Machine Learning: un nuovo approccio al Data Mining

Introduzione al corso

Fabio Mardero

9 ottobre 2019

edulife
apprendere per crescere insieme

t2i

Fabio Mardero

Data scientist e addetto valutazioni riserve danni IAS presso Cattolica Assicurazioni

Collaboratore e formatore all'interno del gruppo di studio "Tarallucci, Vino e Machine Learning"

Laureato in fisica e in scienze statistiche e attuariali

fabio.mardero@gmail.com

Io?

Non sono un programmatore!

Vedo la programmazione come uno strumento per mettere a terra idee e analisi sui dati

Obiettivi del corso

1

Costruire delle solide basi sulla materia

1. *"Assaggiare" tutti gli aspetti del Machine Learning, da quelli più tecnici a quelli prettamente legati al business*
2. *Pochi algoritmi ma molto utilizzati e facilmente applicabili*

Obiettivi del corso

2

Sviluppare un processo logico e universale che consenta di approcciare con successo:

1. *qualsiasi problema per il quale è possibile applicare il Machine Learning (applicazione)*
2. *qualsiasi aspetto teorico e pratico della materia e ogni sua evoluzione (apprendimento)*

Obiettivi del corso

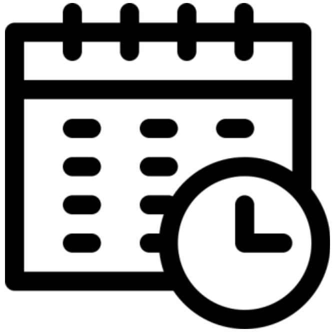
3

Costruire con le proprie mani e con il proprio codice un piccolo case study che possa mimare una soluzione in miniatura ad un problema aziendale/di interesse personale

Organizzazione del corso

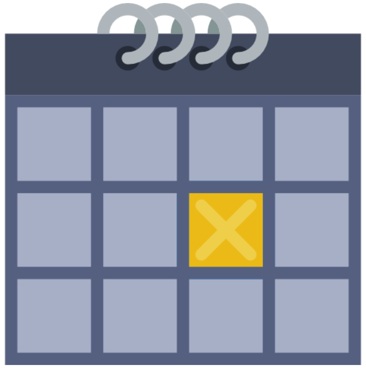
10 incontri, dal 9 ottobre al 11 novembre

Lunedì e mercoledì dalle 18.30 alle 21.30

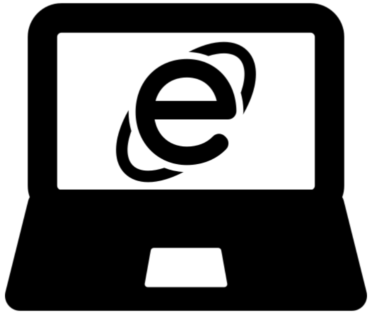


Struttura del corso

1. *Introduzione al Machine Learning, alla teoria e agli strumenti*
   [2 incontri]
2. *Casi pratici volti all'apprendimento di tecniche di ML*
   [6 incontri]
3. *Contestualizzazione della tecnologia in azienda*
   [2 incontri]



Materiale necessario per seguire il corso



1. Un PC, con la possibilità di installare alcuni programmi (~ 1 GB libero)
2. Un account google.com

Disclaimer

«Se la pizza è italiana allora il Machine Learning è anglosassone.»

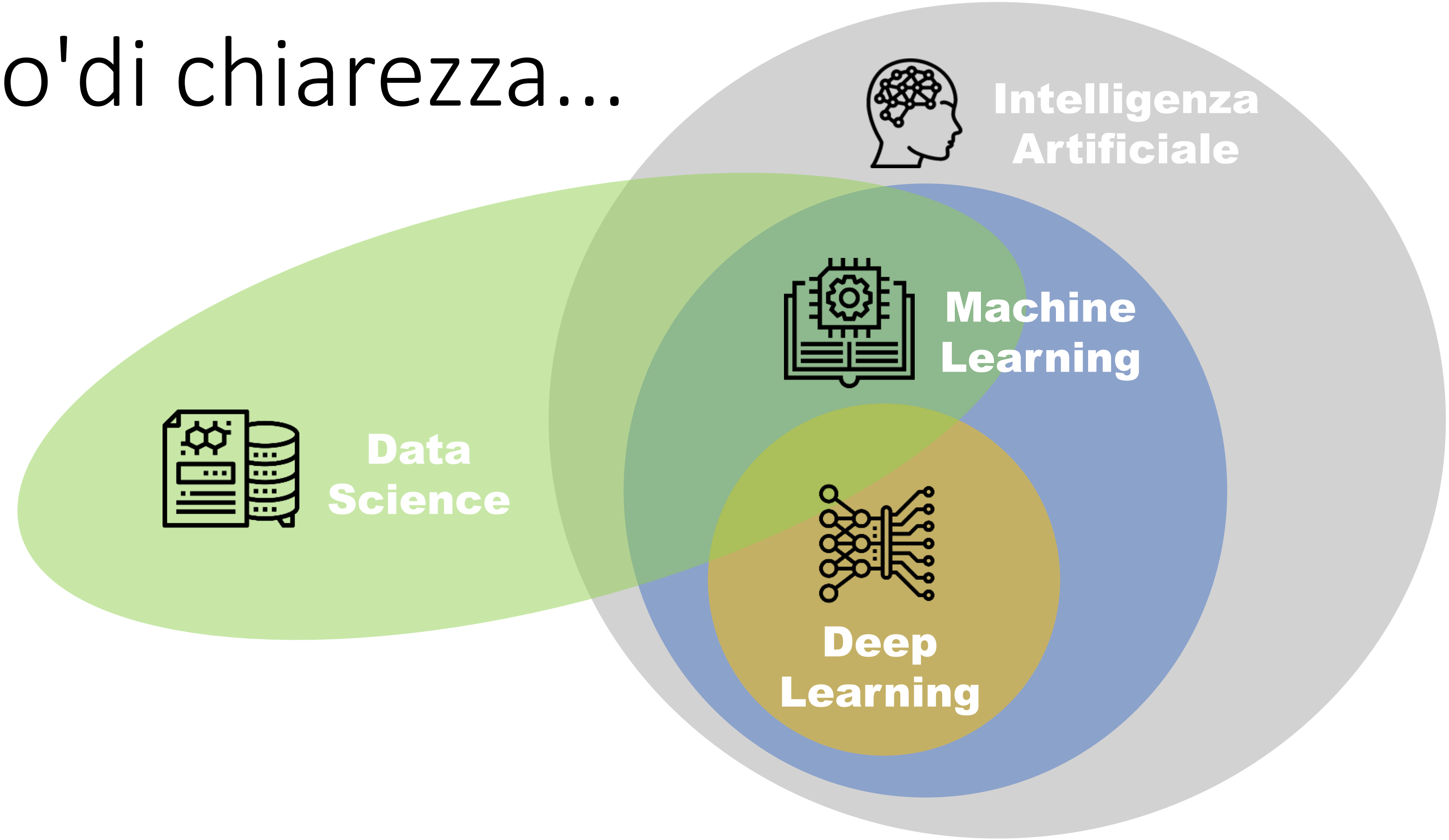
Le migliori fonti, la documentazione e le pubblicazioni sull’argomento sono in inglese

Ciò non significa che nel panorama italiano non esistano documenti/libri validi

Machine Learning: un nuovo approccio al Data Mining

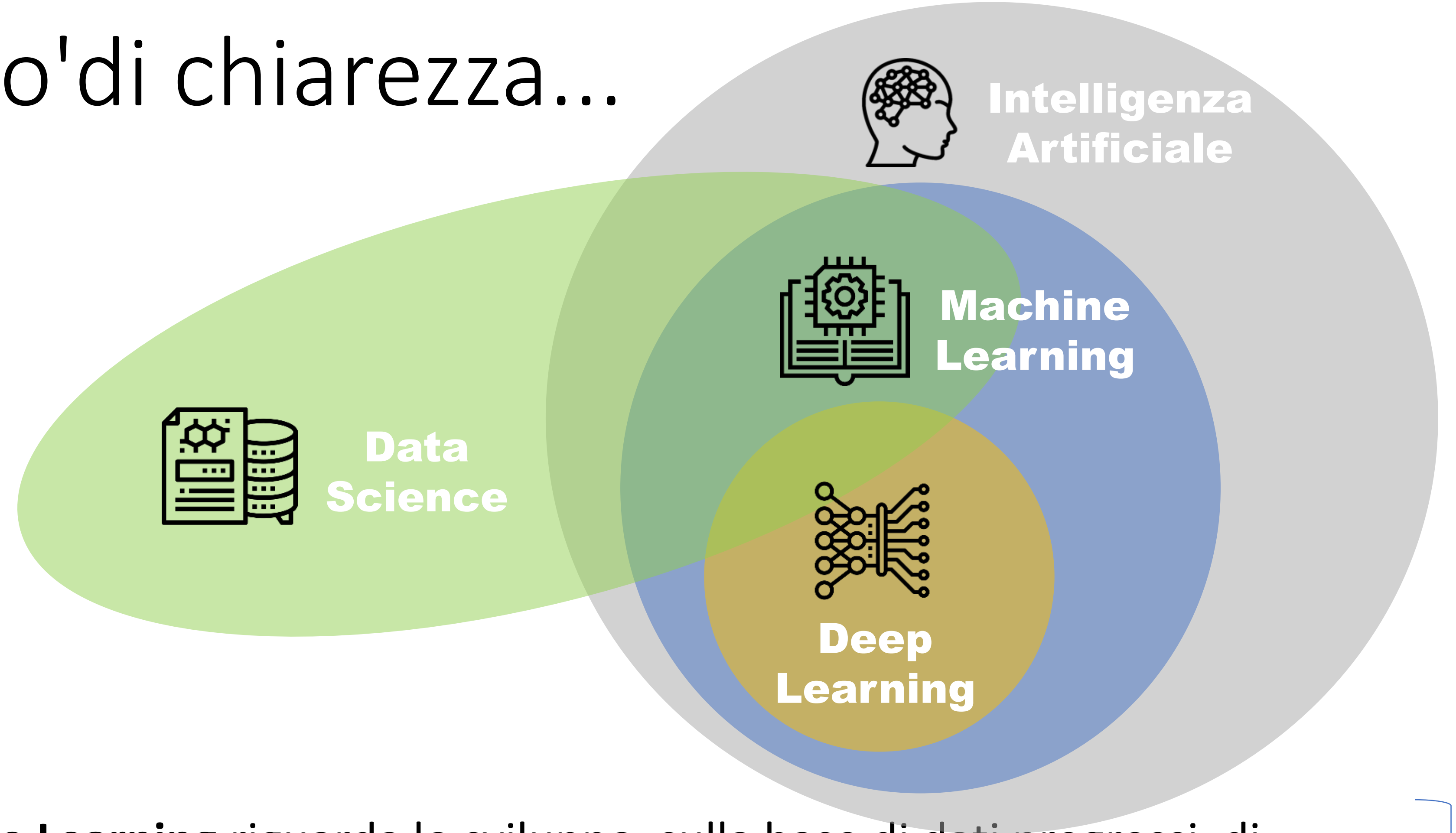
Introduzione e contestualizzazione

Un po'di chiarezza...



L'**Intelligenza Artificiale** consiste nella progettazione di sistemi hardware e software capaci di fornire all’elaboratore elettronico prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell’intelligenza umana

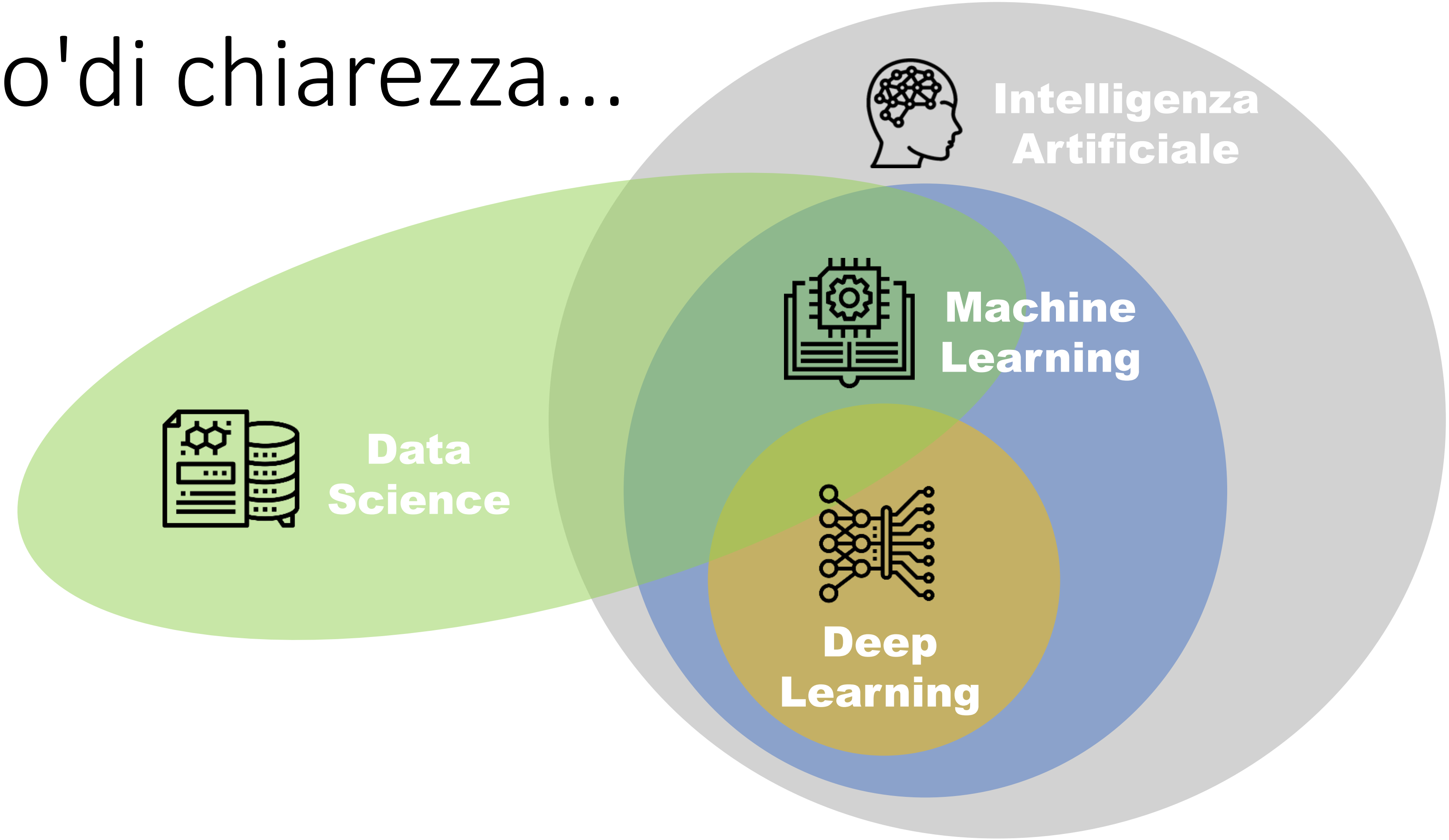
Un po'di chiarezza...



Il **Machine Learning** riguarda lo sviluppo, sulla base di dati pregressi, di modelli matematici in grado di fornire previsioni o su una variabile di interesse (noto A input prevedere B output) o sulle relazioni che intercorrono tra i dati stessi

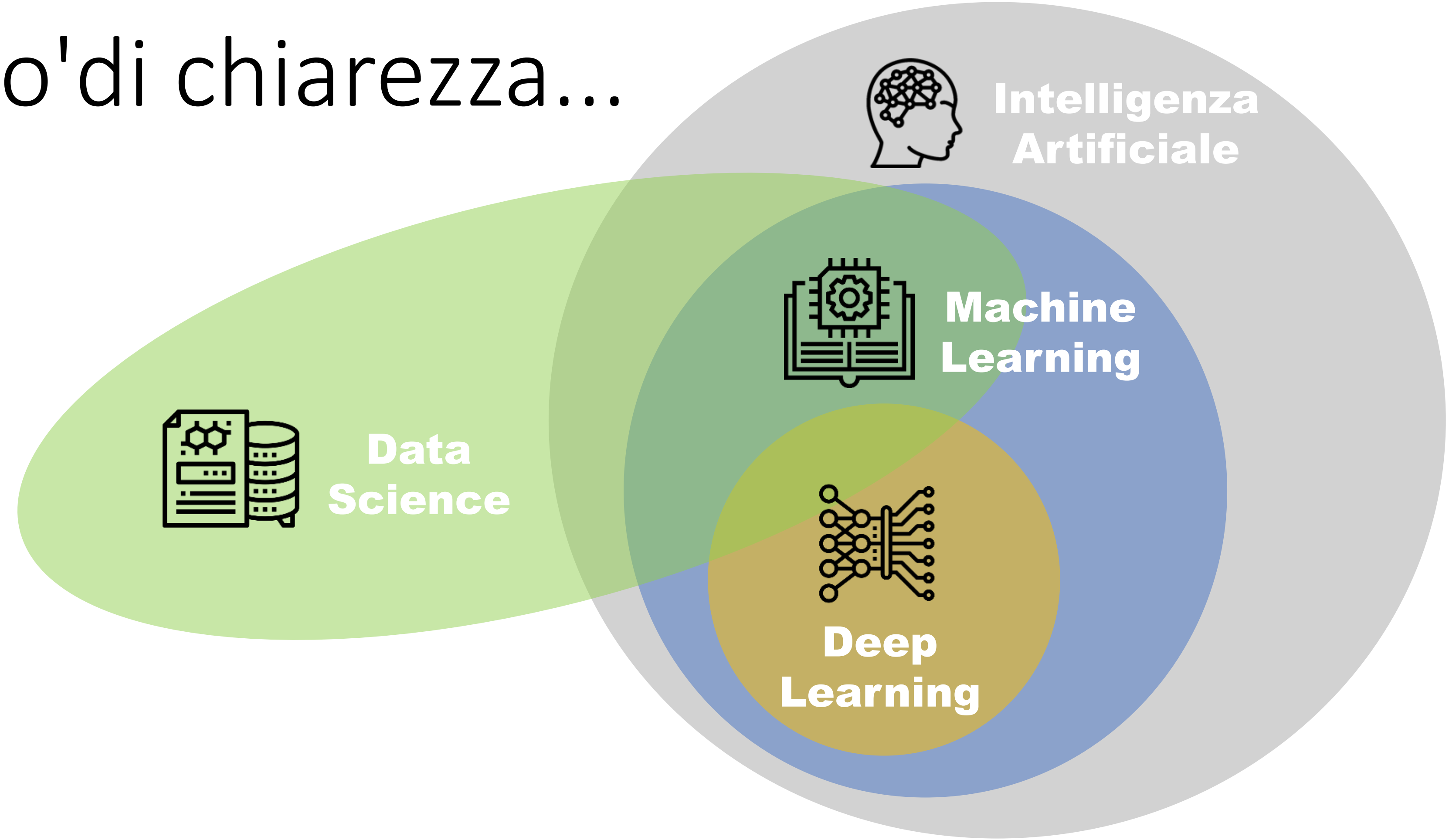
Output: Software

Un po'di chiarezza...



Il **Deep Learning** include specifici modelli di machine learning, le reti neurali, che si sono guadagnati una sotto-classificazione dato il loro particolare funzionamento e i notevoli risultati che hanno raggiunto in tempi recenti

Un po'di chiarezza...



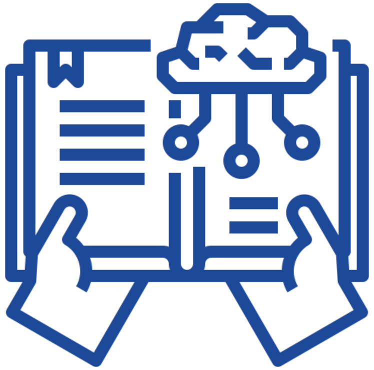
La **Data Science** consiste nell'estrazione dai dati di informazioni significative e/o utili al business

Output: Decisioni strategiche

Machine learning

Il significato

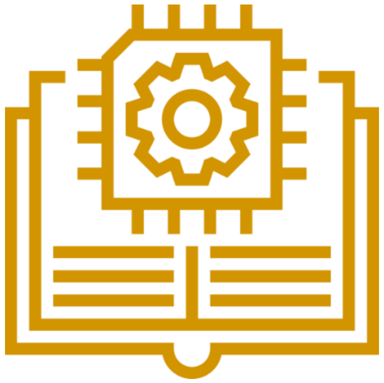
"Disciplina che utilizza metodi statistici per migliorare progressivamente, in seguito alla disponibilità di informazioni pregresse, le performance di un algoritmo in un dato compito."



Machine learning

Il significato

Un modello di machine learning è quindi in grado di "**apprendere**" dai **dati** allo scopo di eseguire, nel miglior modo possibile, un dato **compito**



Machine Learning
Elementi chiave

STEP 1

i dati

1. *qualsiasi formato, anche immagini, audio o testo*
2. *La quantità di dati necessaria per allenare un modello di Machine Learning dipende dall'algoritmo utilizzato e dalla difficoltà del compito*
3. *Potrebbero essere necessarie alcune pre-elaborazioni*

Machine Learning
Elementi chiave

STEP 2

l'apprendimento

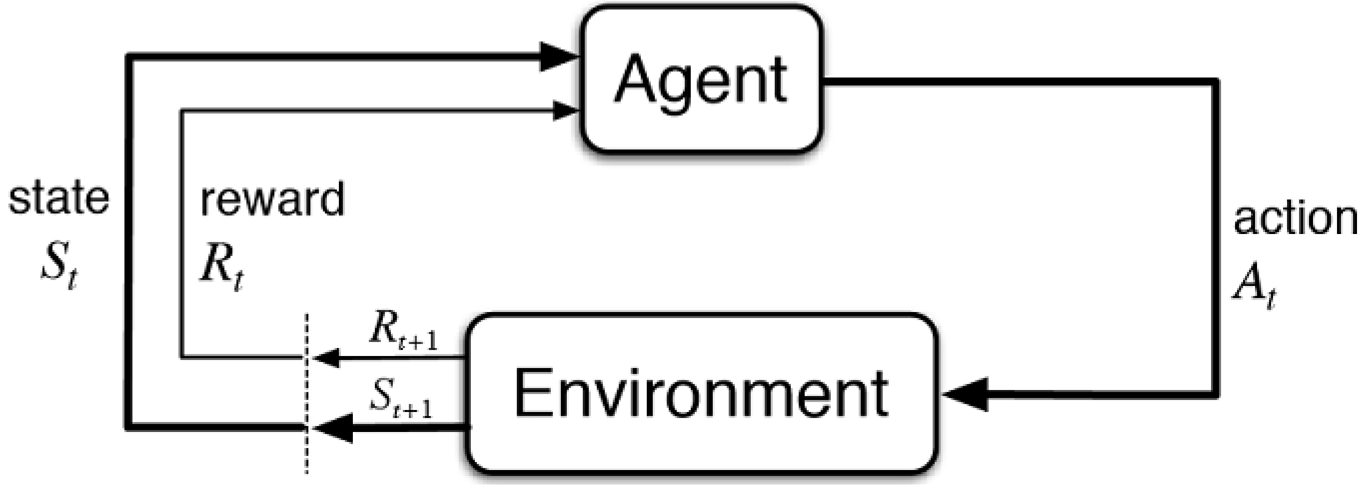
un modello apprende quando modifica la sua struttura, o i suoi parametri, per ridurre gli errori delle sue previsioni

1. *Apprendimento per rinforzo*
2. *Apprendimento supervisionato*
3. *Apprendimento non supervisionato*

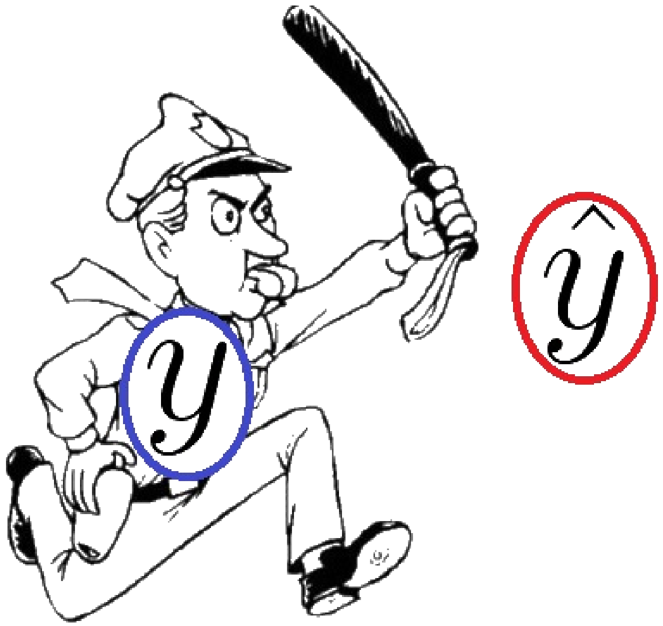
Machine Learning
Apprendimento per rinforzo

- *L'agente interagisce con l'environment e ogni sua azione modifica l'ambiente*
- *Il modello interagisce con il sistema e ogni sua previsione modifica il suo stato*

Nel tempo, non necessariamente ad ogni interazione con l'ambiente, l'agente riceve un feedback sul suo comportamento. Egli modifica quindi le sue future azioni, sulla base delle precedenti, tentando di massimizzare quelle che hanno portato a risultati positivi e minimizzando quelle risultate negative.
L'apprendimento dipende quindi da un sistema di rewards e punishments



Machine Learning
Apprendimento supervisionato



Il modello subisce l'ambiente

Nel caso dell'apprendimento supervisionato il modello mira a predire il comportamento di una o più variabili osservate rispetto alle altre

Indicata con $\hat{y}$ la previsione e con y il valore osservato, il modello apprende a minimizzare l'errore tra $\hat{y}$ e y. L'apprendimento è, informalmente, ``supervisionato'' dai valori di $\hat{y}$

Machine Learning
Apprendimento non supervisionato

Il modello subisce l'ambiente ma non è allenato per fornire una previsione

L'apprendimento non supervisionato prevede che l'algoritmo ricerchi strutture informative (*pattern*) tra i dati

Machine Learning
Elementi chiave

STEP 3

il compito

definisce su cosa il modello è allenato e con quali intenzioni, ad esempio fornire previsioni o trovare pattern di aggregazione dei dati

1. *Regressione*
2. *Classificazione*
3. *Clustering*
4. *…*

Machine Learning
Elementi chiave

STEP 3

il compito

Si riconoscono due casi:
- si individuano delle variabili più importanti, dette *variabili target/risposta*, rispetto alle altre, chiamate *variabili esplicative/covariate/features*
- tutte le variabili sono intese come significative (o potenzialmente tali)

Dato un insieme di dati, spetta all'osservatore decidere come intende interpretarli e se assegnare particolare importanza a qualcuna delle variabili disponibili

Machine Learning

Elementi chiave

STEP 3

il compito

Si riconoscono due casi:

1. **Compiti di regressione o classificazione**
   Mirando a fornire una previsione accurata delle variabili target, il modello spiega il fenomeno che genera $\hat{y}$

2. **Compiti legati all'estrazione di informazione dai dati e ad una loro rappresentazione, ad esempio il clustering**
   Ad esempio si individuano relazioni tra i dati contenuti in un dataset

Un esempio – Regressione Lineare

X	Y
+0.10	5.5
-0.40	4.1
-1.95	1.7
+1.17	8.1
-0.25	4.3
…	…

$\hat{Y}$
6.0
5.3
1.9
8.4
5.7
…

$$\hat{y} = mx + b$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Fit at iteration 0

"Linear Regression: The Easier Way"
Sagar Sharma

L'algoritmo ricerca i parametri che permettono di minimizzare la funzione d'errore calcolata sui dati noti

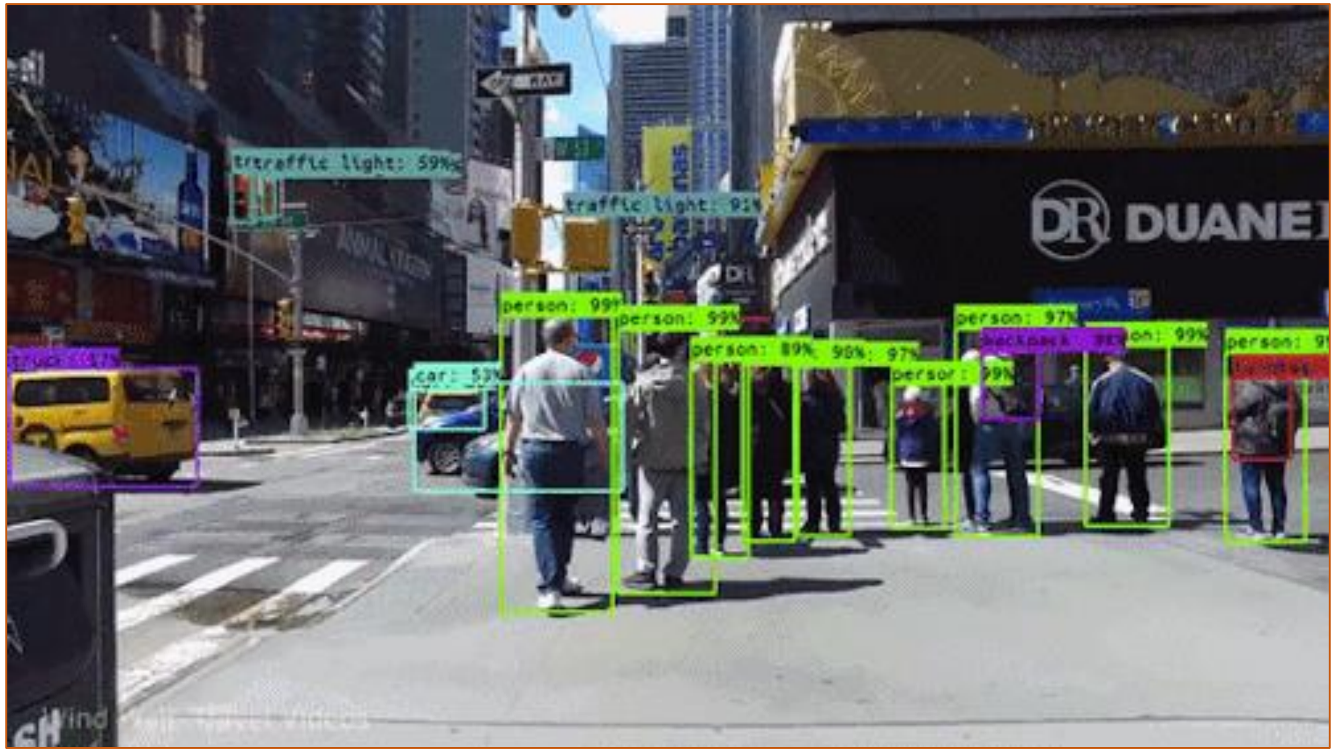
Capacità del Machine Learning

Un algoritmo di ML può:

- utilizzare sensori più precisi dei sensi umani
- comandare dispositivi in qualsiasi condizione con una precisione maggiore a quella umana
- agire più velocemente di un operatore
- lavorare in background anche 24h/24h (non si stanca mai)
- raggiungere per molti compiti una precisione che supera di gran lunga quella di un persona

Capacità del Machine Learning

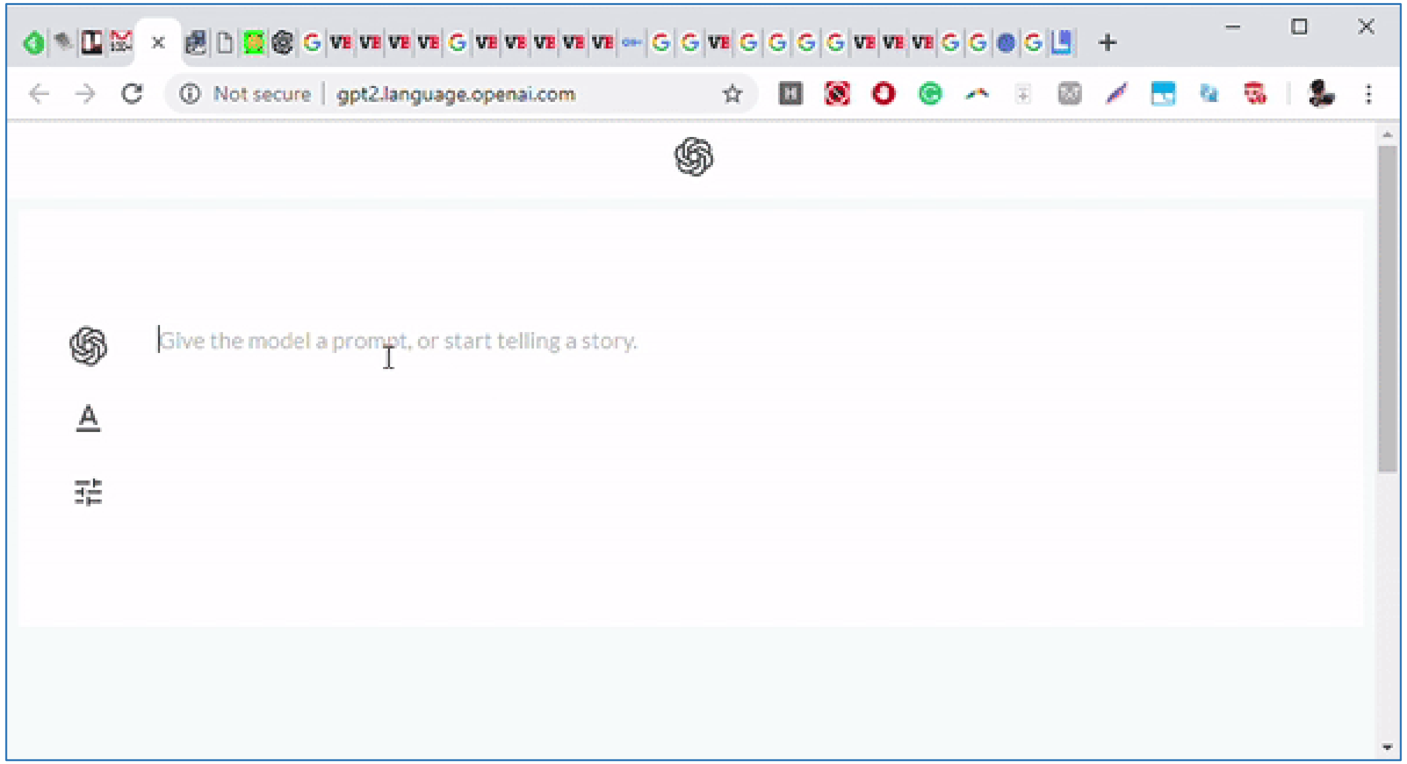
Rilevazione di oggetti



Capacità del Machine Learning

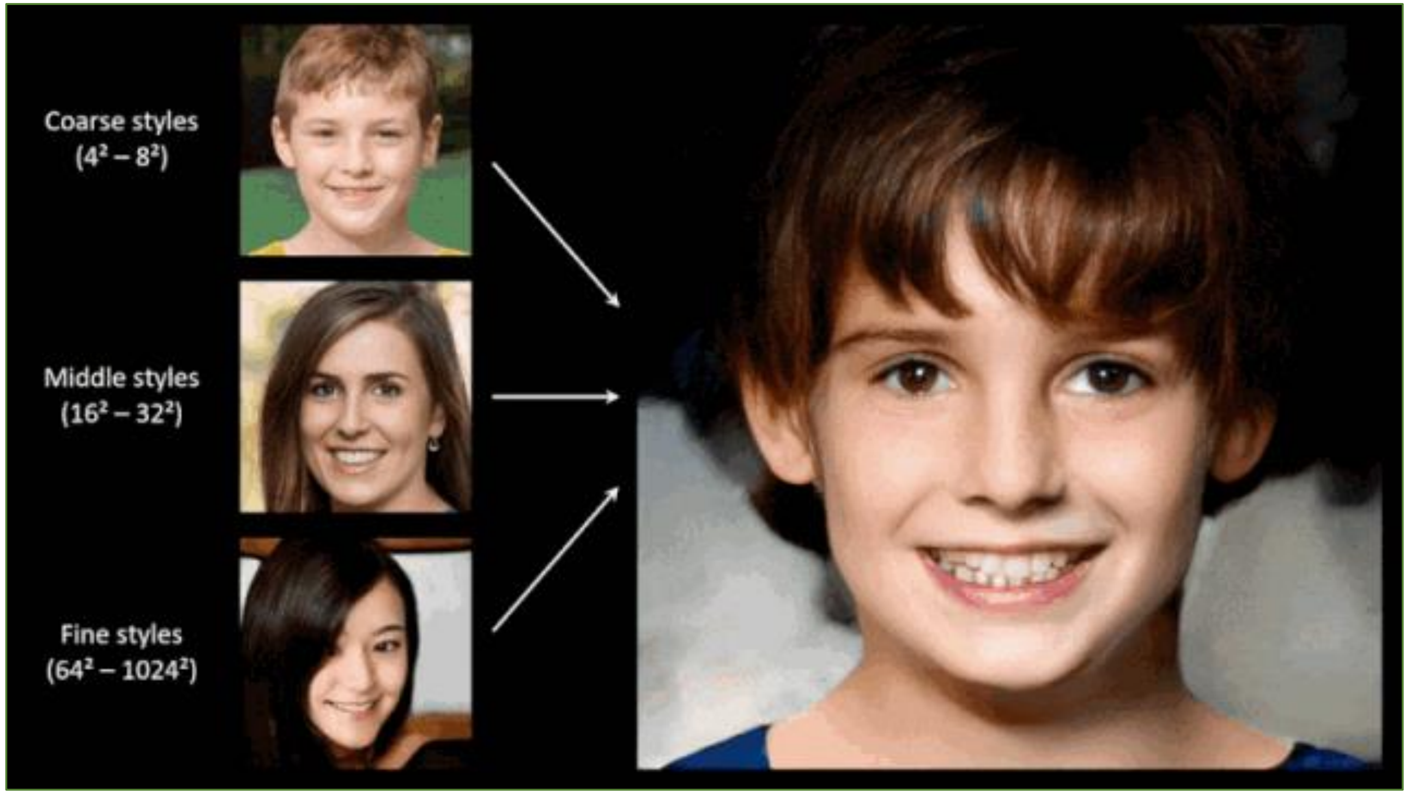
OpenAI

Generazione di testo



Capacità del Machine Learning

Generazione di volti



Capacità del Machine Learning
Alcune demo

Tramite webcam insegna all’algoritmo la comprensione dei gesti

https://teachablemachine.withgoogle.com/

Capacità del Machine Learning
Alcune demo

SEMANTRIS

Word association games powered by machine learning

Replica come l’algoritmo ha modellizzato la similarità tra diversi vocaboli

https://research.google.com/semantris/

Capacità del Machine Learning
Alcune demo

https://quickdraw.withgoogle.com/

Capacità del Machine Learning
Alcune demo

E ancora…

- Pix2Pix
  https://affinelayer.com/pixsrv/
- Replicatore dello stile di scrittura
  http://www.cs.toronto.edu/~graves/handwriting.html
- Motore di ricerca
  https://books.google.com/talktobooks/

Limiti del Machine Learning

Basta qualche secondo!

Limiti del Machine Learning

"Gli algoritmi di Machine Learning non sono in grado di svolgere compiti che richiederebbero per una comune persona più di qualche secondo per essere completati." Cit. Andrew Ag

- Potenziale esposizione ad attacchi informatici
- Richiedono domain knowledge da parte degli sviluppatori
- Potenziali bias e inefficienze

Big Data e Machine Learning

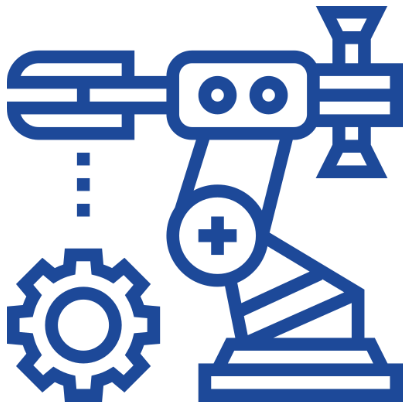
La massiccia raccolta di dati, anche tramite sistemi connessi e IoT, è estremamente legata all'implementazione di algoritmi di Machine Learning

È possibile sviluppare modelli di Machine Learning anche senza disporre di centinaia di GigaByte di dati (Big Data)

Si possono ottenere vantaggi economici per l'azienda già per la sola applicazione del Machine Learning, in particolare nelle aree dove prima non era previsto l'utilizzo di alcun algoritmo

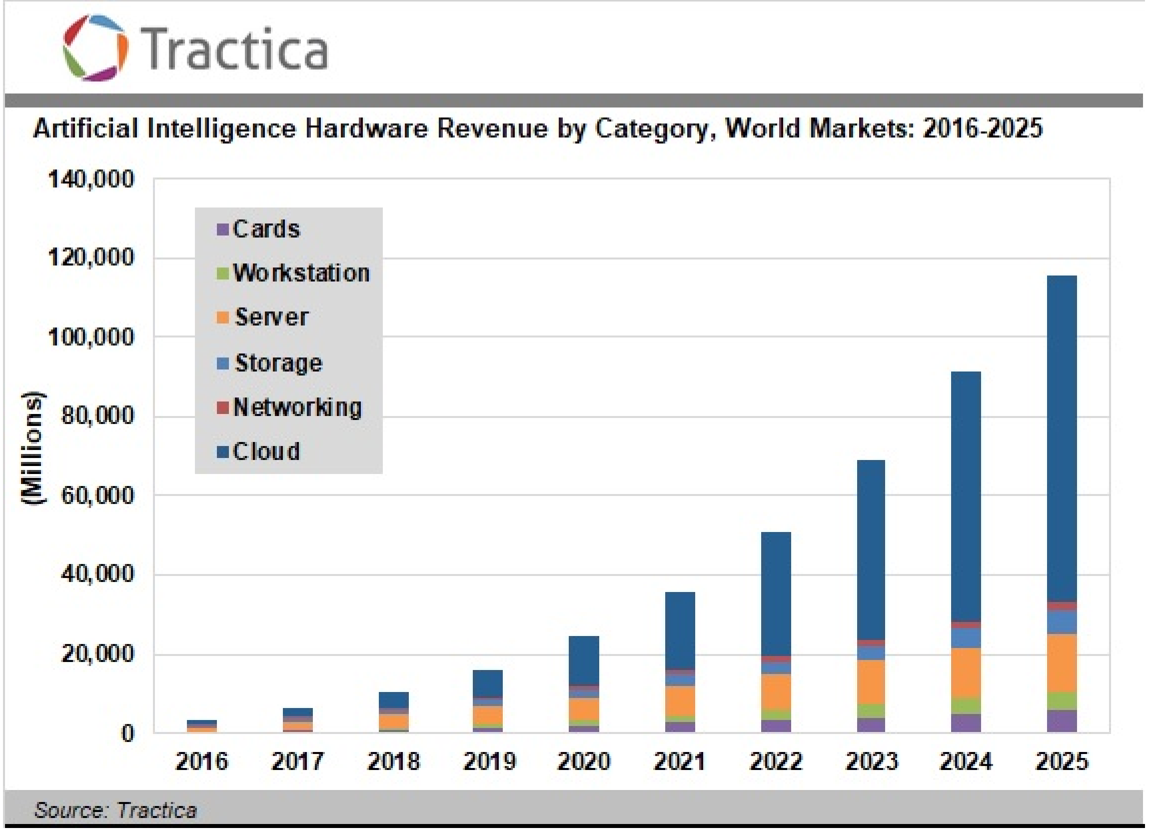
Industria 4.0 e Machine Learning

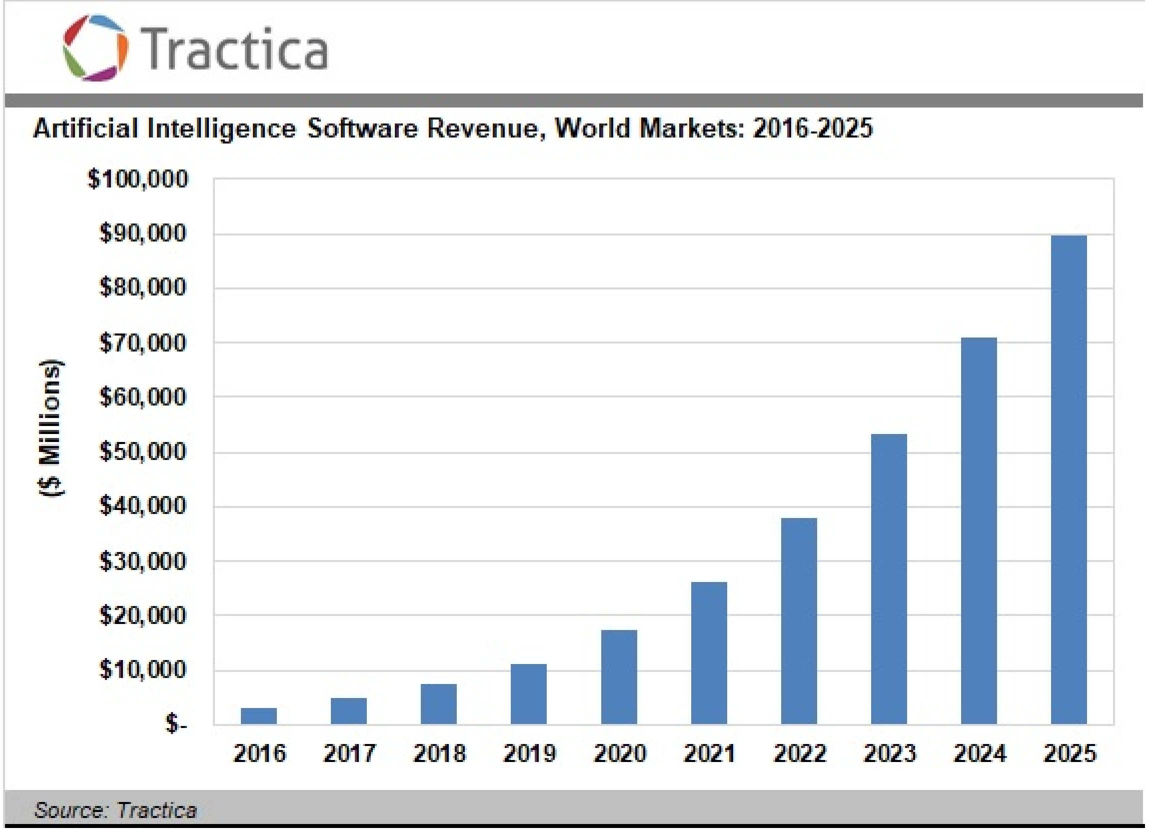
Il Machine Learning è strettamente legato all’Industria 4.0 in quanto consente all'impresa di implementare o evolvere i propri processi rendendoli più veloci ed efficienti, anche grazie all'automazione



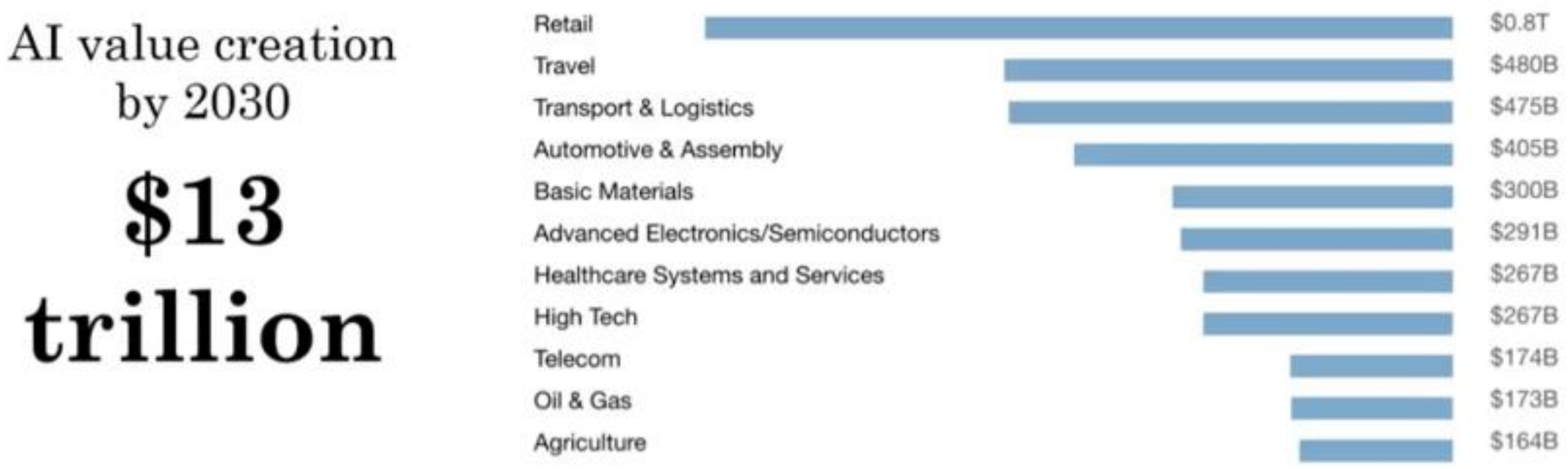
Il mercato del Machine Learning

Nel presente e in futuro





Il mercato del Machine Learning
In futuro



"Notes from the AI frontier: Modeling the impact of AI on the world economy"
September 2018, mckinsey.com

Il mercato del Machine Learning
Nel presente e in futuro

*Il mercato mondiale della **data annotation** è stato valutato nel 2018 pari a più di 316 milioni di USD e si stima che nel 2025 possa arrivare a valere oltre 1.6 miliardi di USD (*)*

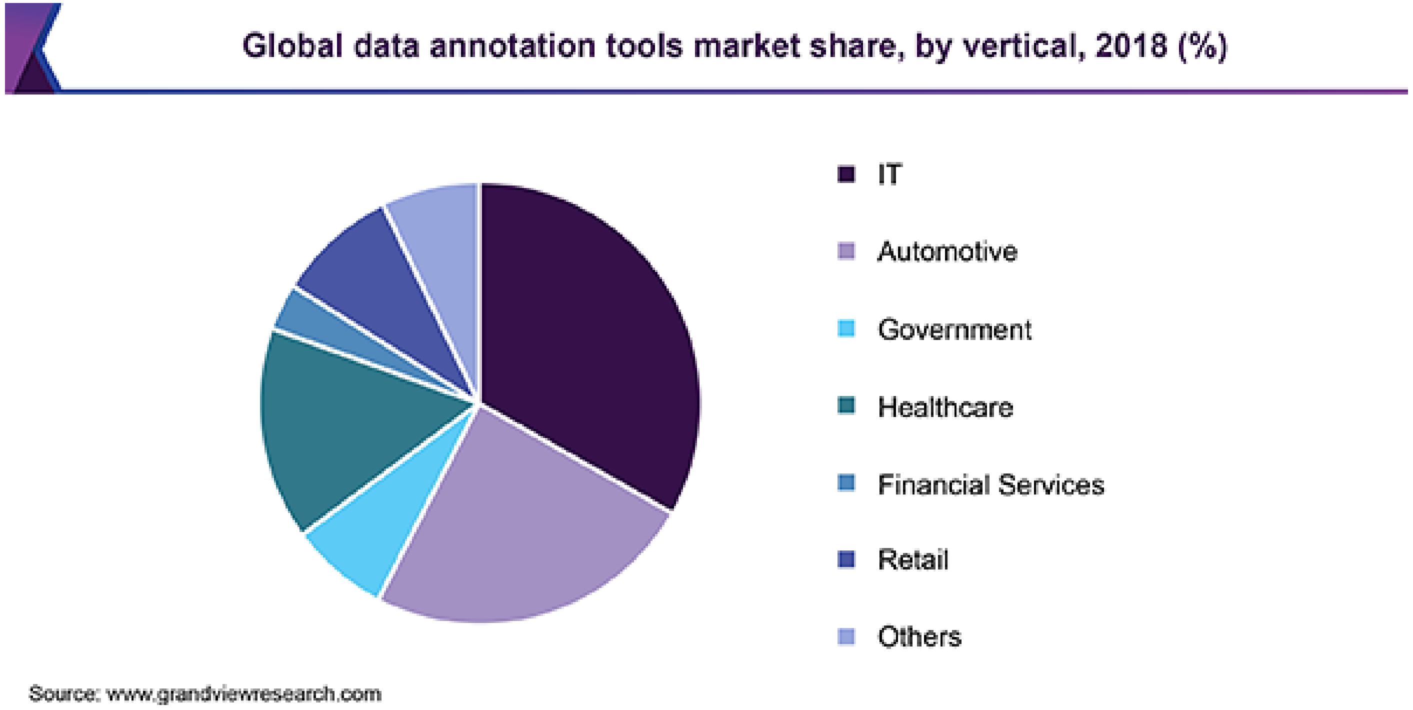
Secondo le previsioni, la crescita sarà trainata soprattutto dal mercato dell’automotive, quello retail, e i settori healthcare

Esempi di aziende interamente dedicate alla data annotation:

- **Testin**
- **Scale AI** (valutata circa 100 milioni di dollari US)
- **Mighty AI** (acquisita nel 06/2019 da Uber)

(*) “Data Annotation Tools Market Size, Share & Trends Analysis Report, 2019 – 2025“, grandviewresearch.com

Il mercato del Machine Learning
Nel presente

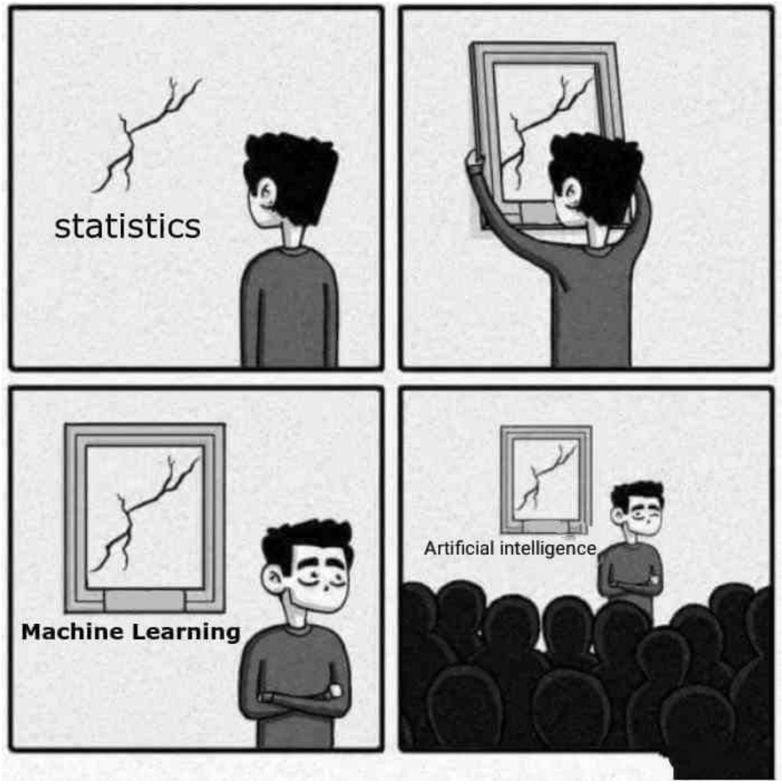


"Desperate Venezuelans are making money by training AI for self-driving cars"
Agosto 2019, technologyreview.com

Machine Learning: un nuovo approccio al Data Mining

Introduzione teorica

Introduzione teorica al Machine Learning
La statistica



Definiamo incerto/aleatorio tutto ciò che non possiamo verificare in maniera deterministica, per mancanza di informazioni o per proprietà intrinseca del fenomeno

Introduzione teorica al Machine Learning

Le due interpretazioni statistiche

Statistica frequentista

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

Statistica bayesiana

$$P(H_0|E) = \frac{P(E|H_0)P(H_0)}{P(E)}$$

Introduzione teorica al Machine Learning

La statistica

Protagoniste della statistica sono le variabili aleatorie cioè funzioni che da un evento causale restituiscono un valore «numerico»

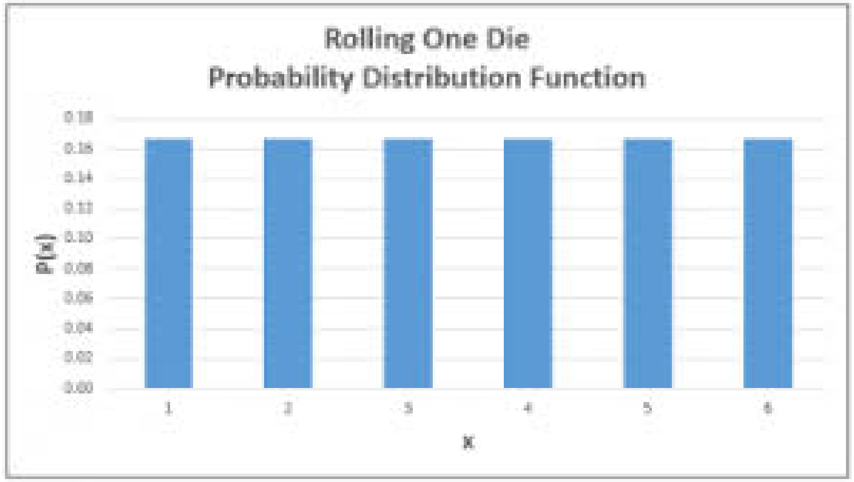
Spazio campionario

$$X\colon \Omega \to \mathbb{R}$$

Variabile aleatoria

Le variabili aleatorie sono caratterizzate da una distribuzione di probabilità (PDF) che descrive la relazione tra ogni possibile risultato della v.a. con la probabilità di ottenere tale valore

→ La PDF può essere discreta o continua



Introduzione teorica al Machine Learning

Esempi di PDF

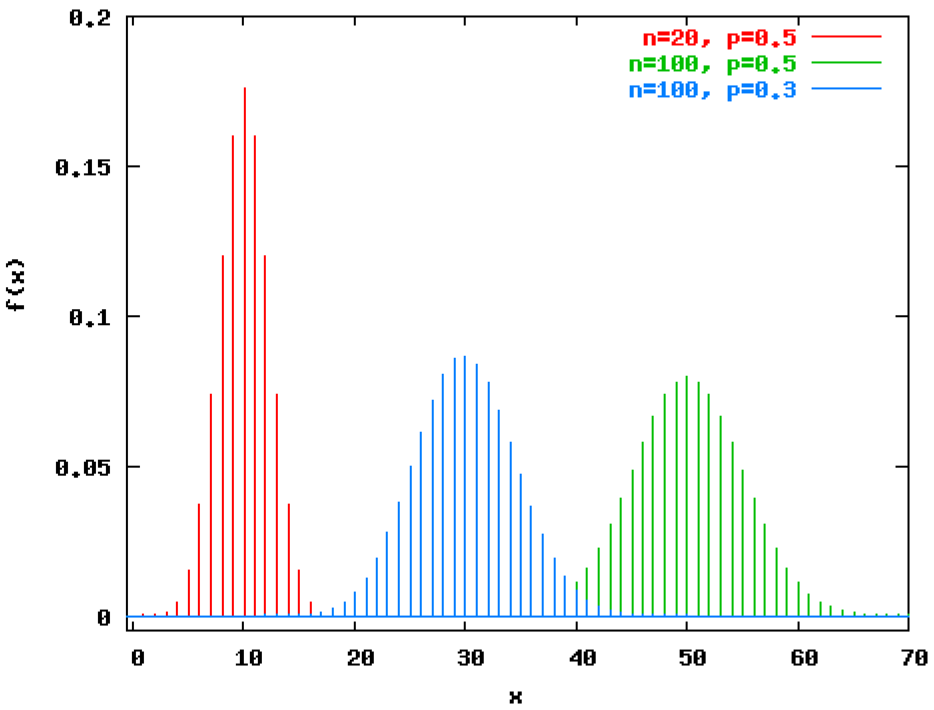
Variabile di Bernoulli

$$P(X=1)=p,$$
$$P(X=0)=q=1-p$$

***processo** (stocastico)*
di Bernoulli

Distribuzione Binomiale

$$S_n = X_1 + X_2 + \ldots + X_n$$



"Binomial – CPN Tools", *cpntools.org*

→ *i parametri della distribuzione sono n e p*

Introduzione teorica al Machine Learning

Esempi di PDF

"Galton Board", *store.fourpines.com*

Distribuzione Gaussiana/Normale

«Presa una distribuzione binomiale, assumendo

$$n \to +\infty$$
$$np \to +\infty$$

Allora si ottiene una distribuzione di probabilità continua detta gaussiana»

Funzione di densità di probabilità

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{con} \quad x \in \mathbb{R}$$

→ *i parametri della distribuzione sono μ e σ*

Introduzione teorica al Machine Learning

Esempi di PDF

Distribuzione Gaussiana/Normale

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

68,3% = P{ μ - 1,00 σ < X < μ + 1,00 σ }

95,0% = P{ μ - 1,96 σ < X < μ + 1,96 σ }

95,5% = P{ μ - 2,00 σ < X < μ + 2,00 σ }

99,0% = P{ μ - 2,58 σ < X < μ + 2,58 σ }

99,7% = P{ μ - 3,00 σ < X < μ + 3,00 σ }

→ *distribuzione a «coda corta»*

Introduzione teorica al Machine Learning

Esempi di PDF

Distribuzione Gaussiana/Normale

Teorema del limite centrale

Assumendo certe condizioni, la somma di n variabili aleatorie con media e varianza finite tende a una distribuzione normale al crescere di n

$$Y_n = \frac{\sum_{j=1}^{n} X_j - n\mu}{\sigma\sqrt{n}} \qquad Y_n \xrightarrow{D} Y \sim N(0,1)$$

Introduzione teorica al Machine Learning

Esempi di PDF

Processo stocastico di Poisson

$\{N(t), t \geq 0\}$ *con* $N(t)$ *numero di eventi che si verificano tra* 0 *e* t

Ipotesi:

- $N(0) = 0$
- $N(t_1) - N(s_1), \dots, N(t_n) - N(s_n)$ *indipendenti (per ogni intervallo disgiunto)*
- $P(N(t+h) - N(t) = 1) \approx \lambda h$ *e* $P(N(t+h) - N(t) > 1) \approx 0$ *con le approssimazioni sempre più vere tanto quanto* h *«più piccolo» (tendente a 0)*

$\Downarrow$

Distribuzione di Poisson

$$N(t+\tau) - N(t) \sim Poisson(\lambda\tau)$$

$$P(N(t+\tau) - N(t) = k) = \frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!}$$

→ *Descrive la probabilità che in un dato intervallo temporale si verifichino* k *eventi indipendenti*

Introduzione teorica al Machine Learning
Esempi di PDF

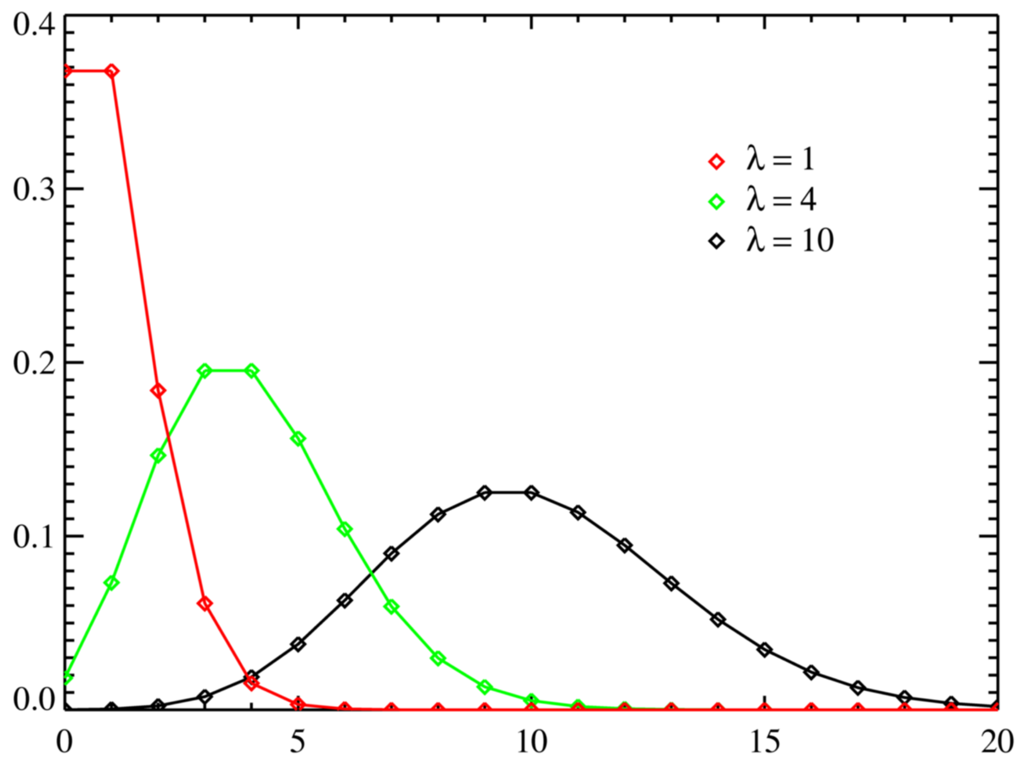
Distribuzione di Poisson

La probabilità di un evento in un piccolo intervallo di tempo è proporzionale alla durata dell'intervallo stesso:

$$P(N(t+h)-N(t)=1) \approx \lambda h$$

→ La costante di proporzionalità λ è detta intensità del processo

La probabilità che accada più di un evento in un piccolo intervallo di tempo è trascurabile



"Poisson Distribution / Poisson Curve: Simple Definition", *statisticshowto.datasciencecentral.com*

Introduzione teorica al Machine Learning

Proprietà di PDF

Molto spesso non è possibile conoscere completamente la distribuzione di probabilità della variabile aleatoria, ma si possono ricavare informazioni utili anche tramite indicatori sintetici che ne rilevino qualche notevole proprietà

Valore atteso

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Varianza

$$\sigma_X^2 = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$$

→ *σ è detta «deviazione standard»*

Introduzione teorica al Machine Learning

Proprietà di PDF

Per le distribuzioni di probabilità viste finora, quanto valgono le seguenti quantità?

Valore atteso

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Varianza

$$\sigma_X^2 = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$$

Introduzione teorica al Machine Learning

Inferenza statistica

Molto spesso, purtroppo…

se parametrico : $\{f_{\boldsymbol{\theta}}(x) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$

Spazio parametrico

Introduzione teorica al Machine Learning

Apprendimento supervisionato

Dataset: $\mathcal{D} = \{ (\boldsymbol{x}_i, y_i) \}_{i=1}^{N}$ con $\boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Variabile target descritta tramite un modello predittivo:

$$Y = \underbrace{f(\boldsymbol{x})}_{\text{Componente deterministica}} + \underbrace{\varepsilon}_{\text{Componente stocastica}}, \quad \boldsymbol{x} \in \mathcal{X}$$

Ipotesi induttiva

$$f : \mathcal{X} \longrightarrow \mathcal{Y}$$
$$\boldsymbol{x} \longmapsto f(\boldsymbol{x}) = \hat{y}$$

Modello parametrico

$$\mathcal{F} = \{ f(\boldsymbol{x}; \boldsymbol{\theta}) ; \ \boldsymbol{\theta} \in \Theta \}, \quad \text{per ogni } \boldsymbol{x} \in \mathcal{X}$$

Introduzione teorica al Machine Learning

Apprendimento supervisionato

Loss function:

$$
\begin{aligned}
L : \mathcal{Y} \times \mathcal{Y} &\longrightarrow \mathbb{R} \\
(y_1, y_2) &\longmapsto L(y_1, y_2)
\end{aligned}
$$

$$
\begin{aligned}
f : \mathcal{X} &\longrightarrow \mathcal{Y} \\
\boldsymbol{x} &\longmapsto f(\boldsymbol{x}) = \hat{y}
\end{aligned}
$$

$$\boxed{L(\hat{y}, y) = L\left(f(\boldsymbol{x}; \boldsymbol{\theta}), y\right), \quad \text{per ogni } (\boldsymbol{x}, y) \in \mathcal{D}}$$

Introduzione teorica al Machine Learning

Apprendimento supervisionato

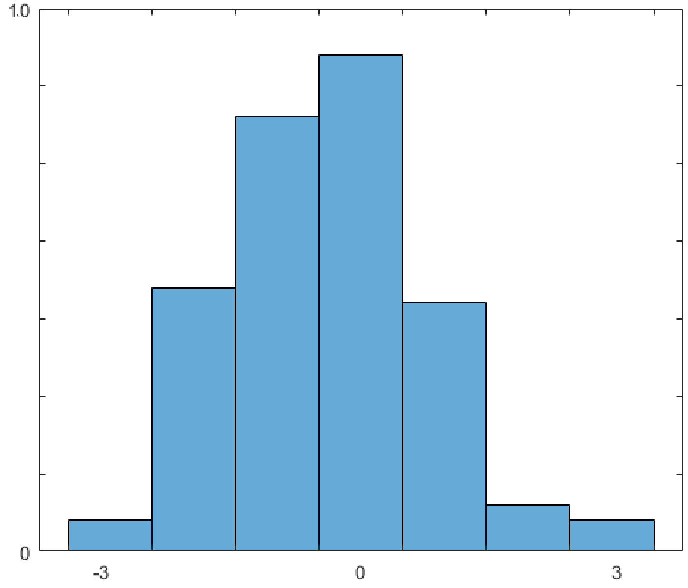
Risk function:

$$R(\boldsymbol{\theta}) = \mathbb{E}\left[L\left(f(\boldsymbol{X};\boldsymbol{\theta}), Y\right)\right]$$
$$= \int L\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\, dF(\boldsymbol{x}, y)$$

approssimando

$$F_{emp}(\boldsymbol{x}, y) = \frac{1}{N}\sum_{i=1}^{N} I_{(\boldsymbol{x}\leq\boldsymbol{x}_i, y\leq y_i)}$$

Distribuzione di probabilità empirica



$$R(\boldsymbol{\theta}) \approx R_{emp}(\mathcal{D}, \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} L\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i\right)$$

Introduzione teorica al Machine Learning

Apprendimento supervisionato

$$R(\boldsymbol{\theta}) \approx R_{emp}(\mathcal{D}, \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} L\left(f(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i\right)$$

Alcuni esempi:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Mean Squared Error

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|$$

Mean Absolute Error

Introduzione teorica al Machine Learning

Apprendimento supervisionato

Empirical Risk Minimization (ERM)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} R_{emp}(\mathcal{D}, \boldsymbol{\theta})$$

Modello parametrico

$$\mathcal{F} = \left\{ f\left(\boldsymbol{x}; \boldsymbol{\theta}\right);\ \boldsymbol{\theta} \in \Theta \right\}, \quad \text{per ogni } \boldsymbol{x} \in \mathcal{X}$$

$$R(\boldsymbol{\theta}) \approx R_{emp}(\mathcal{D}, \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} L\left(f(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i\right)$$

Introduzione teorica al Machine Learning

Apprendimento supervisionato

Empirical Risk Minimization (ERM)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} R_{emp}(\mathcal{D}, \boldsymbol{\theta})$$

Risoluzione del problema di ottimizzazione

$$\nabla_\theta = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_p}\right)^T$$

gradiente

$$0 = \nabla_{\boldsymbol{\theta}} R_{emp}(\mathcal{D}, \boldsymbol{\theta})$$

$$0 = \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} L\left(f(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i\right)$$

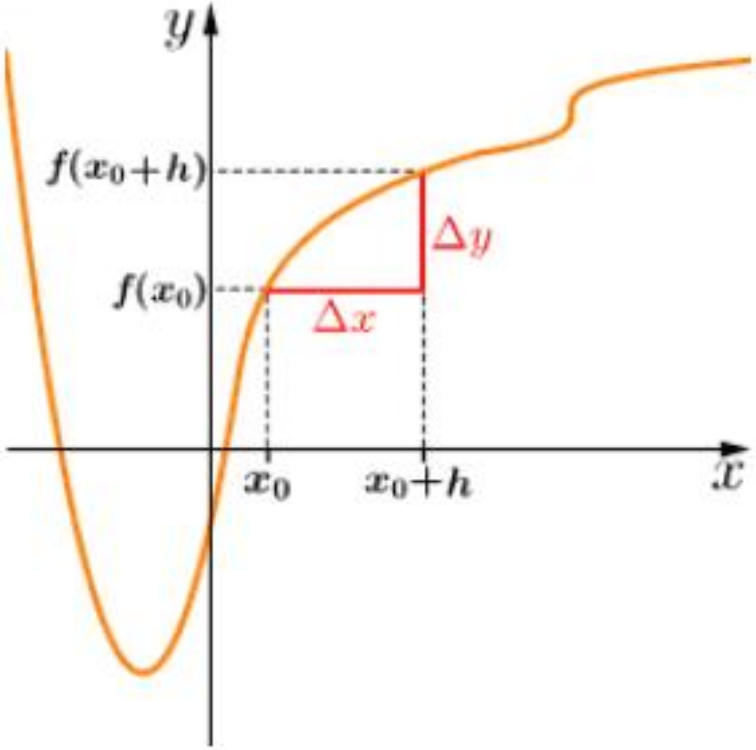
$$0 = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial L\left(t, y_i\right)}{\partial t} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i; \boldsymbol{\theta})$$

Introduzione teorica al Machine Learning

Definizione di rapporto incrementale

Consideriamo una funzione $f : \mathbb{R} \to \mathbb{R}$, $y = f(x)$ di variabile reale a valori reali. Definiamo il rapporto incrementale della funzione f nel punto x_0 nel modo seguente:

$$\frac{\Delta y}{\Delta x} := \frac{f(x_0 + h) - f(x_0)}{h}$$



dove il simbolo := nella formula indica che l'uguaglianza è una definizione.

Nella formula del rapporto incrementale è presente il rapporto tra la differenza delle ordinate $f(x_0+h)$, $f(x_0)$, ossia le ordinate corrispondenti alle ascisse x_0+h e x_0 mediante f, e la differenza delle relative ascisse x_0+h e x_0, che è evidentemente h.

Il rapporto che abbiamo indicato con

$$\frac{\Delta y}{\Delta x}$$

si chiama *rapporto incrementale*, e il nome si giustifica per il fatto che è un rapporto di differenze calcolate a partire da un incremento: h, per l'appunto. La lettera greca Δ (Delta) si usa solitamente in Matematica e in Fisica per indicare una variazione o differenza, il che giustifica la notazione $\Delta y/\Delta x$.

Introduzione teorica al Machine Learning

Derivata di una funzione in un punto

youmath.it

Consideriamo la solita funzione $y = f(x)$ ed un punto x_0 nel suo dominio. Ci sono diversi simboli usati per denotare la derivata di una funzione in un punto:

$$f'(x_0) \quad ; \quad \frac{df}{dx}(x_0) \quad ; \quad D(f(x))|_{x=x_0}$$

Tutti questi simboli si riconducono alla medesima definizione

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

In altri termini, la derivata di una funzione in un punto è il limite del rapporto incrementale al tendere dell'incremento h a zero.

Tutto qui? In effetti no, possiamo dare altre due definizioni. Chiamiamo derivata sinistra nel punto x_0 il limite del rapporto incrementale calcolato da sinistra

$$f'_-(x_0) = \lim_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h}$$

e diciamo derivata destra nel punto x_0 il limite del rapporto incrementale calcolato da destra

$$f'_+(x_0) = \lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h}$$

Introduzione teorica al Machine Learning

Regole di derivazione

$$(k \cdot f)'(x) = k \cdot f'(x) \quad k \in \mathbb{R}$$

$$(f \pm g)'(x) = f'(x) \pm g'(x)$$

$$(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$$

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{\left(g(x)\right)^2}$$

Chain rule: $(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$

Constant Rule : $\frac{d}{dx}[c] = 0$

Power Rule : $\frac{d}{dx}\left[x^n\right] = n \cdot x^{n-1}$

mathspadilla.com/macsII/Unit6-Derivatives/derivation_rules.html

Introduzione teorica al Machine Learning

Teorema di Fermat (1 / 3)

Enunciato e dimostrazione del teorema di Fermat

youmath.it

Sia $y = f(x)$ una funzione con dominio $Dom(f) \subseteq \mathbb{R}$. Se $x_0 \in Dom(f)$ è un punto estremante per f, e la funzione è derivabile in quel punto, allora si ha che

$$f'(x_0) = 0$$

Dimostrazione

Prima di tutto osserviamo che per ipotesi f(x) è derivabile nel punto x_0, dunque vale la condizione

$$\lim_{x \to x_0^-} f'(x) = \lim_{x \to x_0^+} f'(x)$$

Dimostriamo il teorema nel caso in cui x_0 sia un punto di massimo relativo; il caso in cui è un punto di minimo si dimostra in maniera del tutto analoga.

Poiché x_0 è un punto di massimo relativo, dato un incremento h vale

$$f(x_0 + h) - f(x_0) \leq 0$$

Introduzione teorica al Machine Learning

Teorema di Fermat (2 / 3)

Infatti se x_0 è un punto di massimo spostandoci sull'asse delle ascisse troveremo, localmente, valori della funzione più piccoli di $f(x_0)$.

Dividiamo la disuguaglianza per h. Otteniamo:

youmath.it

- se h è positivo

$$\frac{f(x_0+h)-f(x_0)}{h} \leq 0$$

- se h è negativo

$$\frac{f(x_0+h)-f(x_0)}{h} \geq 0$$

Ora: se passiamo al limite per $h \to 0$ in entrambe le disuguaglianze, otteniamo

$$\lim_{h \to 0} \frac{f(x_0+h)-f(x_0)}{h} \leq 0 \quad (h > 0)$$

$$\lim_{h \to 0} \frac{f(x_0+h)-f(x_0)}{h} \geq 0 \quad (h < 0)$$

Introduzione teorica al Machine Learning

Teorema di Fermat (3 / 3)

I due limiti sono rispettivamente limite destro e limite sinistro della derivata prima,

$$f'_+(x_0) = f'_-(x_0)$$

Per l'ipotesi di derivabilità di f in x_0 in due limiti devono coincidere, quindi essendo

$$f'_+(x_0) \leq 0 \quad \text{e} \quad f'_-(x_0) \geq 0$$

l'unico caso possibile è

$$f'_+(x_0) = 0 = f'_-(x_0)$$

Ossia

youmath.it

$$f'(x_0) = 0$$

Introduzione teorica al Machine Learning

Derivata parziale

Partiamo dal concetto di derivata parziale del primo ordine. Come sempre prendiamo una funzione f definita in un aperto non vuoto $D \subset \mathbb{R}^2$ e un punto $(x_0, y_0) \in D$ diremo che la funzione è derivabile parzialmente rispetto alla variabile x nel punto (x_0, y_0) se esiste finito il limite in una variabile:

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

youmath.it

diremo invece che essa è derivabile parzialmente rispetto ad y nel punto (x_0, y_0) se esiste finito il limite:

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{k \to 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}$$

Tendenzialmente si utilizzano tre simboli per indicare la derivata parziale:

$\frac{\partial f}{\partial x}(x, y) \quad f_x(x, y) \quad \partial_x f(x, y)$ indicano la *derivata parziale prima rispetto alla variabile x della funzione f;*

$\frac{\partial f}{\partial y}(x, y) \quad f_y(x, y) \quad \partial_y f(x, y)$ indicano la *derivata parziale prima rispetto alla variabile y della funzione f;*

Introduzione teorica al Machine Learning

Problema di ottimizzazione

Un esempio…

$$\mathcal{R}(\theta_1, \theta_2) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg\min_{(\theta_1,\theta_2)} \mathcal{R}(\theta_1, \theta_2)$$

problema di ottimizzazione

$$\begin{cases} \dfrac{\partial \mathcal{R}}{\partial \theta_1}(\hat{\theta}_1, \hat{\theta}_2) = 0 \\ \dfrac{\partial \mathcal{R}}{\partial \theta_2}(\hat{\theta}_1, \hat{\theta}_2) = 0 \end{cases}$$

Introduzione teorica al Machine Learning

Problema di ottimizzazione

Un esempio…

$$\frac{\partial \mathcal{R}}{\partial \theta_1}(\hat{\theta}_1, \hat{\theta}_2) = 0$$

$$\mathcal{R}(\theta_1, \theta_2) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\frac{\partial}{\partial \theta_1}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\Big)^2\right] = 0$$

Introduzione teorica al Machine Learning

Problema di ottimizzazione

$$\frac{d}{dx}A(B(x))^2 = 2A(B(x))\frac{d}{dx}\left[A(B(x))\right]$$

Introduzione teorica al Machine Learning

Problema di ottimizzazione

Un esempio…

$$0 = \frac{\partial}{\partial \theta_1}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\right)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \theta_1}\left[\left(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\right)^2\right]$$

$$0 = \frac{2}{N}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\right)\frac{\partial}{\partial \theta_1}\left(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\right)$$

Introduzione teorica al Machine Learning

Problema di ottimizzazione

Un esempio…

$$0 = \frac{2}{N}\sum_{i=1}^{N}\Big(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\Big)\frac{\partial}{\partial\theta_1}\Big(y_i - f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2)\Big)$$

Regressione lineare: $f(\mathbf{x}_i, \hat{\theta}_1, \hat{\theta}_2) = \hat{\theta}_1\mathbf{x}_i + \hat{\theta}_2$

$$\Downarrow$$

$$0 = -\frac{2}{N}\sum_{i=1}^{N}\mathbf{x}_i\left(y_i - \left(\hat{\theta}_1\mathbf{x}_i + \hat{\theta}_2\right)\right)$$

Introduzione teorica al Machine Learning

Problema di ottimizzazione

Un esempio con la regressione lineare…

$$0 = -\frac{2}{N}\sum_{i=1}^{N}\mathbf{x}_i\left(y_i - \left(\hat{\theta}_1\mathbf{x}_i + \hat{\theta}_2\right)\right)$$

Svolgendo calcoli analoghi per l'altro parametro,
si ottengono due equazioni in due incognite

…

Per la regressione lin. si possono ottenere ***soluzioni esatte,***
generalmente è necessario implementare algoritmi
per la ricerca di ***soluzioni approssimate***

Introduzione teorica al Machine Learning

Problema di ottimizzazione

Non sempre i problemi di ottimizzazione sono «liberi»

In alcuni casi è necessario imporre dei vincoli ai valori che i parametri possono assumere

Si parla di problemi di ottimizzazione «vincolati»

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg \min_{(\theta_1, \theta_2)} \mathcal{R}(\theta_1, \theta_2)$$

$$\mathcal{R}(\theta_1, \theta_2) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$\text{con} \quad (\theta_1)^2 + (\theta_2)^2 \leq 1$$

Regolarizzazione RIDGE

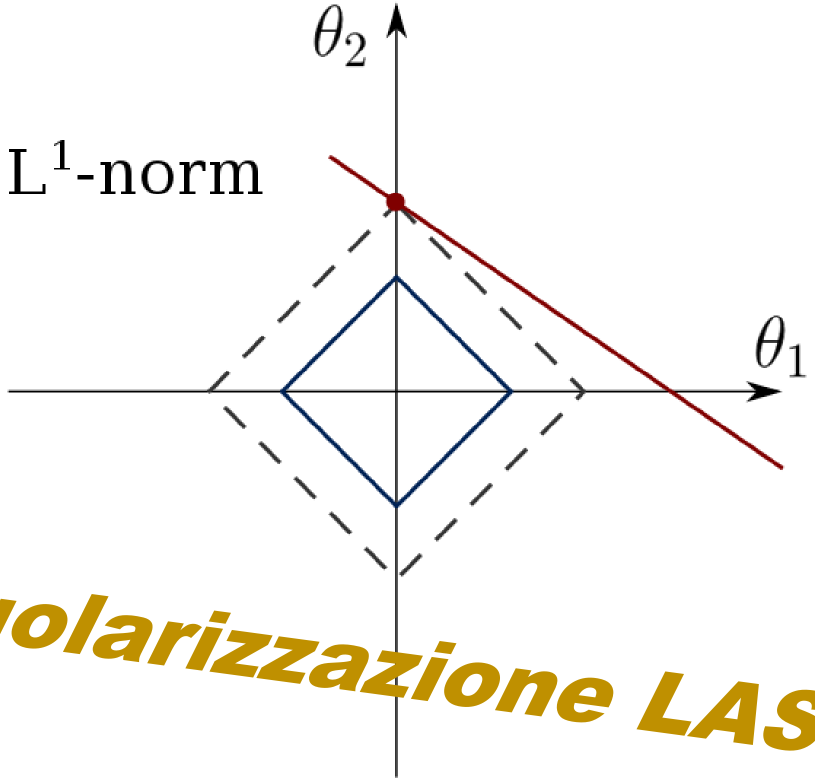
Introduzione teorica al Machine Learning

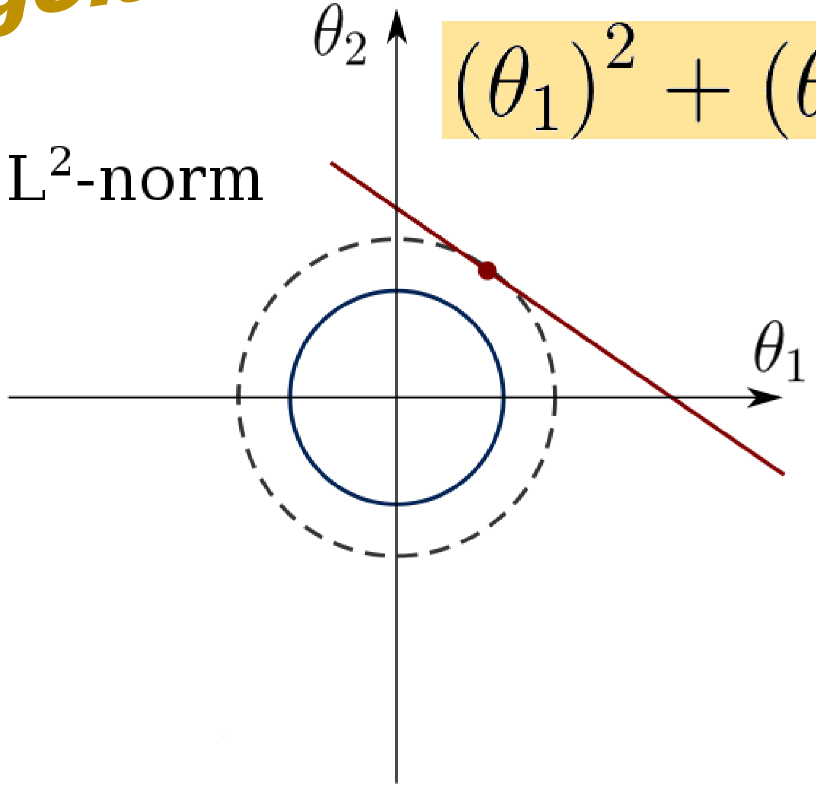
Problema di ottimizzazione vincolato

Un esempio…

Regolarizzazione RIDGE

$$(\theta_1)^2 + (\theta_2)^2 \leq 1$$





Regolarizzazione LASSO

Introduzione teorica al Machine Learning

Moltiplicatori di Lagrange

Il metodo dei moltiplicatori di Lagrange è una tecnica per studiare i massimi e minimi vincolati di una funzione a più variabili in riferimento ad un vincolo espresso mediante una o più equazioni, che individuano il vincolo come luogo geometrico di zeri.

youmath.it

problema vincolato

problema libero

Introduzione teorica al Machine Learning

Moltiplicatori di Lagrange

youmath.it

Moltiplicatori di Lagrange in due variabili con un vincolo

Sia $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ una funzione definita su un aperto $A \subseteq \mathbb{R}^2$, e sia $g(x, y) = 0$ un vincolo espresso sotto forma di luogo geometrico. Supponiamo che $f, g \in C^1(A)$, ossia che siano funzioni che ammettono derivate parziali continue su A.

Condizione necessaria ma non sufficiente affinché $(x_0, y_0) \in A$ sia un punto di estremo relativo per f rispetto al vincolo $g(x, y) = 0$ è che sussistano le seguenti condizioni:

1) $g(x_0, y_0) = 0$ e che inoltre il gradiente di g in (x_0, y_0) non sia nullo: $\nabla g(x_0, y_0) \neq 0$

2) Definita la *funzione lagrangiana*

$$L(x, y, \lambda) = f(x, y) - \lambda g(x, y)$$

esista un valore reale λ_0 tale per cui sia nullo il gradiente di L in (x_0, y_0, λ_0)

$$\nabla L(x_0, y_0, z_0)$$

In particolare la variabile λ è detta *moltiplicatore di Lagrange*.

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…

$$\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) \text{ subject to } h(\mathbf{x}) = 0$$

con

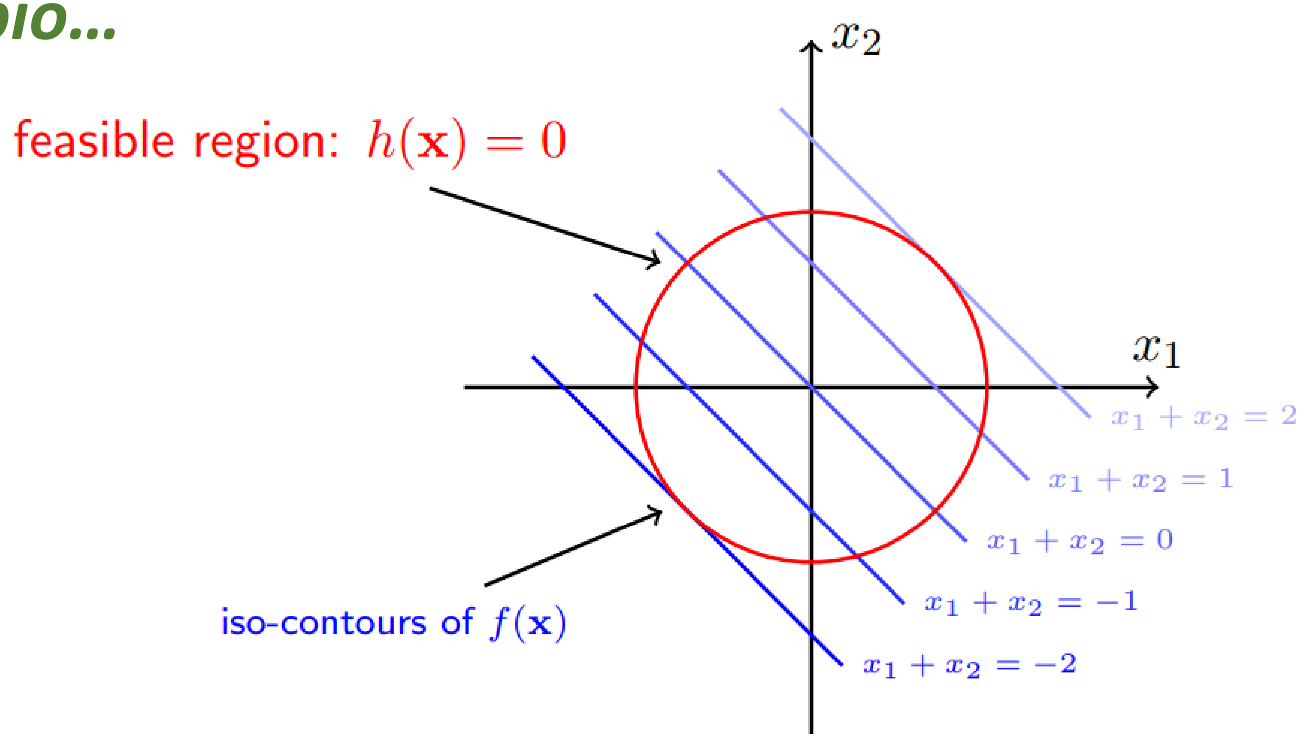
$$f(\mathbf{x}) = x_1 + x_2 \textbf{ and } h(\mathbf{x}) = x_1^2 + x_2^2 - 2$$

"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", deleeuwpdx.net/pubfolders/dual/KKT.pdf

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…



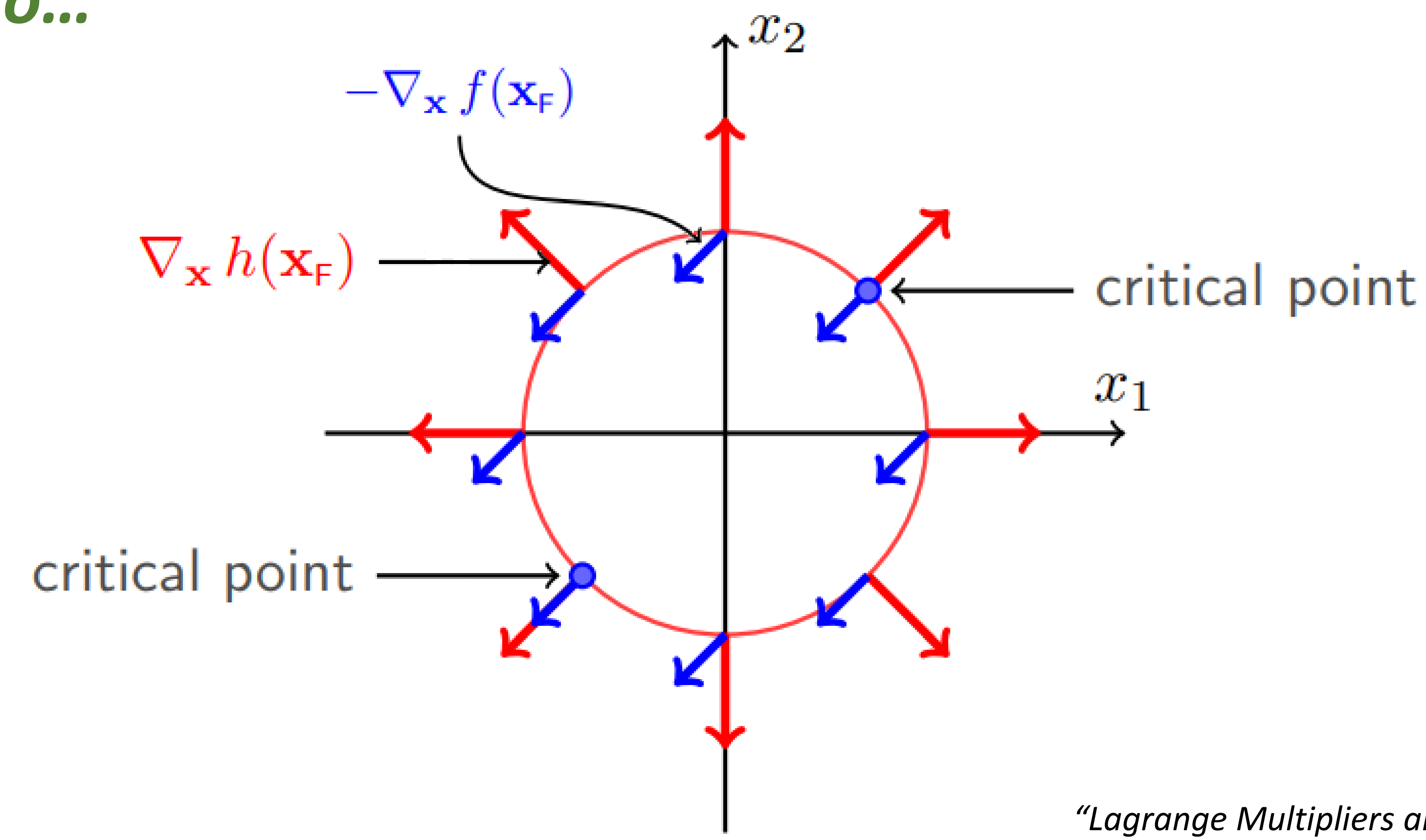
$$\boxed{h(\mathbf{x}) = x_1^2 + x_2^2 - 2}$$

"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", deleeuwpdx.net/pubfolders/dual/KKT.pdf

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…



"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", *deleeuwpdx.net/pubfolders/dual/KKT.pdf*

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) \text{ subject to } h(\mathbf{x}) = 0}$$

Define the **Lagrangian** as

$$\boxed{\mathcal{L}(\mathbf{x},\mu) = f(\mathbf{x}) + \mu\, h(\mathbf{x})}$$

Then $\mathbf{x}^*$ a local minimum $\iff$ there exists a unique μ^* s.t.

1. $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*,\mu^*) = \mathbf{0}$
2. $\nabla_{\mu}\mathcal{L}(\mathbf{x}^*,\mu^*) = 0$
3. $\mathbf{y}^t(\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^*,\mu^*))\mathbf{y} \geq 0 \quad \forall \mathbf{y} \text{ s.t. } \nabla_{\mathbf{x}}h(\mathbf{x}^*)^t\mathbf{y} = 0$

"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", *deleeuwpdx.net/pubfolders/dual/KKT.pdf*

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Una generalizzazione per vincoli con disuguaglianze:

Condizioni KKT

$$\min_{x \in \mathbb{R}} f(x) \quad \text{con} \quad g(x) \leq 0 \quad \Leftrightarrow \quad \min_{x \in \mathbb{R}} \left[f(x) + \lambda g(x) \right], \ \lambda \geq 0$$

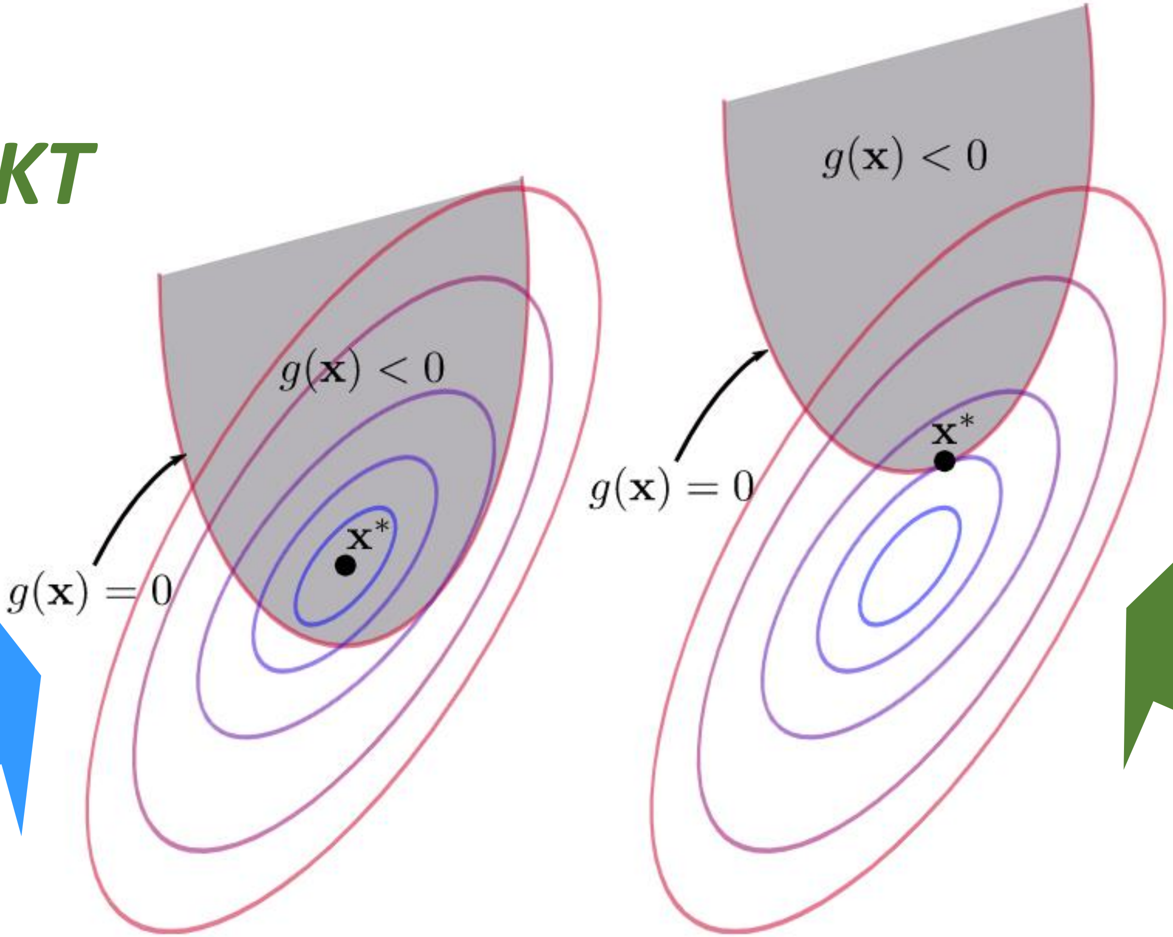
"Karush-Kuhn-Tucker (KKT) conditions", onmyphd.com

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Condizioni KKT

Se il punto di minimo libero è dentro la regione di vincolo allora il problema di ottimizzazione può essere trattato come libero



Se il punto di minimo globale libero è fuori la regione di vincolo, allora il punto che risolve il problema di ottimizzazione vincolato si trova sul bordo della regione

"Karush-Kuhn-Tucker (KKT) conditions", onmyphd.com

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) \leq 0$$

con

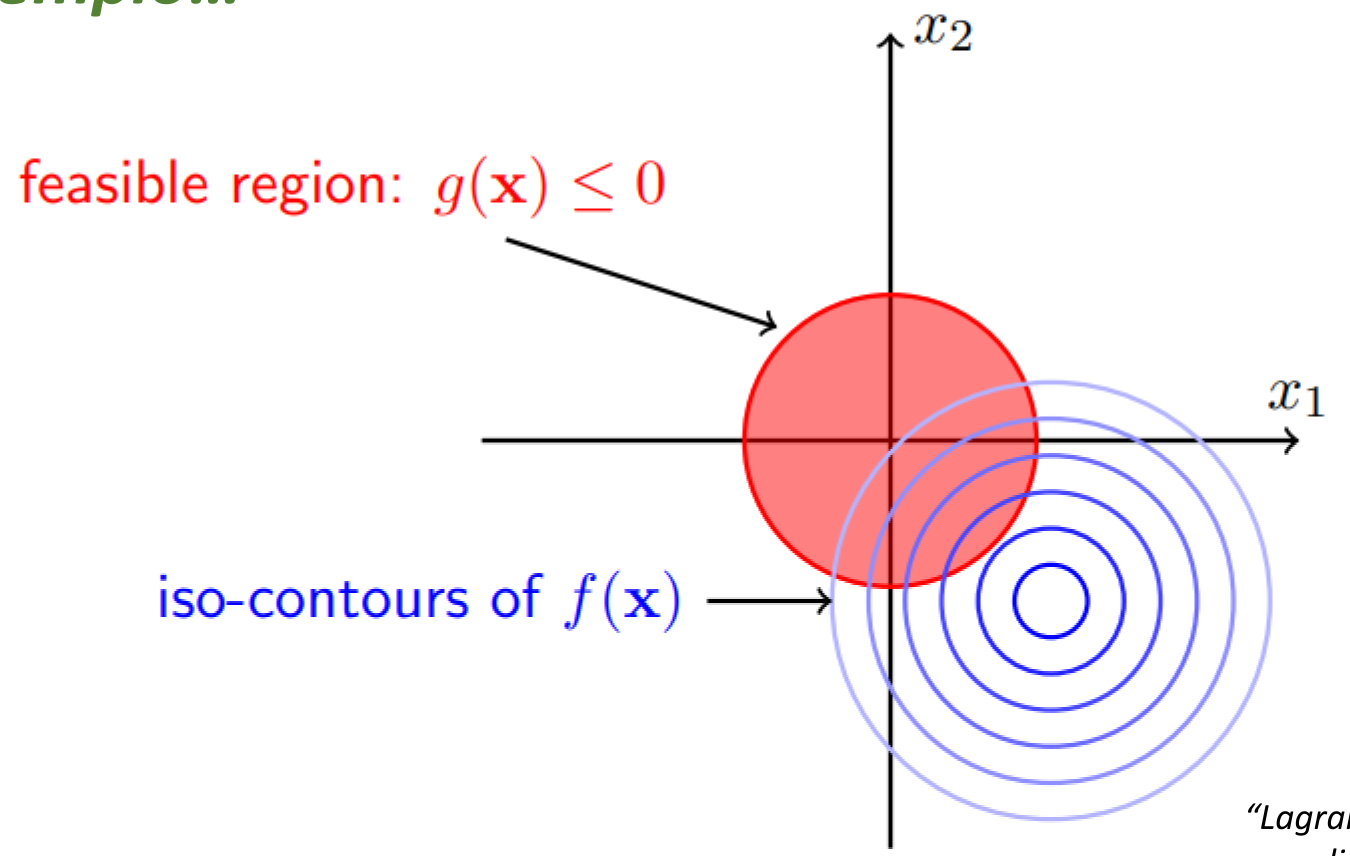
$$f(\mathbf{x}) = (x_1 - 1.1)^2 + (x_2 - 1.1)^2 \quad \textbf{and} \quad g(\mathbf{x}) = x_1^2 + x_2^2 - 1$$

"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", *deleeuwpdx.net/pubfolders/dual/KKT.pdf*

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…

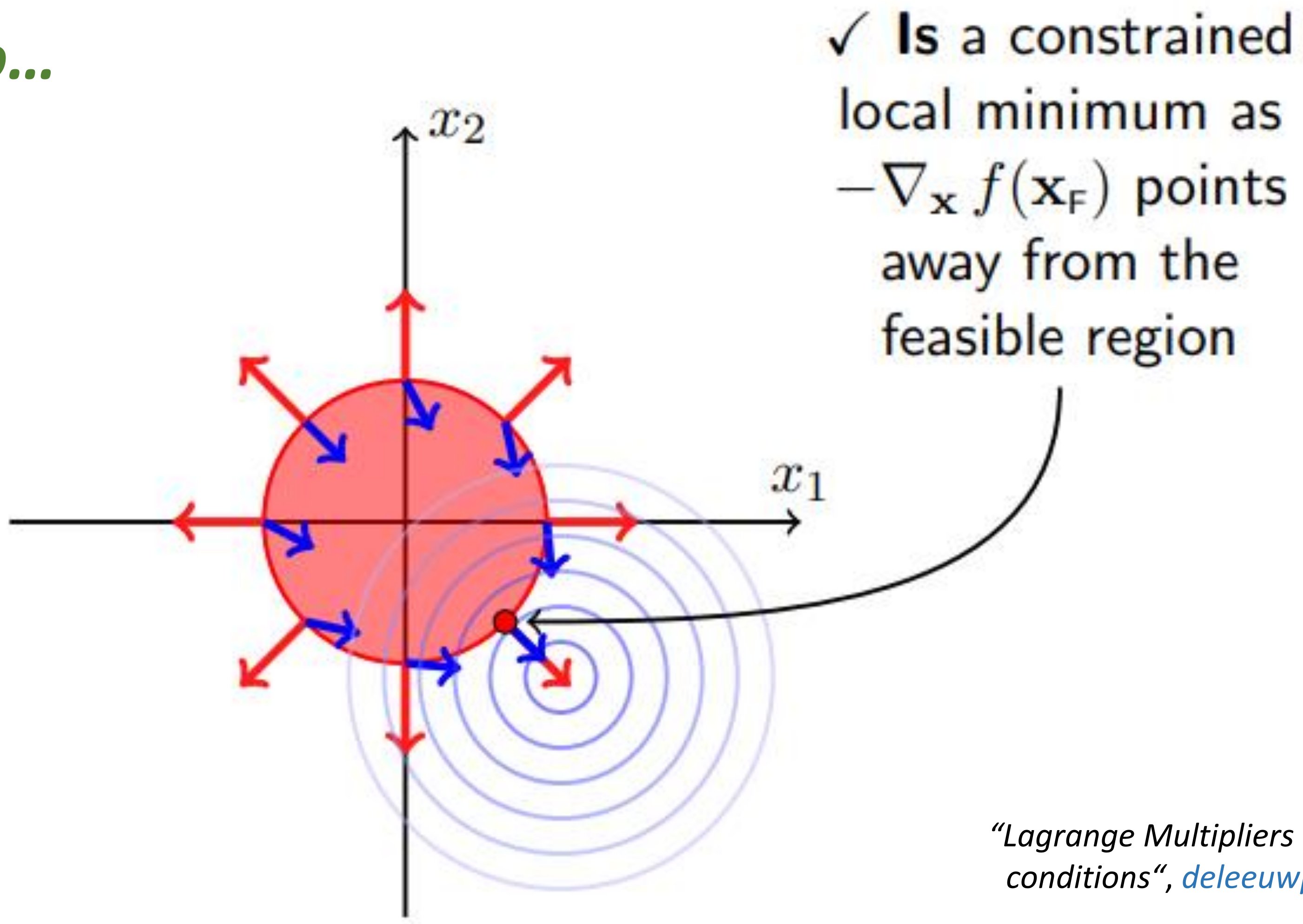


"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", *deleeuwpdx.net/pubfolders/dual/KKT.pdf*

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…



"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", *deleeuwpdx.net/pubfolders/dual/KKT.pdf*

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio…

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) \text{ subject to } g(\mathbf{x}) \leq 0}$$

Define the **Lagrangian** as

$$\boxed{\mathcal{L}(\mathbf{x},\lambda) = f(\mathbf{x}) + \lambda\, g(\mathbf{x})}$$

Then $\mathbf{x}^*$ a local minimum $\iff$ there exists a unique λ^* s.t.

1. $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*,\lambda^*) = \mathbf{0}$
2. $\lambda^* \geq 0$
3. $\lambda^*\, g(\mathbf{x}^*) = 0$
4. $g(\mathbf{x}^*) \leq 0$
5. Plus positive definite constraints on $\nabla_{\mathbf{xx}}\,\mathcal{L}(\mathbf{x}^*,\lambda^*)$.

"Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions", deleeuwpdx.net/pubfolders/dual/KKT.pdf

Introduzione teorica al Machine Learning

Problema di ottimizzazione vincolato

Un esempio… **Regolarizzazione RIDGE**

$$\min_{(\theta_1,\theta_2)} \mathcal{R}(\theta_1,\theta_2) \quad \text{con} \quad (\theta_1)^2+(\theta_2)^2 \leq t$$

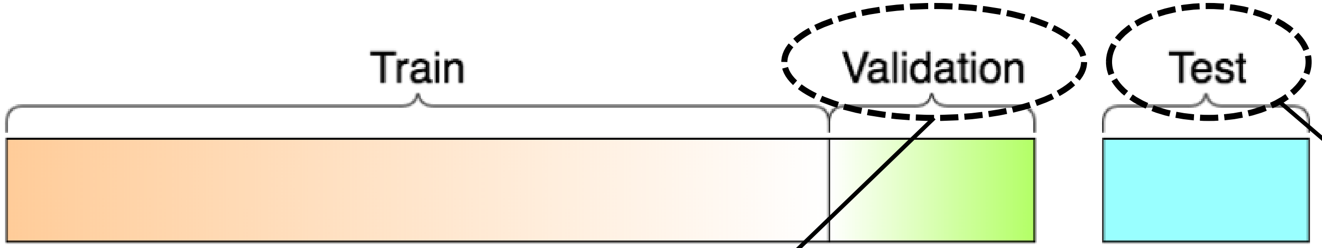
$$\mathcal{R}(\theta_1,\theta_2) = \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$$

$$\min_{(\theta_1,\theta_2)} \left[\mathcal{R}(\theta_1,\theta_2) + \lambda\left[(\theta_1)^2+(\theta_2)^2\right]\right], \ \lambda \geq 0$$

Introduzione teorica al Machine Learning

Apprendimento supervisionato

Validazione delle capacità predittive del modello



necessario solo se il modello possiede *iper-parametri*

«dovrebbe» essere rappresentativo dell'intera popolazione

Introduzione teorica al Machine Learning

Problematiche nel campionamento

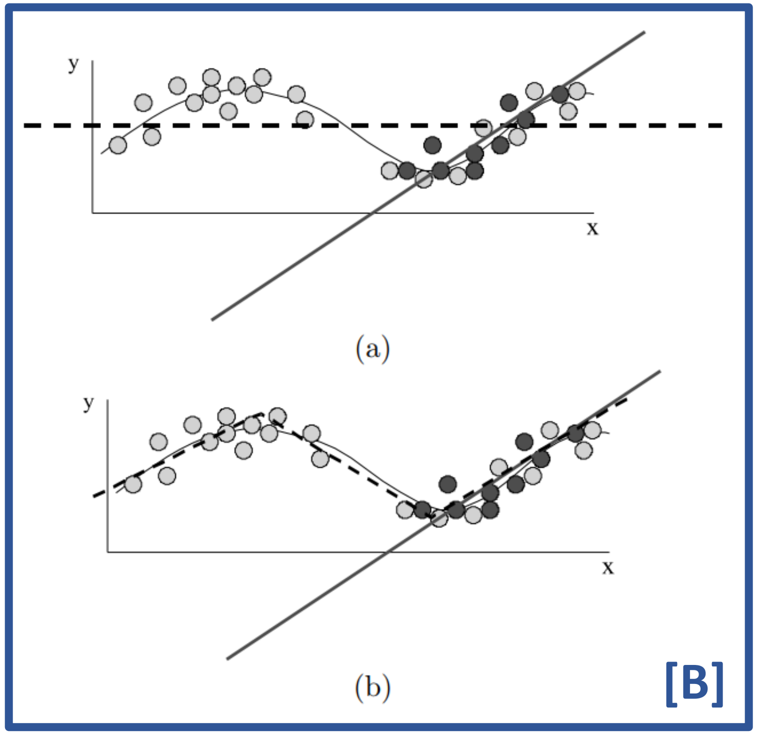
Problemi

A. ***Errori nel partizionamento del dataset***

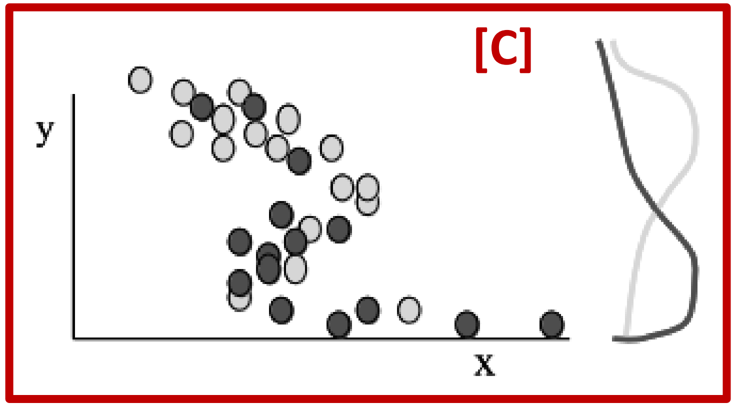
B. ***Covariate shift***

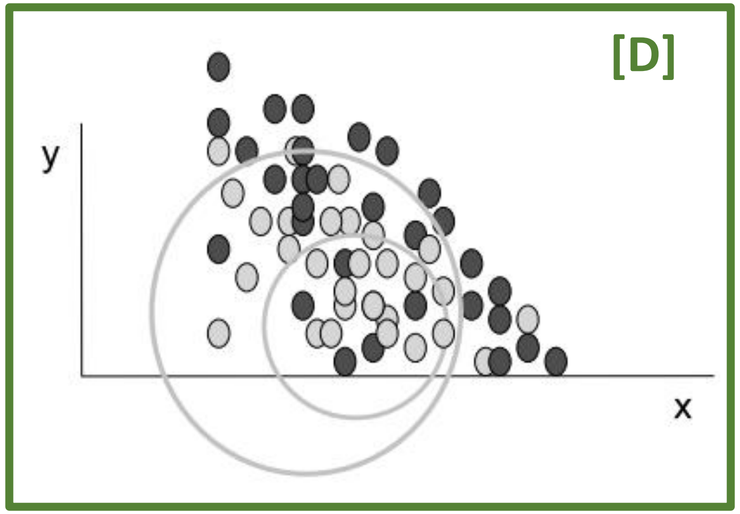
C. ***Probability shift***

D. ***Selection bias***



"When training and test sets are different: characterizing learning transfer", *Storkey, Amos (2013)*

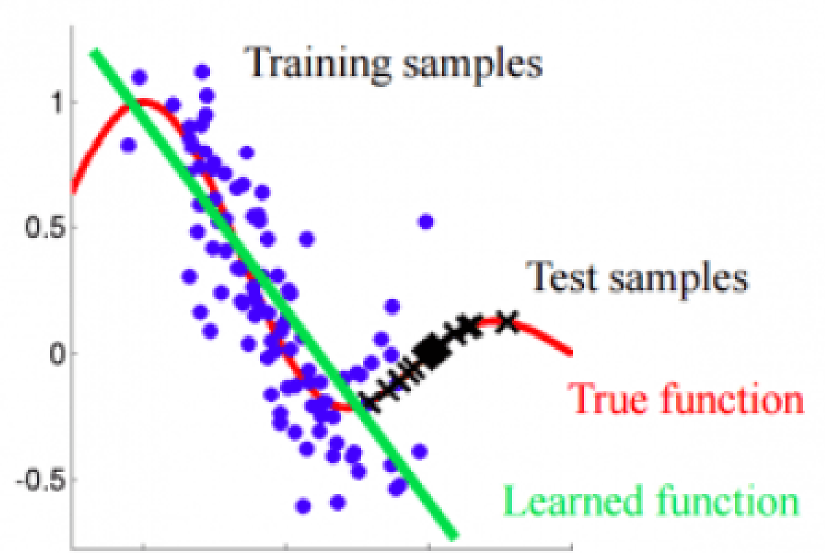




Introduzione teorica al Machine Learning

Problematiche nel campionamento

Un esempio di … ***Errore nel partizionamento del dataset***

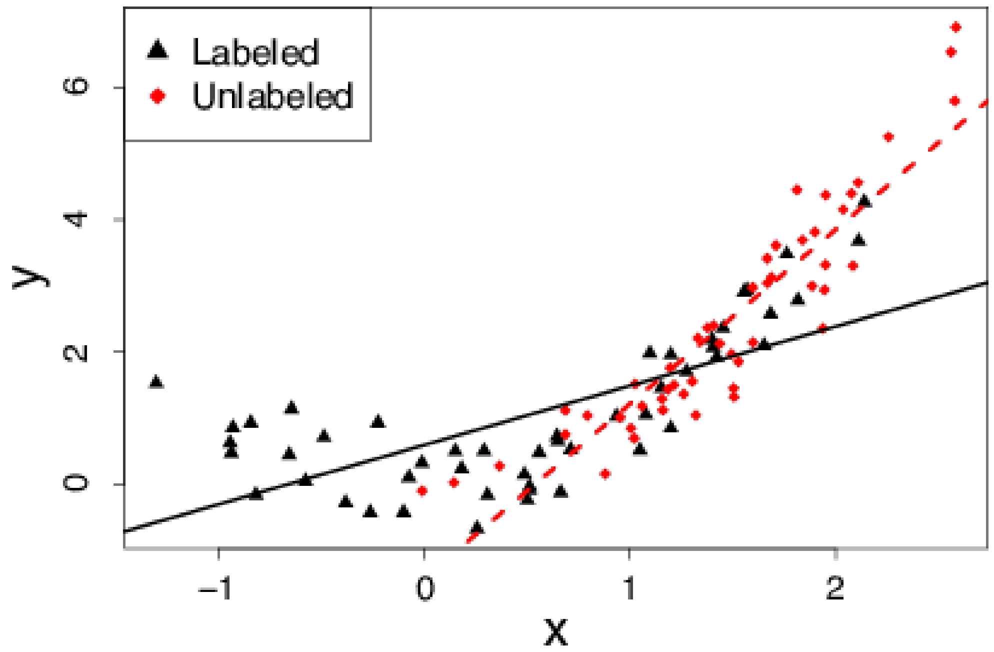


"Dataset Shift in Classification: Approaches and Problems", *Francisco Herrera (IWANN)*

Introduzione teorica al Machine Learning

Problematiche nel campionamento

Un esempio di … ***Covariate Shift***

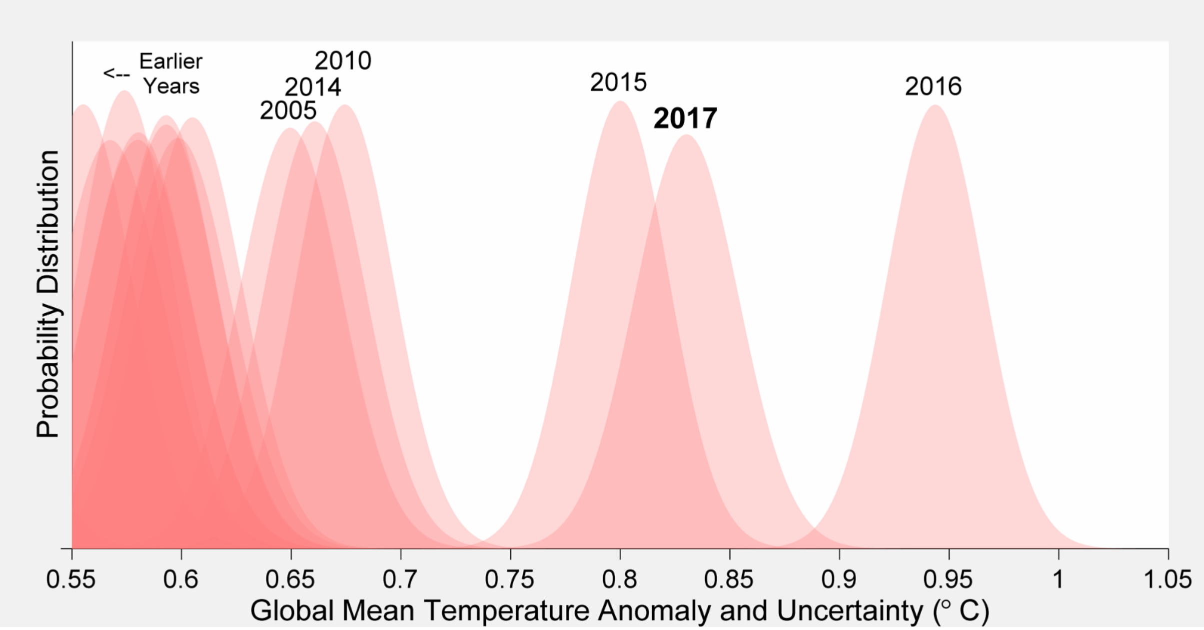


"Toy example for covariate shift in linear regression",
researchgate.net

Introduzione teorica al Machine Learning

Problematiche nel campionamento

Un esempio di … *Probability Shift*



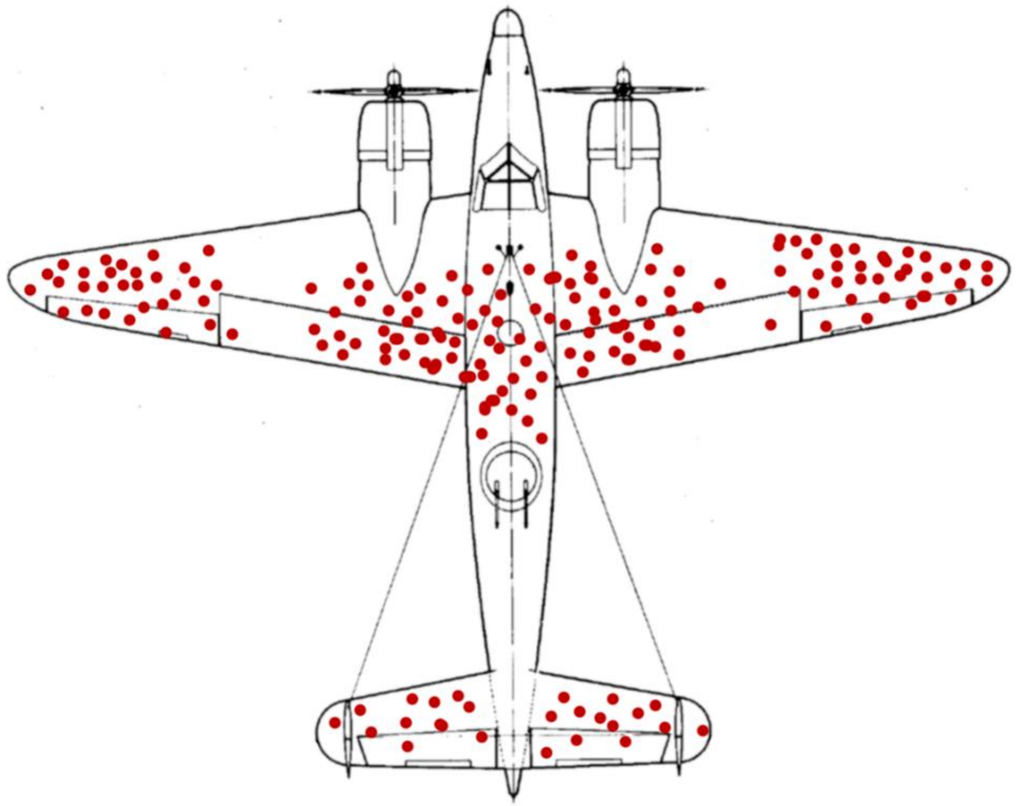
"Based on Berkeley Earth's estimates of the global annual average temperature increase relative to 1951-1980.",
berkeleyearth.org/global-temperatures-2017

Introduzione teorica al Machine Learning

Problematiche nel campionamento

Un esempio di … ***Selection Bias***

in particolare di … ***Survivorship Bias***



“Planes coming home from battle have bullet holes everywhere but the engine and cockpit, so we should put armor everywhere but the engine and cockpit.”

“Damage taken by planes able to come back after the fight. Image shows hypothetical data.“, *en.wikipedia.org*

Introduzione teorica al Machine Learning
Bias-Variance trade-off

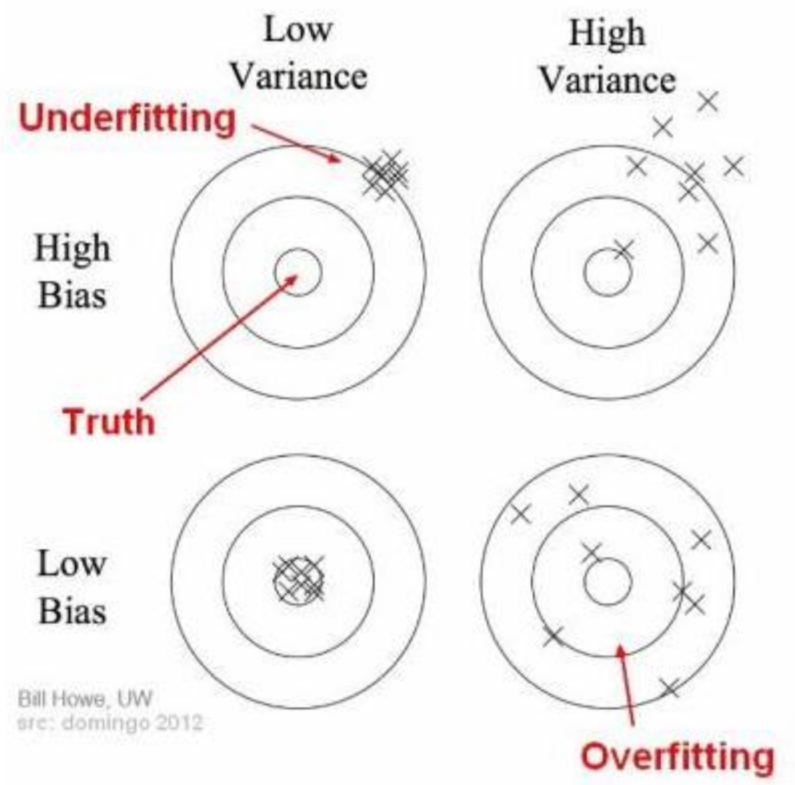
L'errore quadratico medio prodotto dal modello si può scrivere come

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$

che può essere decomposto

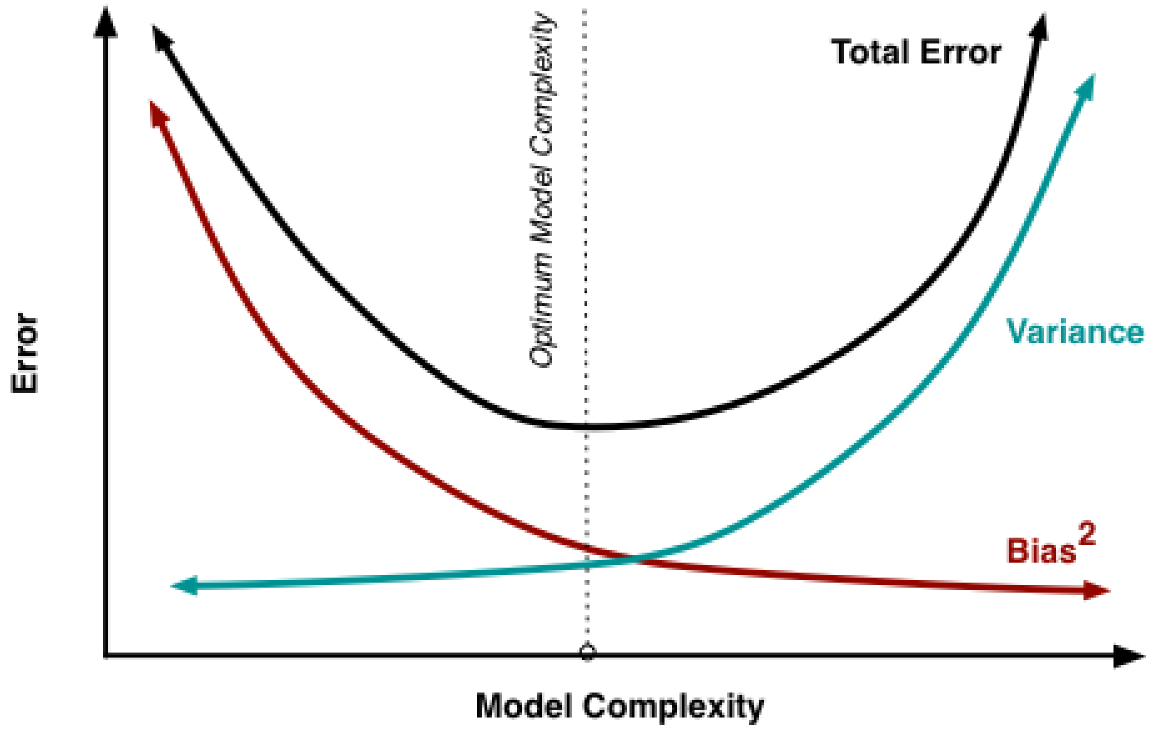
$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$



Introduzione teorica al Machine Learning

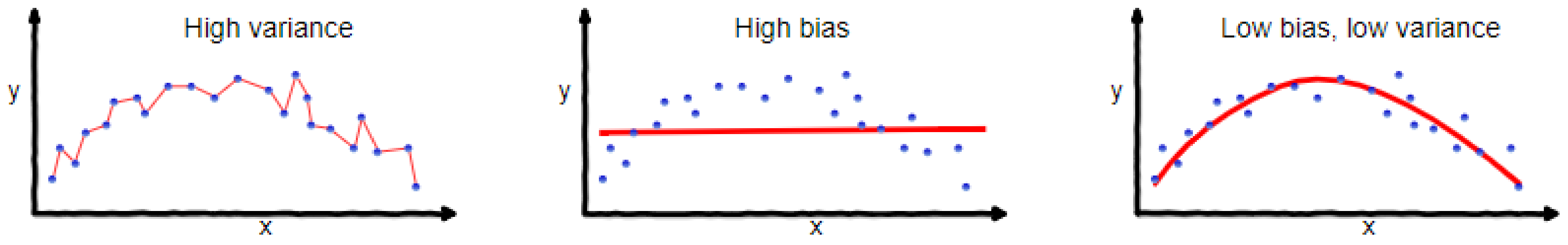
Bias-Variance trade-off



"Understanding the Bias-Variance Tradeoff", *Bryan White*

Introduzione teorica al Machine Learning

Bias-Variance trade-off



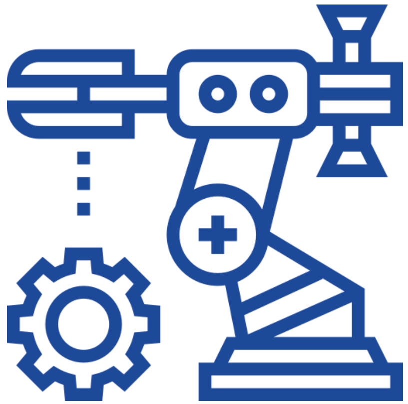
overfitting **underfitting** **Good balance**

Introduzione teorica al Machine Learning

Bias-Variance trade-off

Da Wikipedia:

- **Linear and Generalized linear** models can be regularized to decrease their variance at the cost of increasing their bias
- In **artificial neural networks**, the variance increases and the bias decreases as the number of hidden units increase (regularization applied)
- In ***k*-nearest neighbor** models, a high value of *k* leads to high bias and low variance
- In **decision trees** the depth of the tree determines the variance (pruned to control variance)



Grazie dell'attenzione

Fabio Mardero

fabio.mardero@gmail.com

apprendere per crescere insieme