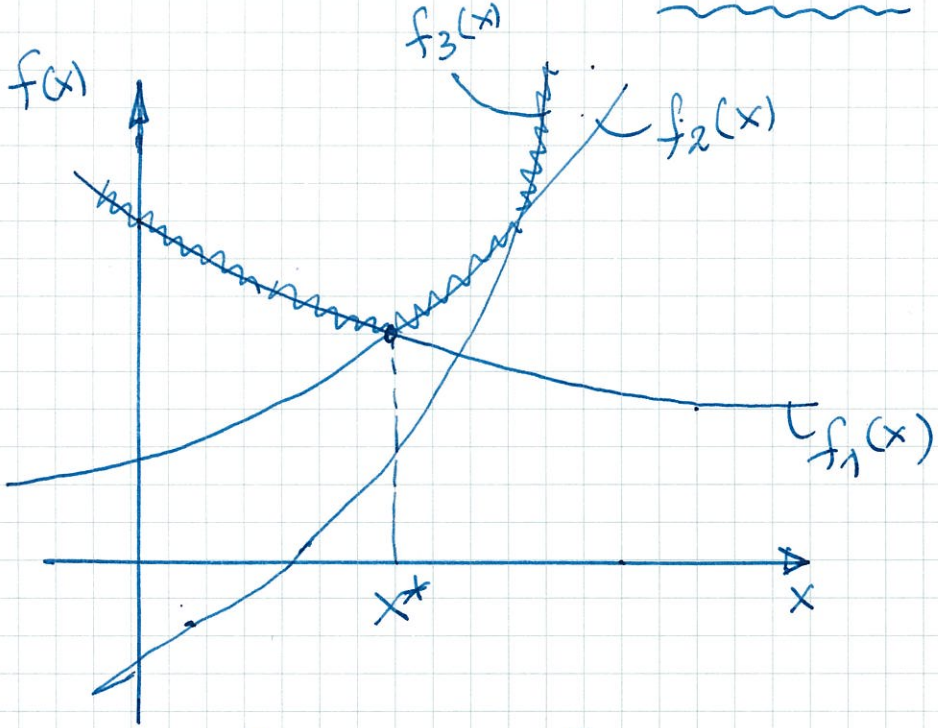
La posizione ottima è x^*

$$x^* = \arg\min_{x \in [a,b]} \text{dist}(x)$$

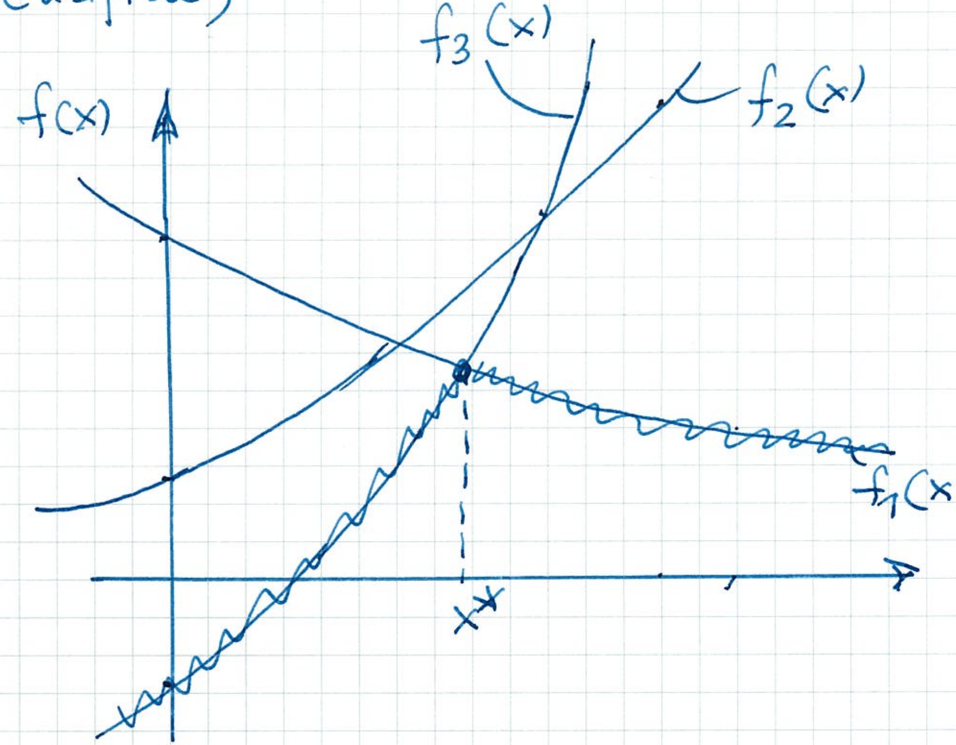
$$= \arg\min_{x \in [a,b]} \max(x-a,\ b-x)$$

$$x^* - a = b - x^* \implies x^* = \frac{a+b}{2}$$

Osservazione: gli strumenti dell'analisi classica non consentono di caratterizzare il punto di minimo (in esso la funzione $\max(x-a,\ b-x)$ non è derivabile (cuspide))



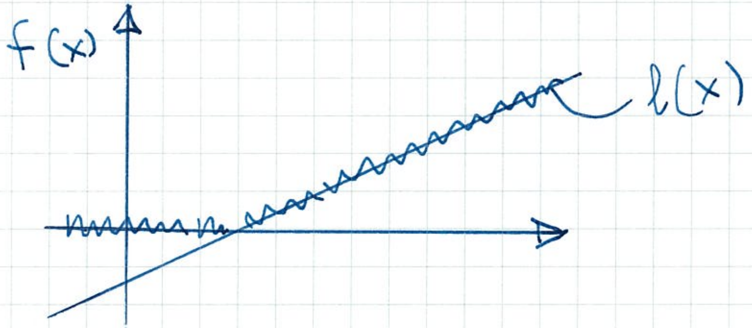
$$f(x) = \max_{i=1,\dots,3} f_i(x)$$



$$f(x) = \min_{i=1,\dots,3} f_i(x)$$

$x^* = \arg\min_x f(x) = \arg\min_x \max_{1\le i\le 3} f_i(x)$; $x^* = \arg\max_x f(x) = \arg\max_x \min_{1\le i\le 3} f_i(x)$

$f(x) = \max(0, \ell(x))$ con $\ell(x)$ funzione lineare



$f(x)$ è una funzione lineare a tratti e non derivabile ovunque

Nota. $f(x) \geq 0 \;\; \forall x$

$\Rightarrow \quad \min f(x) \geq 0$

Funzione di errore complessiva nel problema di classificazione

$$f(w, \gamma) = \sum_{i=1}^{m} \underbrace{\max\left(0, a^{(i)^T} w - \gamma + 1\right)}_{e_i(w,\gamma)} + \sum_{\ell=1}^{k} \underbrace{\max\left(0, -b^{(\ell)^T} w + \gamma + 1\right)}_{e_\ell(w,\gamma)}$$

Risolvere il problema $\min\limits_{w \in \mathbb{R}^n, \gamma \in \mathbb{R}} f(w, \gamma)$ e calcolare la soluzione ottima (w^*, γ^*)

Se $f(w^*, \gamma^*) > 0 \Rightarrow$ Non esiste un iperpiano di separazione e $H(w^*, \gamma^*)$ è un quasi-separatore

Se $f(w^*, \gamma^*) = 0$, allora tutti i punti sia di A che di B sono stati ben classificati e $H(w^*, \gamma^*)$ è l'iperpiano cercato.

$H(w^*, \gamma^*)$ è un classificatore: se perviene un nuovo campione $c \in \mathbb{R}^n$, se è $c^T w^* \leq \gamma^*$ $\Rightarrow$ c viene classificato di classe A, altrimenti di classe B.

Definendo $\xi_i = \max\left(0, a^{(i)T} w - \gamma + 1\right)$, $i = 1, \dots, m$

e $\zeta_l = \max\left(0, -b^{(l)T} w + \gamma + 1\right)$, $l = 1, \dots, k$

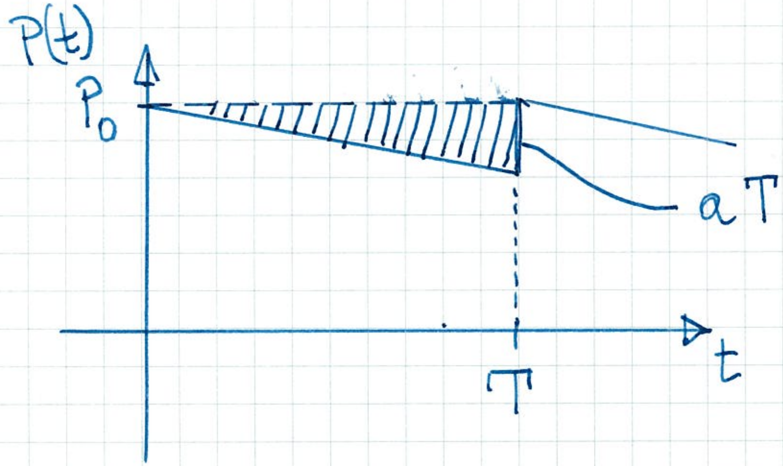
il problema di calcolare $\min\limits_{w, \gamma} f(w, \gamma)$ diventa

$$\begin{cases} \min\left(\sum_{i=1}^{m} \xi_i + \sum_{l=1}^{k} \zeta_l\right) \\ \xi_i \geq 0, \quad i = 1, \dots, m \\ \xi_i \geq a^{(i)T} w - \gamma + 1, \quad i = 1, \dots, m \\ \zeta_l \geq 0, \quad l = 1, \dots, k \\ \zeta_l \geq -b^{(l)T} w + \gamma + 1, \quad l = 1, \dots, k \end{cases}$$

Un problema di Programmazione Lineare

Un problema di gestione di un impianto turbogas per la produzione di energia elettrica

L'impianto è dotato di filtri che vanno di tanto in tanto sostituiti, pena il degrado delle prestazioni. Ogni quanto tempo è conveniente sostituirli?



$P(t)$ potenza erogata al tempo t (Kw)

P_0: potenza nominale erogata con filtri nuovi

T: intervallo di tempo tra due successivi rinnovi

A : costo dei filtri (€)

C : prezzo di vendita (€/Kwh) energia

Legge di degrado della potenza : $P(t) = P_0 - at$

$a > 0$, (Kw/t)

L'area tratteggiata, essendo ottenuta per prodotto Potenza × Tempo, rappresenta l'energia non prodotta (e quindi non venduta) per effetto del deterioramento dei filtri

Costi sostenuti all'interno di un ciclo:

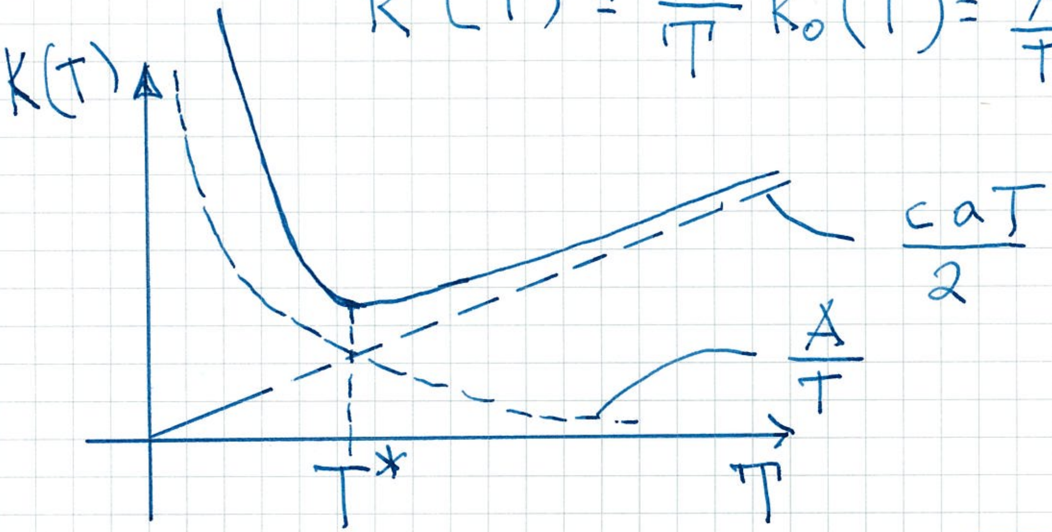
$$K_0(T) = \underset{\substack{\uparrow \\ \text{acquisto} \\ \text{filtri}}}{A} + \underset{\substack{\uparrow \\ \text{Costo KWh}}}{c} \cdot \text{Area}$$

Energia "persa"

$$K_0(T) = A + c\,\frac{aT^2}{2}$$

Costo nell'unità di tempo $K(T) = K_0(T) \cdot \left(\begin{array}{c}\text{numero di cicli} \\ \text{nell'unità di tempo}\end{array}\right)$

$$K(T) = \frac{1}{T} K_0(T) = \frac{A}{T} + c\,\frac{aT}{2}$$



$K(T)$ è una funzione convessa (somma di due convesse)

$$\Downarrow$$

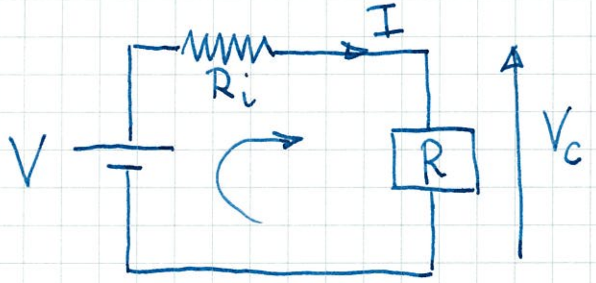
$K'(T) = 0$ fornisce la soluzione ottima

$$K'(T) = -\frac{A}{T^2} + \frac{ca}{2}\Bigg|_{T=T^*} = 0 \implies T^* = \sqrt{\frac{2A}{ca}}$$

Se cresce A, T si riduce
Se cresce c, T si incrementa

Esercizio

Calcolo dell'adattamento della resistenza



Determinare la resistenza R del carico che consente di massimizzare il trasferimento di potenza

La potenza trasferita $P = RI^2 = R\left(\frac{V}{R_i+R}\right)^2 = \frac{RV^2}{(R_i+R)^2}$

$$\max_R \frac{RV^2}{(R_i+R)^2}$$

$$\frac{dP}{dR} = \frac{V^2(R_i+R)^2 - RV^2 2(R+R_i)}{(R+R_i)^4} = \frac{V^2R_i^2 - V^2R^2}{(R+R_i)^4}$$

$$\left.\frac{dP}{dR} = 0\right|_{R=R^*} \Rightarrow R^* = R_i$$

Provare, calcolando $\frac{d^2P}{dR^2}$, che R^* è proprio un massimo

Problemi di approssimazione. (Best fitting)

Assegnati dei punti di $\mathbb{R}^n$, $x^{(1)}, x^{(2)}, \dots x^{(k)}$ e dei corrispondenti valori numerici $y_1, y_2, \dots, y_k$ si vuole individuare una funzione $f: \mathbb{R}^n \to \mathbb{R}$ tale che $f(x^{(i)}) \simeq y_i$, $i = 1, \dots, k$

Questo è un problema di approssimazione. Se si richiede che $f(x^{(i)}) = y_i$, $\forall i$, il problema è detto di interpolazione.

Risoluzione in due step: 1 Identificazione modello; 2 Stima dei parametri

1 Scelta della tipologia di funzione che meglio potrebbe approssimare i dati (es. parabola, esponenziale, sinusoidale ----)

2 Calcolo dei parametri (es. ampiezza della sinusoide, esponente etc.)

Es. Si ipotizzi che la relazione tra y e x sia esponenziale:

$$y = a_1 e^{a_2 x} \qquad (x \text{ è qui uno scalare})$$

Si tratta di stimare i valori di a_1 e a_2 che forniscono la migliore approssimazione.

Identificare il modello equivale a definire una funzione $F(a_1, \dots, a_m, x)$ di cui vanno stimati i parametri $a_1, \dots, a_m$ (nell'esempio è $F(a_1, a_2, x) = a_1 e^{\bar{a}_2 x}$)

La stima avviene minimizzando una misura dell'errore nei punti campionati. Ad esempio

(1) $$\min_{a_1, a_2, \dots, a_m} \sum_{i=1}^{K} \left(y_i - F(a_1, a_2, \dots, a_m, x^{(i)}) \right)^2$$

nell'esempio $$(a_1^*, a_2^*) = \arg \min_{(a_1, a_2)} \sum_{i=1}^{K} \left(y_i - a_1 e^{a_2 x_i} \right)^2$$

Stima a minimi quadrati

Alternative: a) $$\min_{a_1, \dots, a_m} \sum_{i=1}^{K} \left| y_i - F(a_1, a_2, \dots, a_m, x^{(i)}) \right|$$

b) $$\min_{a_1, \dots, a_m} \max_{1 \le i \le K} \left| y_i - F(a_1, \dots, a_m, x^{(i)}) \right|$$

a) Approssimazione L_1 ; b) Approssimazione di Chebishëf