Deep Learning

Daniele Avolio

October 10, 2023

Contents

1 Introduzione

Info esame: Tecnicamente, per ora rimane un progetto e potrebbe essere di gruppo. Ancora non ci sono informazioni precisissime.

2 Deep Learning 101

In questo corso affronteremo diverse tematiche, il che può sembrare assurdo se ci si pensa.

- Classificazione (binaria, cioè sì o no)
- Multi-class classification (non più binaria)
- Regressione (il guessing viene fatto su un valore numerico)
- Gestione di immagini e riconoscimento
- Serie numeriche (predizioni di mercato e trend)
- Classificazione di testi

2.1 Architetture e strumenti nel deep learning

- Autoencoder
  - Tutti i possibili tipi
  - Qui si fa anche **Clustering** e **Anomaly detection**
- Architetture generative
  - Tutti i possibili tipi
- XAI: Explainable AI

2.2 Libri utili

- "Deep Learning in Python"
- "Tensorflow tutorial"

2.3 Strumenti che useremo

- Tensorflow
  - High-level più di altri
- Keras
  - High-level API basato su Tensorflow
  - Ci saranno cose che non possiamo fare con Keras perché è troppo ad alto livello

2.4 Schema generale di un problema di deep learning

Abbiamo delle coppie $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ dove x_i è un vettore di features e y_i è un valore numerico (regressione) o una classe (classificazione).

$$y_i = f(x_i)$$

Non conosciamo la funzione f, quindi dobbiamo impararla.

$$y = \alpha x + \beta$$

Con una rete neurale puoi approssimare praticamente qualsiasi funzione.

Una rete neurale permette di collegare un input di dati a una funzione di output.

Abbiamo diversi tipi di reti neurali a seconda del tipo di problema che vogliamo risolvere. È importante essere in grado di selezionare l'architettura giusta per risolvere il problema.

Definiamo alcuni concetti che useremo:

- N: rete neurale
- w: valori dei pesi della rete neurale
- f: funzione di output della rete neurale

$$f \in N(w)$$

2.5 Perché si usa il termine "Tensore"?

Un *tensore* non è altro che una matrice.

- 0D tensor: scalar
- 1D tensor: vector
- 2D tensor: matrix
- 3D tensor: tensor

2.6 AI vs DL

- AI: è un ampio insieme di tecniche per risolvere problemi che richiedono "intelligenza".
  - Esempio: Stockfish, un programma che gioca a scacchi.
- DL: È un sottoinsieme di AI che si concentra sull'*astrazione*.
  - L'astrazione consiste nel fornire una funzione che traduce dati di input in dati di output senza conoscere la funzione stessa.
  - È un approccio induttivo: fornisci un input e ti aspetti un output, senza conoscere la funzione che li collega.
  - Questo è completamente diverso dall'AI basata sulla logica.

3 Introduzione alle Reti Neurali

3.1 Il modello di McCulloch-Pitts

Un modello pensato dai due tizi qui presenti,

- Warren McCulloch
- Walter Pitts

Era formato da:

- Un insieme di neuroni
- Un insieme di connessioni tra i neuroni
- Un insieme di pesi associati alle connessioni
- Una funzione di attivazione
- Una funzione di output

Quindi immaginiamo $x_1, x_2, \ldots, x_n$ che vengono dati come input, e quello che viene fuori è un valore di $y \in \{0, 1\}$. Questo è un modello di classificazione binaria.

In soldoni, in deep learning si usa un vettore di numeri per tirare fuori un altro vettore di numeri.

Nella maggior parte del tempo, però, non lavoriamo solamente con i dati.

- immagini
- Testo
- Audio
- ...

Non ci sono concetti di foto, video, immagini. L'unico concetto che esiste è quello dei numeri.

3.2 Modello di Rosenblatt

Qui il modello è leggermente diverso. Gli elementi in questo modello sono i seguenti:

- Valori di input: x_i

- Funzione di attivazione: ϕ
- Pesi degli archi: w_i
- Bias: b

$$h(x|w,b) = h(\sum_{i=1}^{l} w_i \cdot x_i - b) = h(\sum_{i=1}^{l} w_i \cdot x_i) = sign(w^T x) \tag{1}$$

Nota: Quando il **bias** non viene specificato, allora si assume che sia 0.

I pesi w_i sono collegati archi che vanno da x_i al prossimo neurone. Sia il valore di input che il peso sono **numeri reali**. Non sono lo stesso valore, hanno solamente il formato di *reale* che è uguale tra loro. Ciò che viene fatto è solamente la somma della prodotto tra ogni peso w_i e x_i **meno** il bias.

La funzione di attivazione: Dipende. Ogni funzione che ha *2 stati* va bene per noi. Una funzione di attivazione può essere una qualsiasi che in un punto ha valore 1 e in un altro ha valore -1.

3.3 Esempio con 2 neuroni

.

Immaginiamo di avere un piano cartesiano con una retta che interseca in 2 punti.

$$sign(w_1 \cdot x_1 + w_2 \cdot x_2) \tag{2}$$

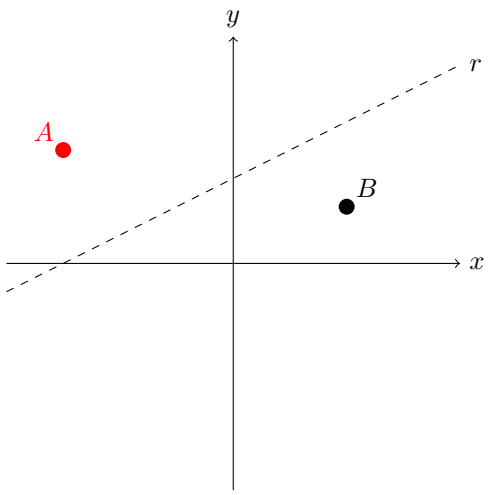
Il parametro della funzione non è altro che **l'equazione di una retta.**

In particolare, se consideriamo la reta che separa gli spazi del piano, vediamo che la retta è **capace di separare 2 punti nel piano.**

Cosa abbiamo:

- Un neurone
- 2 valori in input
- 2 pesi

Con la funzione di attivazione **sign** si avrà come output una retta che separa



dei punti.

Nota: Quando è utile avere un valore output che non è una separazione di punti in un piano? Nel caso della **regressione**.

Oltre la funzione sign:

- Funzione di attivazione lineare
- Funzione di attivazione sigmoid
- Funzione di attivazione Tanh

3.4 Rappresentazione le funzioni logiche

:

3.4.1 AND

Immaginiamo di avere:

- $x_1, x_2 \in \{0, 1\}$
- Bias= -30
- Funzione di attivazione: Logistica o Sigmoid

Come facciamo a rappresentare un AND?

$$h(x) = g(-30 + 20x_1 + 20x_2) \tag{3}$$

x_1	x_2	$h(x)$
0	0	1
0	1	0
1	0	0
1	1	1

Lo stesso ragionamento vale per:

- OR
- NOT
- (NOT x_1) AND (NOT x_2)

3.4.2 Il problema dello XOR

Il percettrone non può imparare regioni che non sono linearmente separabili.

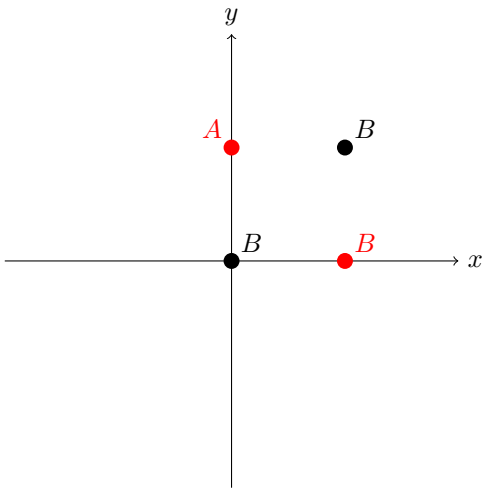


Figure 1: XOR non possibile con 1 percettrone

Come vediamo qui non possoamo tracciare una retta per dividere i due punti. In questo caso ci serve una funzione **non lineare**.

La soluzione a questo problema è **aggiungere LAYER** alla rete neurale. Aggiungere un layer significa aggiungere un neurone successivamente ad un altro.

Maggiore è il numero di layer, maggiore diventa la potenza espressiva della rete. Praticamente possiamo catturare qualsiasi cosa aggiungendo layer alla rete. Questo è **il vero potere del deep learning**.

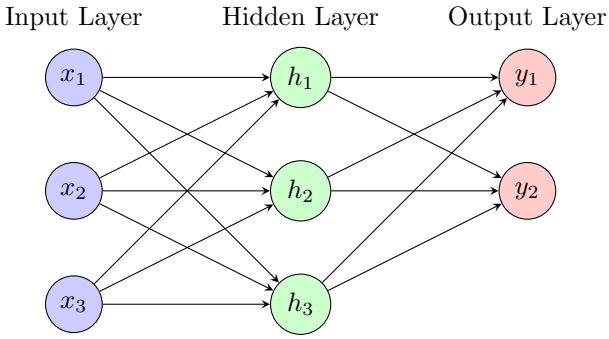


Figure 2: Neural Network con 1 hidden layer

3.5 Gli ingredienti di una rete neurale

3.5.1 Il grafo g

Un grafo $g = \{N, E\}$ è un grafo diretto pesato con label.

Ogni nodo $i \in N$ viene chiamato **neurone** o **percettrone**. Per ogni nodo ci sono 2 targhette:

- Un valore a_i che viene chiamato **attivazione**
- Una funzione di attivazione f_i che viene applicata all'attivazione, che produce un output z_i

Ogni arco $e = \{j \in N \to i \in N\} \in E$ associato con un peso $w_{j,i}$.

Ogni nodo i è anche coinvolto con un arco speciale, con un nodo fantasma, che viene chiamato **bias** b_i.

Nota: $z_i = f_i(a_i)$. E $a_i = b_i + \sum_{j:j\to i\in E} w_{j,i} z_j$

La combinazione dei neuroni connessi costruisce il grafo. I nodi che condividono gli stessi input sono raggruppati in **layers**.

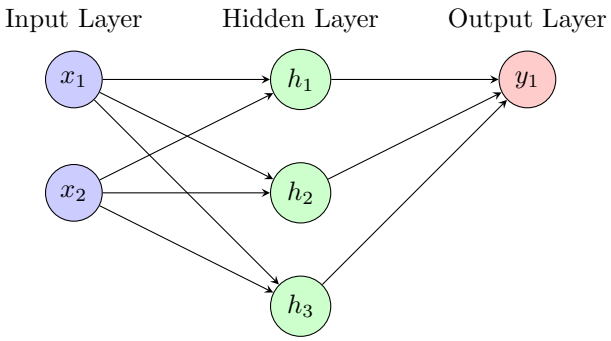


Figure 3: Neural Network con 1 hidden layer

Il risultato finale dipende da tutti i parametri del grafo, e anche dal tipo di funzione di attivazione che viene usata.

> Ma come facciamo a scegliere il corretto valore dei parametri del grafo?

3.5.2 La funzione di loss

Il grafo è un operatore algebrico non lineare: $g(\vec{x}|W, B)$. In questo grafo non sappiamo il valore di **B e W**. La fase di apprendimenti di una rete consiste nel trovare il **migliore** valore per W e per B. Ma cosa intendiamo con **migliore?**.

Formalmente, l'unico modo che abbiamo per definire la conezione di migliore, è quello di *approssimare al meglio la funzione che vogliamo trovare.*

Consideriamo il valore di una funzione F su x_i e il vero valore di y_i. Noi vogliamo minimizzare la differenza tra $F(x_i)$ e y_i.

$$\min_{W,B} \frac{1}{n} \sum_{i=1}^{n} loss(\vec{y_i}, g(x_i|W, B)) \tag{4}$$

con:

- $loss(\vec{y_i}, g(x_i|W, B))$ è una funzione che misura la differenza tra y_i e $g(x_i|W, B)$
- n è il numero di esempi
- W e B sono i parametri del grafo
- $g(x_i|W, B)$ è il valore di output del grafo
- $\vec{y_i}$ è il valore di output vero

La funzione di **loss** dipende dal tipo di task che dobbiamo svolgere.

3.5.3 Funzione di loss per Regressione

Mean Absolute Error:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - g(\vec{x_i}|W,B)| \tag{5}$$

- Considera tutti gli errori con lo stesso peso
- **Non è differenziabile in 0**

Mean Squared Error:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - g(\vec{x_i}|W,B))^2 \tag{6}$$

- Gli errori più grandi hanno un peso maggiore
- **È differenziabile in 0**
- **È più sensibile agli outliers**

Domanda: quale si usa tra le due? Dall'approccio greedy, **usa entrambe**. **Esistono anche altre funzioni:**

- Smooth Absolute Error
- Huber Loss

3.5.4 Funzione di loss per classificazione

Binary Cross Entropy [BCE] Viene usata se $y_i \in \{0,1\}, g(\vec{x}|W,B) \in [0,1]$.

$$BCE = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(g(\vec{x_i}|W,B)) - (1-y_i)\log(1-g(\vec{x_i}|W,B)) \tag{7}$$

Categorical Cross Entropy [CCE]

$$CCE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i,j}\log(g(\vec{x_i}|W,B))_j \tag{8}$$

3.6 L'ottimizzatore o

Come risolviamo il problema di:

$$\min_{W,B}\frac{1}{n}\sum_{i=1}^{n} loss(\vec{y_i}, g(\vec{x_i}|W,B)) \tag{9}$$

Il problema lo risolviamo calcolando il **gradiante** della funzione loss, lo poniamo uugale a 0 e controlliamo se siamo in un punto di **massimo, minimo o punto di sella**.

$$\frac{\partial loss}{\partial W} = 0 \tag{10}$$

Abbiamo bisogno di un metodo iterativo per trovare una soluzione. Ci vuole un'**euristica**. Tipicamente, siamo soddisfatti di un **minimo locale**, e si utilizza, appunto, il **metodo di discesa del gradiente**.

Differenza minimo locale e globale

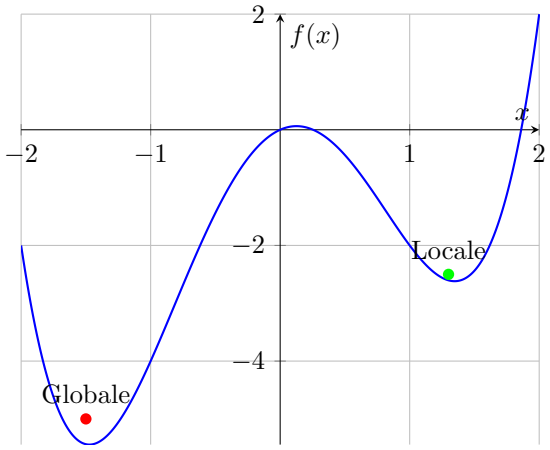


Figure 4: Differenza tra minimo locale e globale

3.6.1 Il metodo di discesa del gradiente

Sia $F(\vec{x})$ una funzione differenziabile.

- $F(\vec{x})$ decresce più veloce nella direzione del gradiente negativo
- $F(\vec{x})$ cresce più veloce nella direzione opposta al gradiente

$$a^{new} = a^{old} - \eta \cdot \nabla F(a^{old}) \tag{11}$$

Il parametro η viene chiamato **learning rate** e determina il comportamento del'ottimizzazione. Da notare che e' l'unico **parametro**.

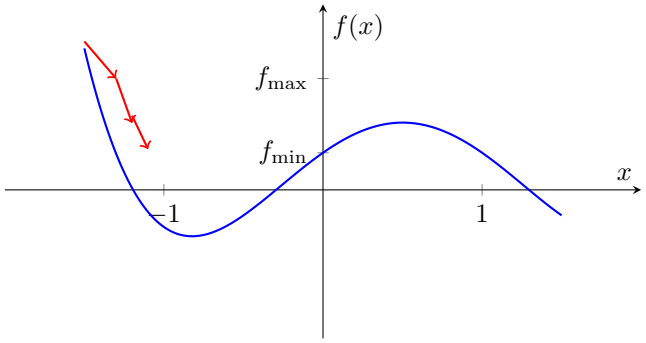


Figure 5: Discesa del gradiente

Ci sono ancora problemi: Questo metodo non è esente da problemi, quindi abbiamo:

- Dipendentemente dal punto di inizio, abbiamo un risultato diverso. Dobbiamo capire che, giustamente, se ci accontentiamo di un minimo locale, non avremo quasi mai lo stesso minimo per ogni allenamento della rete.

Nota: con **topologia** intendiamo il numero di layer e il numero di neuroni per layer.

Diciamo che in generale, gli steps per allenare una rete sono:

```
for each topology T in C_t
    for each η in C_η
        for each initialization I in C_I
            Applica la discesa del gradiente
```

3.6.2 Il metodo di discesa del gradiente stocastico

La differenza del metodo di discesa del gradiente stocastico è che, invece di calcolare il gradiente su tutti i dati, si calcola il gradiente su un sottoinsieme di dati.

Ciò che viene fatto, per ogni step, invece di prendere la derivata di ogni loss function e poi fare i calcoli, prendo **un sample del mio dataset**, tipicamente chiamato **batch**. Ogni volta che calcolo la discesa del gradiente, calcolo la derivata **NON PER TUTTO IL DATASET**, ma solamente del **batch**. Questo OVVIAMENTE non mi assicura che ottimizzare ogni batch mi ottimizza anche l'intero dataset, ma non c'è modo di lavorare sull'intero dasaset, poiché questo è troppo grande e richiede troppo tempo.

Nonostante tutto, utilizzando questa tecnica **si ottengono risultati comunque accettabili**.

Il *workflow* allora cambia in:

```
for each topology T in C_t
    for each η in C_η
```

```
        for each initialization I in C_I
         for each batch size b ∈ C_b
           Applica la discesa del gradiente stocastica
```

3.6.3 Linee guida sul learning rate

Epoca: Un'epoca è un passaggio dell'approssimazione della funzione

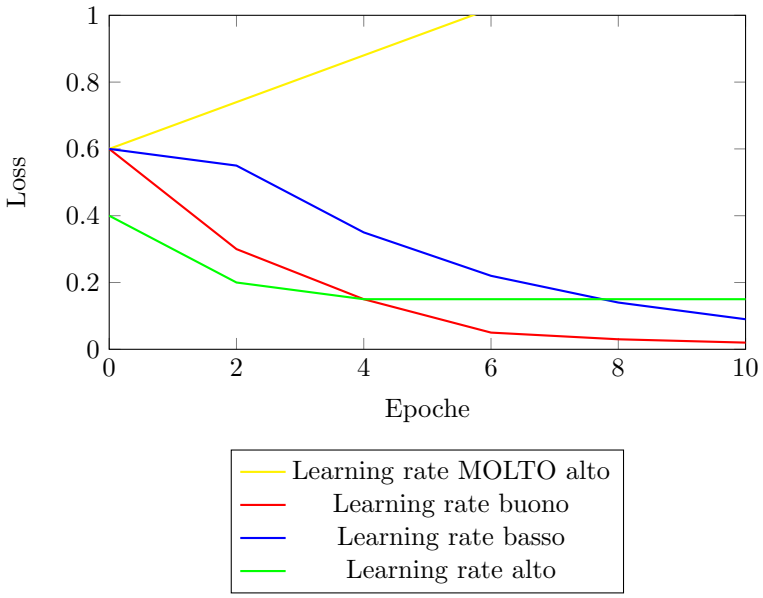


Figure 6: Esempio di andamento della funzione di loss in base al learning rate

Analisi dei learning rate:

- learning rate basso: troppo lento!
- learning rate alto: fermo con un minimo locale immediatamente
- learning rate MOLTO alto: non converge mai
- learning rate buono: decresce sempre e eventualmente converge

Nota: Il learning rate è un parametro che va **tunato**.

- Se il learning rate è troppo alto, non si riesce a convergere poiché si salta il minimo
- Se il learning rate è troppo basso, si rischia di non convergere mai

Il nostro obiettivo è quello di avere un learning rate che sia **giusto** e che permetta di avere un andamento della funzione di loss come quello in rosso, cioé avere una buona diminuizione della loss andando avanti con le epoche. Sempre decrescente e con la distanza minore dall'asse delle x.

Domanda: Vogliamo sempre la loss uguale a 0? **No.** Perché? Perché a differenza che in statistica dove vogliamo avere un errore il minimo. In deep learning non è così. Se la loss è 0, allora vuol dire che la rete ha imparato a memoria i dati, e non è quello che vogliamo. Vogliamo che la rete sia capace di generalizzare.

In particolare, noi **vogliamo fare predizioni**, sopratutto su **istanze ancora non viste.** Se avessimo una loss uguale a 0, probabilmente non saremmo in grado di avere delle predizioni decenti, poiché la rete non è in grado di generalizzare. Avendo la loss non a 0 rendiamo la rete capace di poter introdurre istanze ancora non viste in futuro.

Quindi, per statistica **loss = 0 : nice**, per **deep learning: insomma**.

4 Classificazione Binaria

Descrizione di cosa andremo a fare: utilizziamo il dataset IMDb Dataset. Questo dataset è un Database compilato dagli utenti del sito. Le caratteristiche, in particolare, sono:

- 50.000 recensioni
- circa 50% positive e 50% negative
- *25.000* sono usate per il **training** e *25.000* sono usate per il **testing**. Anche queste sono il 50% positive e 50% negative.

La task che faremo su questo dataset prende il nome di **sentimental analysis**, ovvero analisi del sentimento.

Assunzione: Ciò che stiamo facendo ha senso solamente se assumiamo *che nel futuro avremo punti con una distribuzione abbastanza simile a quelli utilizzati per il training.*

La partizione di **test** è una partizione che non viene utilizzata per la fase di training, ma viene utilizzata successivamente per controllare se il modello si sta comportando bene nella predizione dei valori. Ci sono delle misure che hanno range $[0, 1]$ che ci permettono di capire quanto il modello si sta comportando bene.

Nota: il *test set* non viene fornito quando si lavora nel deep learning, altrimenti verrebbe usato per ottimizzare direttamente il modello. **Non si conosce inizialmente.**

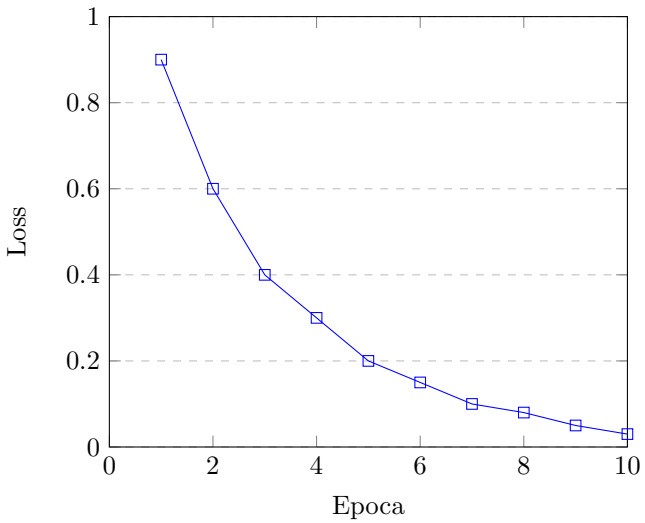


Figure 7: Grafico di una curva di loss che diminuisce all'aumentare delle epoche.

4.1 Caricamento e gestione del Dataset

```
from keras.datasets import imdb
(train_data, train_labels), (test_data, test_labels) = imdb.
    load_data(num_words=10000)
```

Listing 1: Caricamento del dataset IMDb Dataset.

- training_data: sono gli input del training
- training_labels: sono gli output del training
- test_data: sono gli input del test
- test_labels: sono gli output del test

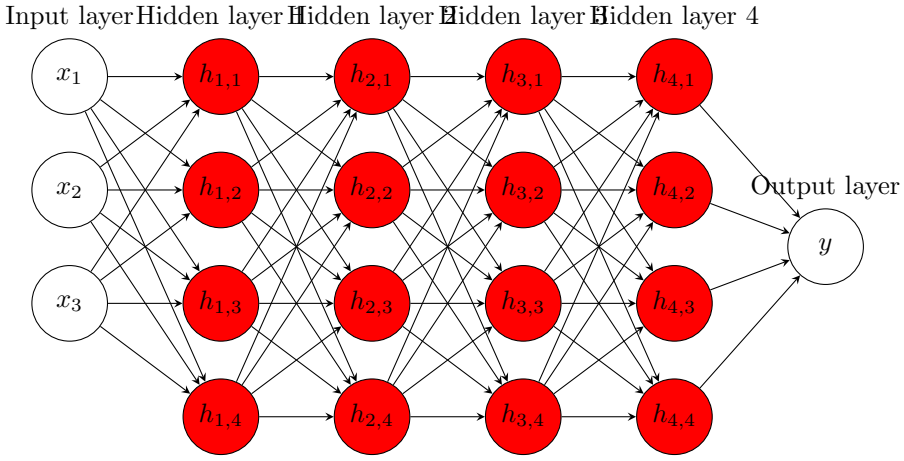


Figure 8: Neural network with 3 input nodes, 4 hidden layers with 3 nodes each, and 1 output layer with 1 node.

Quali sono i primi problemi che vanno gestiti in questo caso? Abbiamo il problema che i **dati sono parole e non numeri**. Quindi dobbiamo trasformare le parole in numeri.

Ogni parola viene **trasformato** in un numero. Ci troviamo però con delle recensioni che sono delle **sequenze di parole**, e quindi abbiamo delle **sequenze di numeri**. La soluzione è quelal di usare un **set di parole codificate.** Mi spiego:

Immaginiamo di avere 10.000 parole encodate. Salvo queste 10.000 parole in un array e associo ad ogni parola un indice di questo array.

0	1	2	3	4	5	6	7	8	9

Ora, cosa manca? **Manca l'ordine**. L'unica cosa che abbiamo è quindi una indicizzazione delle parole, ma non abbiamo alcuna informazione dell'ordine. **Manca anche totalmente la semantica**.

La struttura dell'input è quindi un **livello con 10.000 nodi di input.s**

Input

```
import numpy as np
def vectorize_sequences(sequences, dimension=10000):
# Create an all-zero matrix of shape (len(sequences), dimension)
results = np.zeros((len(sequences), dimension))
for i, sequence in enumerate(sequences):
    results[i, sequence] = 1. # set specific indices of results[i]
    to 1s
return results
# Our vectorized training data
x_train = vectorize_sequences(train_data)
# Our vectorized test data
x_test = vectorize_sequences(test_data)
```

Output

```
# Our vectorized labels
y_train = np.asarray(train_labels).astype('float32')
y_test = np.asarray(test_labels).astype('float32')
```

4.2 Definizione della Rete Neurale

Codice iniziale della definizione della Rete

```
    from keras import models
from keras import layers
model = models.Sequential()
model.add(layers.Dense(16, activation='relu', input_shape=(10000,))
    )
model.add(layers.Dense(16, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
model.compile(optimizer='rmsprop',
              loss='binary_crossentropy',
               metrics=['accuracy'])
```

Il **modello sequenziale** è quello migliore da dove iniziare. Cosa significa però modello sequenziale? Il modello sequenziale ha il concetto di **layer**. Ha la funzione **add** che aggiunge un layer sopra gli altri layer che sono già esistenti.

Primo Layer:

layers.Dense: un layer denso significa che ogni nodo di quel layer è collegato con ogni nodo del layer precedente. **16** è il numero di neuroni che si vogliono

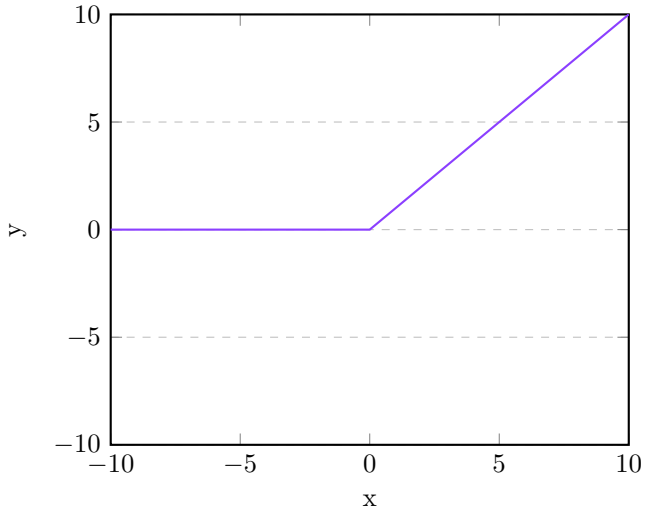
attivare in quel layer. input_shape è la dimensione dell'input. **10000** è la dimensione dell'input. Notare che si ha una **virgola** nell'input dopo il 10.000 e indica che *che stiamo aspettando una sequenza di vettori, ognuno di dimensione 10.000* Cioè, praticamente stiamo dicendo **10.000** sono le parole per ogni recensione, e con la virgola stiamo dicendo che **non sappiamo quante recensioni abbiamo**.E' comodo quando non abbiamo un numero fisso di example del dataset da processare.

Quante connessioni ci sono?

$$16 \cdot 10000 + 16 = 160016$$

dove sono 16 i bias.

Cosa significa relu? ReLu è la funzione di attivazione.



E' una funzione di attivazione artificiale che fino a 0 è 0, e poi cresce linearmente. E' una funzione che si usa molto in deep learning.

Secondo layer:

Nel secondo layer abbiamo altri **16** nodi connessi con i precedenti 16 nodi. In tutto abbiamo

$$16 \cdot 16 + 16 = 272$$

connessioni, con 16 che sono i bias.

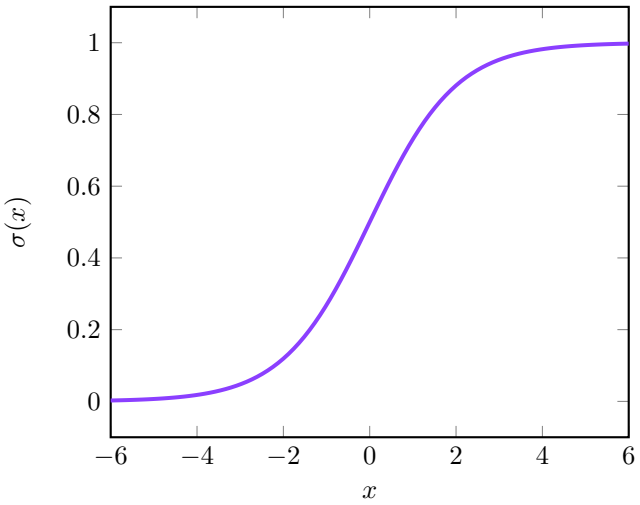
Layer di output:

In questo layer abbiamo un unico nodo connesso con i 16 nodi precedenti. In tutto abbiamo

$$16 \cdot 1 + 1 = 17$$

connessioni, con 1 che è il bias. La funzione di attivazioe è la **sigmoid**, che è una funzione che ha range $[0, 1]$ che indica la probabilità che il dato appartenga alla classe 1.

Diciamo che è stato osservato che la *sigmoid è meglio nell'output*. Perché? Eh funziona cosi a quanto pare lol.



Compilazione del modello

In questa sezione si definisce il **loss** e l'**optimizer**.

L'ottimizzatore sarebbe, praticamente, la *discesa del gradiente*. Ci sono vari ottimizzatori ed è molto buono per te stesso provare gli ottimizzatori e vedere quale funziona meglio per il tuo problema. Letteralmente il deep learning :) Per quanto riguarda la loss abbiamo visto nella scorsa sezione le funzioni di loss che conosciamo.

La **metrica** è quella che viene usata per valutare effettivamente il modello. In questo caso si usa l'**accuracy** che praticamente è la percentuale di classificazioni corrette.

4.3 Plotting del modello

E' un plotting molto base e non molto fancy, ma fa il suo lavoro diciamo

```
from keras.utils import plot_model
plot_model(model, show_shapes=True, show_layer_names=True)
```

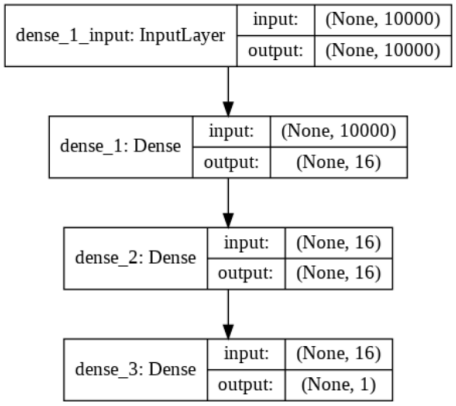


Figure 9: Plotting del modello.

4.4 Training del modello e valutazione

Entra in gioco il concetto di **validation set**.

4.4.1 Validation Set

Quando ho un modello, come faccio a sapere se la rete che sto allenando si sta allenando in modo corretto? Tutto questo sapendo che **non abbiamo accesso al test set**. Pensiamo al **training set** e pensiamo ad un altro **split** interno al training set:

- Training set parziale
- Validation set

Quindi se avessimo 50.000 di grandezza del dataset:

- 25.000 test set
- 25.training set
  - 10.000 validation set
  - 15.000 training set parziale

```
x_val = x_train[:10000]
partial_x_train = x_train[10000:]
y_val = y_train[:10000]
partial_y_train = y_train[10000:]
```

Praticamente è come se usassimo il validation come se fosse un test set. Quindi andremo a plottare 2 curve:

- La loss sulle epoche
- L'accuracy sulle epoche

La loss ci da un'idea di quanto il modello si sta allenando bene. Se ci sono problemi, la figura è strana e non segue un andamento corretto, si modifica. Ma la cosa importante è la **validation accuracy**, che DEVE essere crescere in modo monotono e deve avvicinarsi a 1 il più possibile.

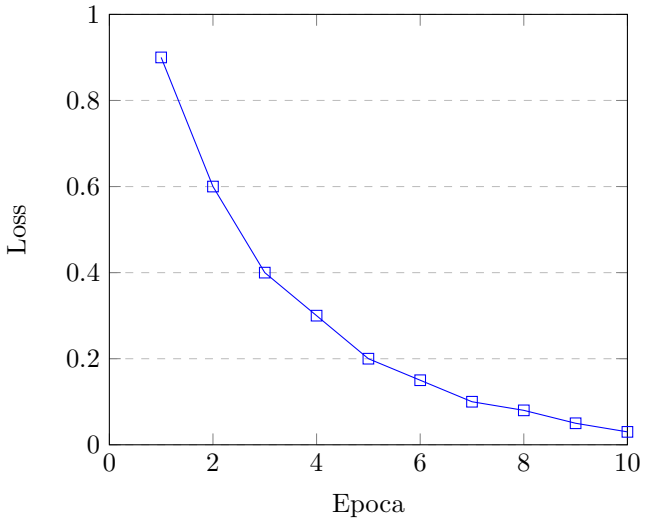


Figure 10: Grafico di una curva di loss che diminuisce all'aumentare delle epoche.

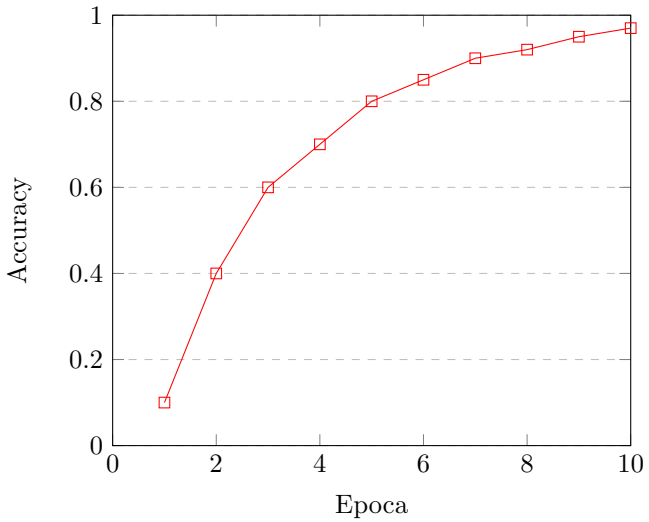


Figure 11: Grafico di una curva di accuracy che aumenta all'aumentare delle epoche.

Ma anche se avessimo queste curve inìcon questa conformazione, **ancora non ci dice niente**. Il validation è in un senso insensato per l'output. Ciò che importerà sarà il **test set** che non è stato usato per il training.

Training code

```
history = model.fit(partial_x_train,
            partial_y_train,
            epochs=20,
            batch_size=512,
            validation_data=(x_val, y_val))
```

Plotting delle curve

Notare che questo è un plotting molto base e non molto fancy, e in futuro utilizzeremo **Tensorboard** che aggiornerà in tempo reale i grafici

```
import matplotlib.pyplot as plt
loss = history.history['loss']
val_loss = history.history['val_loss']
epochs = range(1, len(loss) + 1)
# "bo" is for "blue dot"
plt.plot(epochs, loss, 'bo', label='Training loss')
# b is for "solid blue line"
plt.plot(epochs, val_loss, 'b', label='Validation loss')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```

Notare che *history* è un dizionario e si può accedere come indicato a riga *1 e 2* come accedere a loss e val_loss.

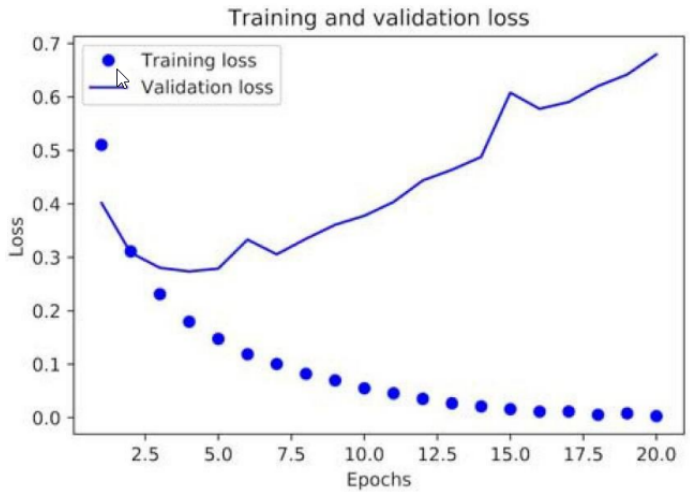


Figure 12: Rete sbagliata

Un grafico del genere ci mostra una rete che **funziona male**!.

```
plt.clf() # clear figure
acc = history_dict['binary_accuracy']
val_acc = history_dict['val_binary_accuracy']
plt.plot(epochs, acc, 'bo', label='Training acc')
plt.plot(epochs, val_acc, 'b', label='Validation acc')
plt.title('Training and validation accuracy')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()
plt.show()
```

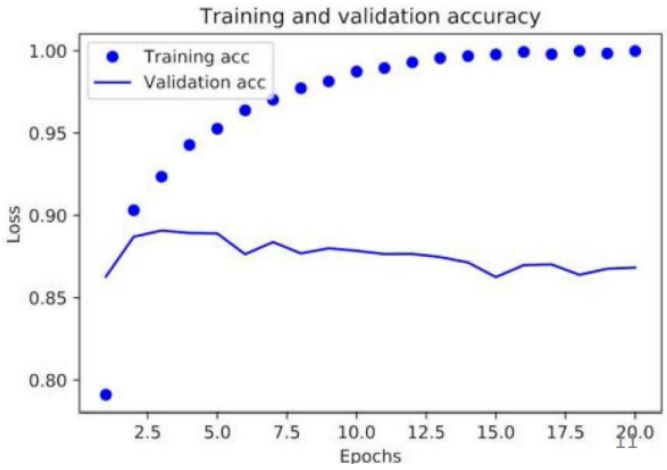


Figure 13: Overfitting

In questo caso la rete ha un problema di **overfitting!**.

4.5 Prediction

```
model.predict(x_test)
```

$array([[0.91966152], [0.86563045], [0.99936908], ..., [0.45731062], [0.0038014], [0.79525089]], dtype = float32$

4.6 Come risolvere problemi di accuracy bassa?

Se quando andiamo a fare la validazione della nostra rete, come sappiamo come modificare qualcosa per rendere la rete migliore? Ci sono alcuni approcci per farlo:

- Cambiare la topologia
- Diminuire le epoche
- Cambiare il numero di nodi per qualche layer
- FORSE cambiare anche il learning rate

Si noti che facendo questo vanno interpretati i grafici per capire come sta andando la rete. Quando cominciamo a vedere che la **validation accuracy NON STA SOLAMENTE CRESCENDO** ma ci sono dei punti in cui diminuisce, siamo sicuri che **c'è qualche problema di mezzo.** Bisogna capire anche su cosa lavorare. Se cambiando i dati spesso la **validation accuracy** mostra problemi, allora bisogna lavorare per quello e cercare di trovare un modo per fare in modo che il problema non si presenti.

RECAPPONE: Se notiamo che nella LOSS ci sono problemi, possiamo tunare il **learning rate** e altri parametri per cercare di risolvere questi problemi. Ma la cosa più importante è **la metrica**. Se notiamo che la **metrica non rispecchia un andamento di effettivo apprendimento**, capiamo che la rete non si sta comportando nel modo corretto e non sta funzionando.

4.7 Early Stopping

E' un metodo che consiste nel vedere quando **la loss** comincia ad avere dei comportamenti sospetti. Quando si **ferma** la rete in un dato momento, si impedisce alla rete di **overfittare** nella maggior parte dei casi.

5 Classificazione Multiclasse

Questa sezione sarà molto corta, poiché praticamente è la stessa cosa della classificazione binaria, ma il layer finale della rete non sarà composto da un solo nodo, ma da **un numero di nodi pari al numero di classi da classificare**.

5.1 Descrizione del dataset - Reuters

Il dataset che verrà utilizzato è il **Reuters Dataset**, che è un dataset di **news** che sono state classificate in **46 categorie**. Ogni categoria ha almeno 10 esempi nel training set.

Preprocessing dell'input

```
import numpy as np
def vectorize_sequences(sequences, dimension=10000):
    results = np.zeros((len(sequences), dimension))
    for i, sequence in enumerate(sequences):
        results[i, sequence] = 1.
    return results
# Our vectorized training data
x_train = vectorize_sequences(train_data)
# Our vectorized test data
x_test = vectorize_sequences(test_data)
```

5.1.1 Come processiamo l'output?

Nel deep learning, **gli output categorici** non vengono mai mappati ad una scala numerica. Si utilizza una tecnica che si chiama **one hot encoding**.

One Hot Encoding:

Consiste nel creare tante variabili numeriche quanti sono i valori della variabile categorica, e per quel dato example viene assegnato

- 1: se il valore della variabile categorica è quello

- 0: se il valore della variabile categorica non è quello, quindi in tutti gli altri casi

Da notare che il one hot encoding viene fatto sia per le labels di output, ma non è sbagliato farlo anche per gli attributi in alcuni casi.

```
def to_one_hot(labels, dimension=46):
    results = np.zeros((len(labels), dimension))
    for i, label in enumerate(labels):
        results[i, label] = 1.
    return results
# Our vectorized training labels
one_hot_train_labels = to_one_hot(train_labels)
# Our vectorized test labels
one_hot_test_labels = to_one_hot(test_labels)

#OPPURE ALLO STESSO MODO

from keras.utils.np_utils import to_categorical

one_hot_train_labels = to_categorical(train_labels)
one_hot_test_labels = to_categorical(test_labels)
```

Nota: nel one hot encoding DOBBIAMO AVERE 1 SOLO VALORE per il dato example, e tutti gli altri non devono essere attivi. La domanda è, quindi, **quale activation function usiamo?**

Immaginiamo questa situazione: Abbiamo y_1, y_2, y_3 che sono i valori di output. Vogliamo solamente uno dei tre. Se normalizzassimo i dati in questo modo:

- $p_1 = \frac{y_1}{y_1+y_2+y_3}$
- $p_2 = \frac{y_2}{y_1+y_2+y_3}$
- $p_3 = \frac{y_3}{y_1+y_2+y_3}$

E se sommiamo $p_1 + p_2 + p_3$ otteniamo 1. Quindi, in questo caso, possiamo usare la **softmax** come activation function.

5.1.2 Softmax activation function

Se prendiamo un vettore $\vec{y} = (y_1, y_2, y_3)$, la softmax è definita come:

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{3} e^{y_j}} \tag{12}$$

In output abbiamo: $\vec{p} = (p_1, p_2, p_3)$, dove p_i è la probabilità che il dato example appartenga alla classe i.

Nota: quando dividiamo qualcosa, abbiamo **sempre** un problema riguardo la stabilità numerica. La divisione è molto critica, poiché **potrebbe essere vicina a 0**. Sappiamo che se usiamo $e_1^y + e_2^y + e_3^y$ difficilmente si avvicina a 0.

```
    from keras import models
    from keras import layers
    model = models.Sequential()
    model.add(layers.Dense(64, activation='relu', input_shape
    =(10000,)))
    model.add(layers.Dense(64, activation='relu'))
    model.add(layers.Dense(46, activation='softmax'))
    model.compile(optimizer='rmsprop',
                  loss='categorical_crossentropy',
                  metrics=['accuracy'])
```

Nota: Dobbiamo avere un numero di nodi di output quanto il numero di classi, ma **altra cosa**, il numero di nodi interni deve essere sicuramente maggiore di 46, altrimenti ci sarebbe una perdita di informazioni.

Nota 2: La funzione di attivazione dell'ultimo layer è una **softmax**.

Validation set e Training set

```

x_val = x_train[:1000]
partial_x_train = x_train[1000:]
y_val = one_hot_train_labels[:1000]
partial_y_train = one_hot_train_labels[1000:]
history = model.fit(partial_x_train,
                    partial_y_train,
                    epochs=20,
                    batch_size=512,
                    validation_data=(x_val, y_val))
```

Loss

```
import matplotlib.pyplot as plt
loss = history.history['loss']
val_loss = history.history['val_loss']
epochs = range(1, len(loss) + 1)
plt.plot(epochs, loss, 'bo', label='Training loss')
plt.plot(epochs, val_loss, 'b', label='Validation loss')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```

Accuracy

```
plt.clf() # clear figure
acc = history.history['acc']
val_acc = history.history['val_acc']
plt.plot(epochs, acc, 'bo', label='Training acc')
plt.plot(epochs, val_acc, 'b', label='Validation acc')
plt.title('Training and validation accuracy')
plt.xlabel('Epochs')
plt.ylabel('Acc')
plt.legend()
plt.show()
```

Nota: L'accuracy è importante dipendentemente dall'utilizzo che bisogna farne. Ad esempio, *in ambito medico* vogliamo **minimizzare i falsi negativi**. Questo implica che la misura che usiamo dipende dall'applicazione che se ne fa.

6 Regressione

Parlando di **regressione**, si ha un problema impostato allo stesso modo di quelli di classificazione binaria o multiclasse, ma l'*output* è un **nuemro reale**.

Nota: quando all'interno della nostra rete abbiamo dei valori categorici, sappiamo che per gestirli utilizziamo la tecnica del **one hot encoding**. Ma quando abbiamo valori numerici, come gestiamo questi dati? Potremmo avere dati che hanno *scale diverse, unità di misura diverse*, ecc.... Questi rendono abbastanza complicato il lavoro della rete. Per questo motivo, è necessario **normalizzare** i dati.

6.1 Boston Housing Price

Il dataset **Boston Housing Price** è un dataset che contiene 506 esempi di case nella zona di Boston. Ogni esempio è composto da 13 *feature* che descrivono la casa e il prezzo della casa. Questo dataset è stato utilizzato per la prima volta nel 1978, ma è ancora utilizzato per testare i modelli di regressione, ed è proprio quello che faremo noi.

```
1  from keras.datasets import boston_housing
2  (train_data, train_targets), (test_data, test_targets) =
   boston_housing.load_data()
```

Listing 2: Caricamento del dataset

6.1.1 Normalizzare i dati

La normalizzazione consiste nel portare i valori numerici di un dataset tutti sulla stessa scala. Abbiamo due modi per fare questo processo di normalizzazione:

- MinMax Normalization: La MinMax Normalization è una tecnica di normalizzazione dei dati che consiste nel portare tutti i valori di un dataset su una scala compresa tra 0 e 1. Questa tecnica è utile quando i dati hanno scale diverse e si vuole portarli tutti sulla stessa scala per facilitare l'elaborazione da parte della rete neurale. Per applicare la MinMax Normalization, si utilizza la seguente formula per ogni valore del dataset: Viene usata ma **non è proprio approrpiata.**

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{13}$$

- Normalizzazione Statistica: Si utilizzano la **media** e la **deviazione standard**. Questa tecnica è utile quando i dati hanno scale diverse e si vuole portarli tutti sulla stessa scala per facilitare l'elaborazione da parte della rete neurale.

```
1          mean = train_data.mean(axis=0)
2          train_data -= mean
3          std = train_data.std(axis=0)
4          train_data /= std
5
6          test_data -= mean
7          test_data /= std
8
```

Questo porta un **centramento** intorno allo 0. **Nota importante:** Da notare le righe *6 e 7* che applicano la normalizzazione dei dati **anche sul test set**. Questa cosa si fa? La risposta è **NON ABBIAMO MAI I TEST DATA.**

Praticamente stiamo facendo l'assunzione che i dati vengano dalla **stessa distribuzione**. Poiché i risultati del test set probabilmente saranno su una scala diversa rispetto a quelli che abbiamo dal training set e i risultati che otteniamo dal nostro modello potrebbero essere su una scala diversa rispetto a quella del training. Quindi, questo va fatto solo se si ha la certezza che i dati provengano dalla stessa distribuzione.

6.1.2 Costruzione della rete

```
1  def build_model():
2      # Because we will need to instantiate
3      # the same model multiple times,
4      # we use a function to construct it.
5      model = models.Sequential()
6      model.add(layers.Dense(64, activation='relu',
7                  input_shape=(train_data.shape[1],)))
8      model.add(layers.Dense(64, activation='relu'))
9      model.add(layers.Dense(1))
10     model.compile(optimizer='rmsprop', loss='mse', metrics=['mae'])
11     return model
```

Piccole note su questo codice:

1. L'ultimo livello ha un solo nodo di output
2. Non ha una funzione di attivazione
3. Sta usando la metrica **MAE**: Mean Absolute Error e la loss **MSE**: Mean Squared Error

6.1.3 Validation con pochi data point

Spiegazione al volo della K-Fold validation: La K-Fold validation è una tecnica di validazione che consiste nel dividere il dataset in **K parti** e utilizzare una di queste parti come **validation set** e le altre come **training set**.

6.1.4 Visualizzare i risultati

```
average_mae_history = [
    np.mean([x[i] for x in all_mae_histories]) for i in range(
    num_epochs)]

import matplotlib.pyplot as plt

plt.plot(range(1, len(average_mae_history) + 1),
    average_mae_history)
plt.xlabel('Epochs')
plt.ylabel('Validation MAE')
plt.show()
```

Nota: Solitamente i primi punti in una task di regressione **sono fuori scala**. Conviene scartarli e non contarli nella big picture finale.

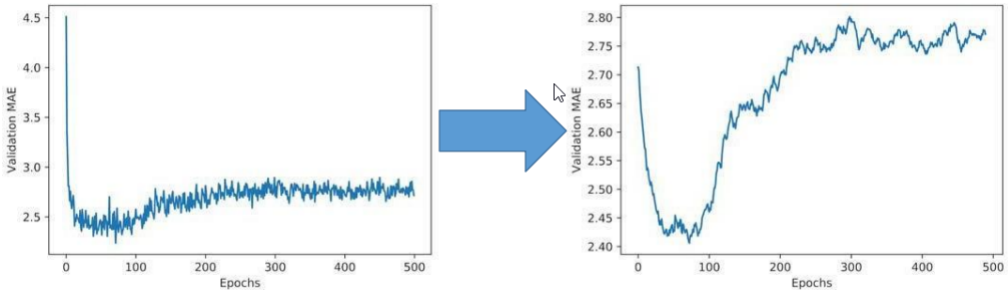


Figure 14: Grafico MAE

6.2 Overfitting

Formalizziamo al volo il problema:

$$\min_{w}[loss(y, NN(x|w))] \tag{14}$$

Cioe vogliamo minimizzare in funzione dei pesi la funzione che ha come parametri la **vera y** e la **y predictata** usando il nostro input **x** e i pesi **w**.

La tecnica che vediamo è la **regolarizzazione**.

6.2.1 Regolarizzazione

La regolarizzazione è una tecnica che non fa altro che **modificare la funzione che vogliamo approssimare**. Questo rende i pesi della rete più piccoli, il che rende la distribuzione dei valori *più regolare.*

$$\min_{w}[loss(y, NN(x|w)) + R(w)] \tag{15}$$

Abbiamo due tipi di regolarizzazione:

- L1 Regularization: *vecchia loss function* $+ \lambda \sum_i |w_i|$, cioè aggiunge la somma dei valori assoluti dei pesi.
- L2 Regularization: *vecchia loss function* $+ \lambda \sum_i w_i^2$, cioè aggiunge la somma dei valori dei pesi al quadrato.

Queste due tecniche rendono *il modello più sensibile la noise e alla varianza dei dati.*

- L2: Rende i pesi più piccoli
- L1: Rende i pesi più sparsi (più 0, in pratica)

```
from keras import regularizers
l2_model = models.Sequential()
l2_model.add(layers.Dense(8, kernel_regularizer=regularizers.l2
    (0.001),
    activation='relu'
    , input_shape=(10000,)))
l2_model.add(layers.Dense(8, kernel_regularizer=regularizers.l2
    (0.001),
            activation='relu'))
l2_model.add(layers.Dense(1, activation='sigmoid'))
```

Nota: si può fare una combinazione tra L1 e L2.

```
from keras import regularizers
# L1 regularization
regularizers.l1(0.001)
# L1 and L2 regularization at the same time
regularizers.l1_l2(l1=0.001, l2=0.001)
```

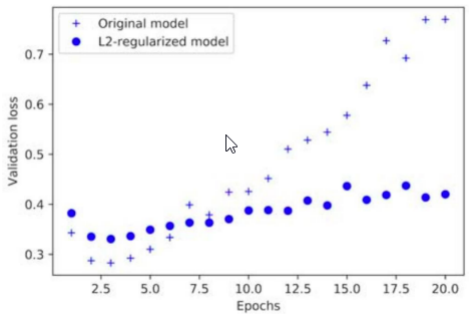


Figure 15: Grafico Regularization L2

Quindi, alla fine, **early stopping** e **diminuire grandezza rete** non vengono fatti con leggerezza e non sono spesso soluzioni applicabili.

6.3 Dropout

In teoria, questa tecnica prevede il rimuovere **durante il training** alcuni nodi della rete in modo casuale seguendo una specifica distribuzione (si setta il valore di 0). L'effetto è quello di rendere la rete più confusa, e quindi più robusta. **Quando si fa il testing** si va ad utilizzare l'intera rete senza troncamenti di connessioni.

Per quale motivo funziona? La rispota è **BOH**. Il fatto è che **funziona** ed è uno standard. Quindi praticamente in ogni NN si utilizza la tecnica del *dropout*.

```
dpt_model = models.Sequential()
dpt_model.add(layers.Dense(16, activation='relu', input_shape
    =(10000,)))
#Probabilita di rendere 0 un nodo del layer con probabilita 0.5
dpt_model.add(layers.Dropout(0.5))
dpt_model.add(layers.Dense(16, activation='relu'))
dpt_model.add(layers.Dropout(0.5))
dpt_model.add(layers.Dense(1, activation='sigmoid'))
dpt_model.compile(optimizer='rmsprop',
    loss='binary_crossentropy',
    metrics=['acc'])
```

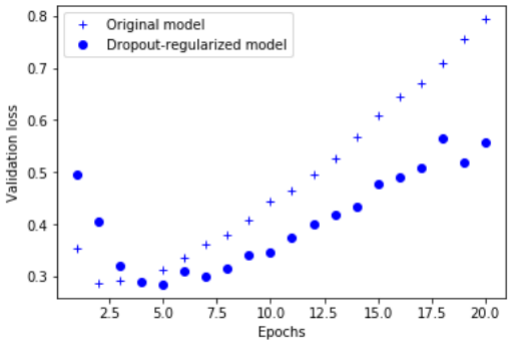


Figure 16: Grafico Dropout

Capitoli di laboratorio

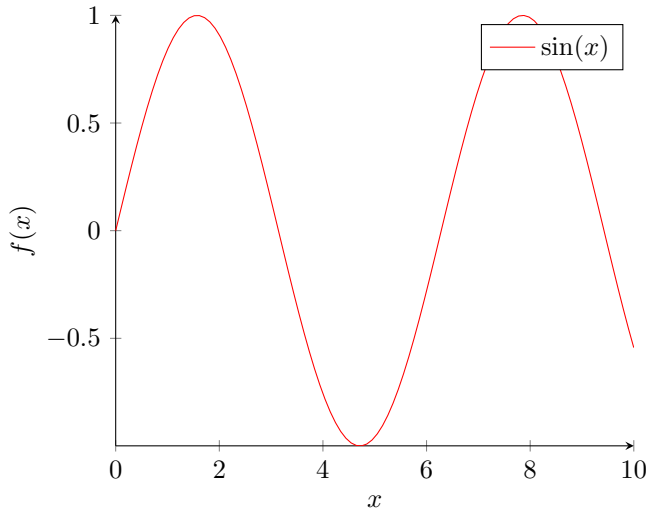
7 Lab: Introduzione Python

7.1 MatPlotLib

7.1.1 Plots

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 x = np.linspace(0, 10, 100)
5 plt.plot(x, np.sin(x))
6 plt.show()
```

Non c'è molto da dire, il codice è autoesplicativo. La funzione `plot` prende in input due array, uno per l'asse delle ascisse e uno per l'asse delle ordinate. In questo caso, `x` è un array di 100 punti equidistanti tra 0 e 10, mentre `np.sin(x)` è un array di 100 punti che rappresentano il seno dei punti di `x`. La funzione `show` mostra il grafico.

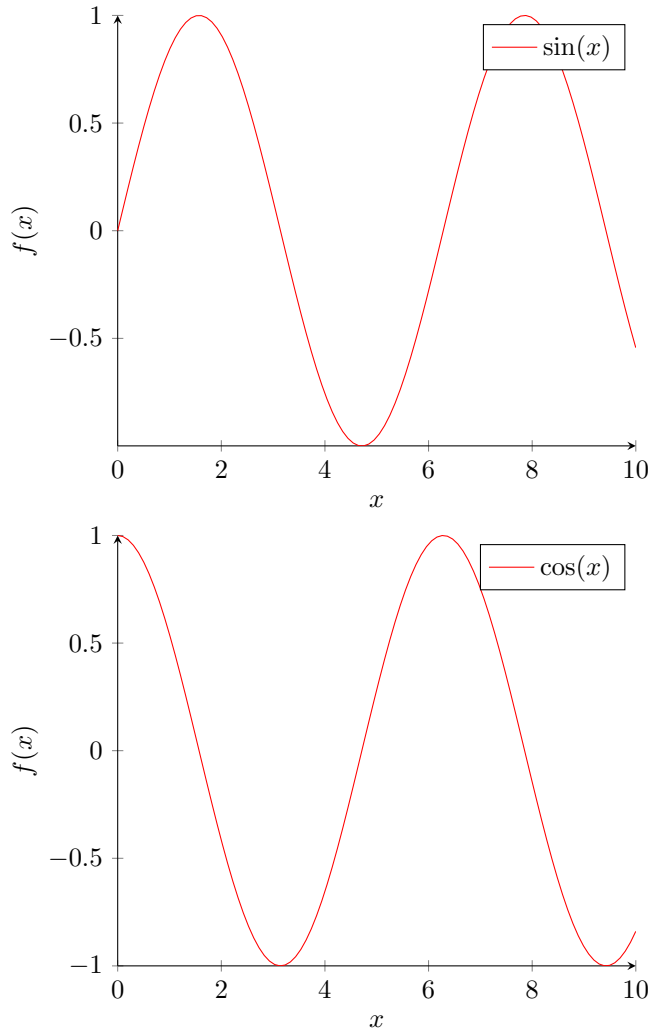


7.1.2 Sub-plots

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 x = np.linspace(0, 10, 100)
5 plt.subplot(2, 1, 1)
6 plt.plot(x, np.sin(x))
```

```
plt.subplot(2, 1, 2)
plt.plot(x, np.cos(x))
plt.show()
```

La funzione `subplot` prende in input tre parametri: il numero di righe, il numero di colonne e l'indice del subplot corrente. Nel caso di questo esempio, il subplot corrente è il primo, quindi viene mostrato il grafico del seno. Poi viene mostrato il secondo subplot, che è quello del coseno.



C'è anche qui poco da dire, il codice è autoesplicativo.

7.2 NumPy

La libreria NumPy è una libreria per Python che permette di lavorare con array multidimensionali. Per importare la libreria, basta scrivere `import numpy as`

`np`.

Mi rompo le palle in maniera assurda di scrivere tutti gli esempi. Quindi questo capitolo penso sia abbastanza inutile.

E' possibile trovare il codice di **numPy** a questo link: `www.ciao.it`

8 Lab: Reti Neurali da zero

8.1 Introduzione

Partiamo dicendo una cosa molto importante: **per quale motivo usiamo la backpropagation?**

$$\frac{y-b}{x} = w \tag{16}$$

Ma possiamo scriverlo come:

$$(y-b) \cdot x^{-1} = w \tag{17}$$

Ora però, se consideriamo:

- y il vettore risultante
- w la matrice dei pesi
- x il vettore di input

Calcolare l'inversa di **x** non è una cosa cosi poco costosa, anzi. Per questo motivo utilizziamo il training delle reti come la backpropagation e la discesa del gradiente.

Ora, andiamo più a fondo. Facciamo un esempio più pratico.

8.2 Esempio pratico

Immaginiamo di avere un dataset di 2 features e una label da identificate.

x_1	x_2	y
1	2	0
2	3	0
3	4	0
4	5	0
5	6	0
6	7	1
7	8	1
8	9	1
9	10	1
10	11	1

Table 1: Dataset di esempio

Ogni livello di una rete neurale può essere rappresentato attraverso le **matrici**.

Ma passiamo alla definizione formale:

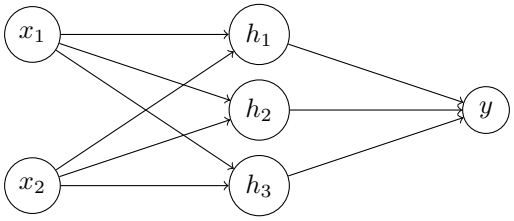


Figure 17: Rete neurale di esempio

Definizione Rete Neurale: Una rete neurale è una tupla:

$$NN = \{g, l, o, i, fpp\} \tag{18}$$

con:

- g: il grafico
- l: la funzione loss
- o: l'ottimizzatore
- i: l'inizializzatore
- fpp: la fix point procedure

Nota 1: l'ottimizzatore intende in quale modo si performa la **discesa del gradiente.**

Nota 2: l'inizializzatore è la funzione che inizializza i pesi della rete neurale. Ci possono essere diversi algoritmi per gli inizializzatori, ma non c'è modo di sapere quale funziona meglio, poiché non si può sapere nelle reti neurali.

8.2.1 Il grafo

Il solito grafo che abbiamo visto in precedenza è un grafo aciclico diretto. Ha dei nodi che sono i percettroni, che hanno degli archi composti dai pesi, che hanno una funzione di attivazione, che prendono in input un valore e ne sputano fuori uno chiamando la funzione di attivazione sull'input.

Nota: la funzione di attivazione è una funzione non lineare.

8.2.2 La funnzione loss

L'obiettivo della funzione loss è quello di misurare la distanza tra il valore predetto e il valore reale.

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \tag{19}$$

8.2.3 L'ottimizzatore

L'ottimizzatore ci permette di trovare una soluzione ottimale **non ottima**, cioé trovare **il minimo della funzione di loss**. La tecnica che si usa è quella delle **discesa del gradiente**.

8.2.4 Discesa del gradiente

La discesa del gradiente sfrutta un parametro chiamato **learning rate** che permette di capire quanto la discesa del gradiente deve essere veloce. Se il learning rate è troppo alto, la discesa del gradiente potrebbe non convergere, se è troppo basso, la discesa del gradiente potrebbe convergere troppo lentamente.

Nota: il learning rate è un ottimizzatore *molto naive*.

8.2.5 Inizializzatore

Il metodo di inizializzazione può cambiare di molto il risultato della rete neurale. In pratica assegna un valore ai *pesi* e ai *bias*. Alcuni metodi sono:

- Inizializzazione a 0: Questo ha alcuni problemi, perché non si ha diversificazione tra i nodi e i nodi nascosti diventano simmetrici.
- inizializzazione costante: stessi problemi della precedente
- Inizializzazione Random: Si fa seguendo una distribuzione uniforme oppure una distribuzione normale.

8.2.6 Fix Point Procedure

La procedura per il training di una rete neurale è un processo che si basa su alcuni steps:

```
1
2  net CustomNeuralNetwork(...)
3  initialize_weights_and_biases (net)
4  optimizer = myOptimizer(...)
5  loss_function = myLoss Function(...)
6  epochs = ... # the number of dataset scans
7  history = [] # a list containing the loss evolution
8  for epoch in range(epochs):
9      optimizer.reset() # it may have an internal status
10     loss = loss_fuction (out_target, net(input)) back_propagation (
       loss, optimizer, net)
11     history.append(loss)
```

8.3 Esempio da zero

8.3.1 La funzione Sigmoid

Spendiamo qualche parola sulla funzione sigmoid, che è una funzione molto importante per le reti neurali.

In particolare, la funzione avrà un valore di attivazione compreso tra 0 e 1 e, in particolare, quando l'input è 0, la funzione ha valore 0.5. Se ha un valore basso, sarà zero e chiaramente se sarà alto avrà valore 1.

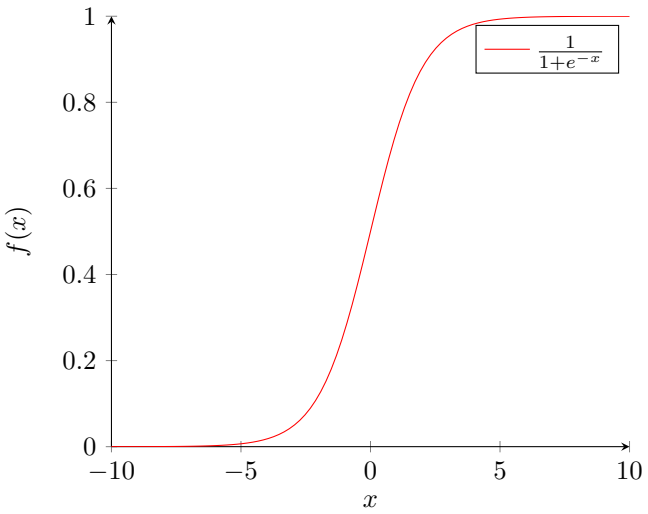


Figure 18: Funzione Sigmoid

Detto questo, **per quale motivo è importante?** Se pensiamo all'*inizializzazione*, che tipi di valori è meglio avere? Avere tutti i valori a zero porterebbe a valori simmetrici e quindi a problemi di convergenza. Vogliamo dei valori di pesi che siano **distribuiti uniformemente intorno allo 0.**

8.4 Tangente Iperbolica TanH

La funzione TanH è una funzione che ha un valore di attivazione compreso tra -1 e 1 e, in particolare, quando l'input è 0, la funzione ha valore 0. Se ha un valore basso, sarà -1 e chiaramente se sarà alto avrà valore 1.

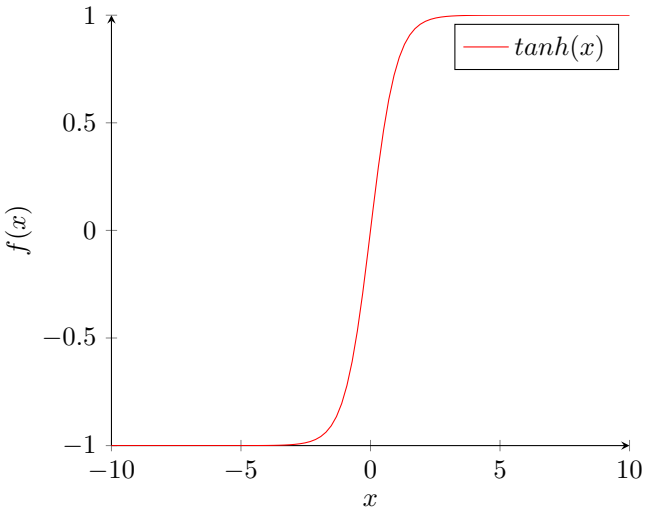


Figure 19: Funzione TanH

8.5 ReLu

La funzione ReLu è una funzione che ha un valore di attivazione pari a 0 quando l'input è negativo e ha un valore di attivazione pari all'input quando l'input è positivo.

Parole di Adornetto: E' buona? Si. E' stabile? Si. Perché? Boh.

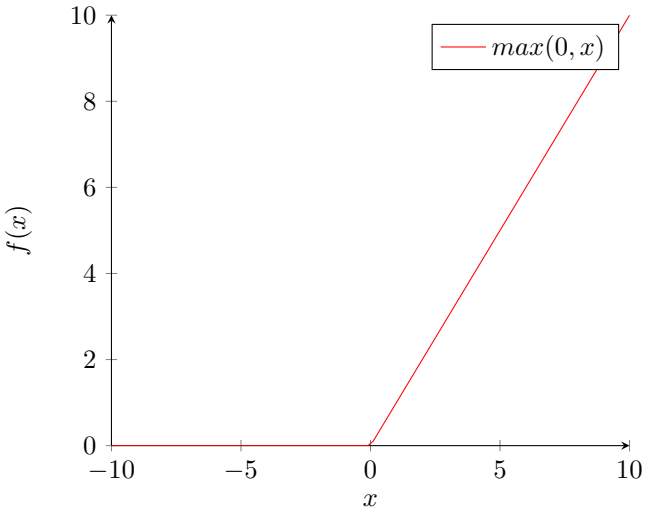


Figure 20: Funzione ReLu

8.6 Scalare i valori

Skip molto avanti riguardo l'esempio, ma si parla di scalare i dati.

Lo scaling dei dati è un'operazione importante nel machine learning perché i dati possono essere su scale diverse e questo può causare problemi durante l'allenamento del modello. Ad esempio, se abbiamo due variabili di input, una che varia da 0 a 1 e l'altra che varia da 0 a 1000, la seconda variabile avrà un impatto molto maggiore sull'output del modello rispetto alla prima. Ciò può portare a problemi come l'overfitting, in cui il modello si adatta troppo ai dati di addestramento e non generalizza bene sui dati di test.

Per risolvere questo problema, si utilizzano tecniche di scaling dei dati per portare tutte le variabili su una scala comune. Ci sono diverse tecniche di scaling, come la normalizzazione e la standardizzazione, che possono essere utilizzate a seconda del tipo di dati e del modello utilizzato.

Le funzioni di attivazione, come la funzione ReLU, possono anche essere influenzate dalla scala dei dati di input. Se i dati non sono scalati correttamente, la funzione di attivazione potrebbe produrre valori che non permettono un allenamento corretto del modello. Lo scopo dello scaling dei dati è quello di mantenere i dati intorno allo 0, in modo che le funzioni di attivazione possano produrre valori che permettono un allenamento corretto del modello.

- Standardizzazione: La standardizzazione è una tecnica di scaling dei dati che assume che i dati siano distribuiti normalmente all'interno di ogni feature e li scala in modo che la distribuzione abbia una media uguale a 0 e una deviazione standard uguale a 1. Questa tecnica funziona bene

quando i dati hanno una distribuzione normale, ma non funziona bene se i dati hanno una distribuzione non normale.

- Normalizzazione: La normalizzazione è una tecnica di scaling dei dati che scala i valori di ogni feature in modo che siano compresi tra 0 e 1. Questa tecnica funziona bene quando i valori di input non hanno una distribuzione normale. Ad esempio, se i valori di input sono compresi tra 0 e 1000, la normalizzazione li porterà su una scala compresa tra 0 e 1.

8.7 Plottare la loss

La funzione loss è una misura dell'errore del modello durante l'allenamento. L'obiettivo dell'allenamento è quello di minimizzare la funzione loss, ovvero di ridurre l'errore del modello. Durante l'allenamento, il modello viene eseguito su un set di dati di addestramento e la funzione loss viene calcolata per ogni esempio di addestramento. L'errore totale del modello è la somma di tutte le funzioni loss per ogni esempio di addestramento.

L'allenamento del modello avviene in epoche, ovvero in cicli di esecuzione su tutti i dati di addestramento. L'obiettivo è che la funzione loss diminuisca ad ogni epoca, ovvero che l'errore del modello diminuisca man mano che il modello viene addestrato su più dati. Se la funzione loss non diminuisce ad ogni epoca, significa che il modello non sta imparando abbastanza dai dati di addestramento e potrebbe essere necessario modificare l'architettura del modello o i parametri di addestramento.

Se la funzione di loss arriva a 0, vuol dire che **siamo in minimo globale**, che non è comune.

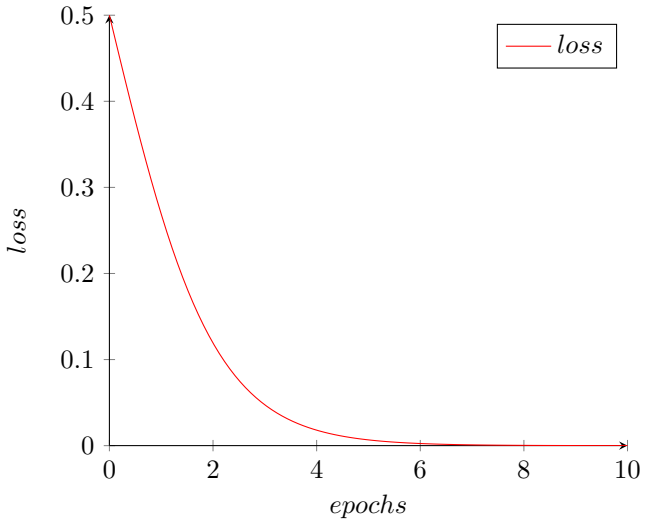


Figure 21: Funzione loss

Se invece finiami con un grafico che arriva fino a 0.1 e poi si stabilizza, potremmo essere in un **minimo locale**, oppure semplicemente la topologia della Rete permette di avere questo risultato come massimo risultato.