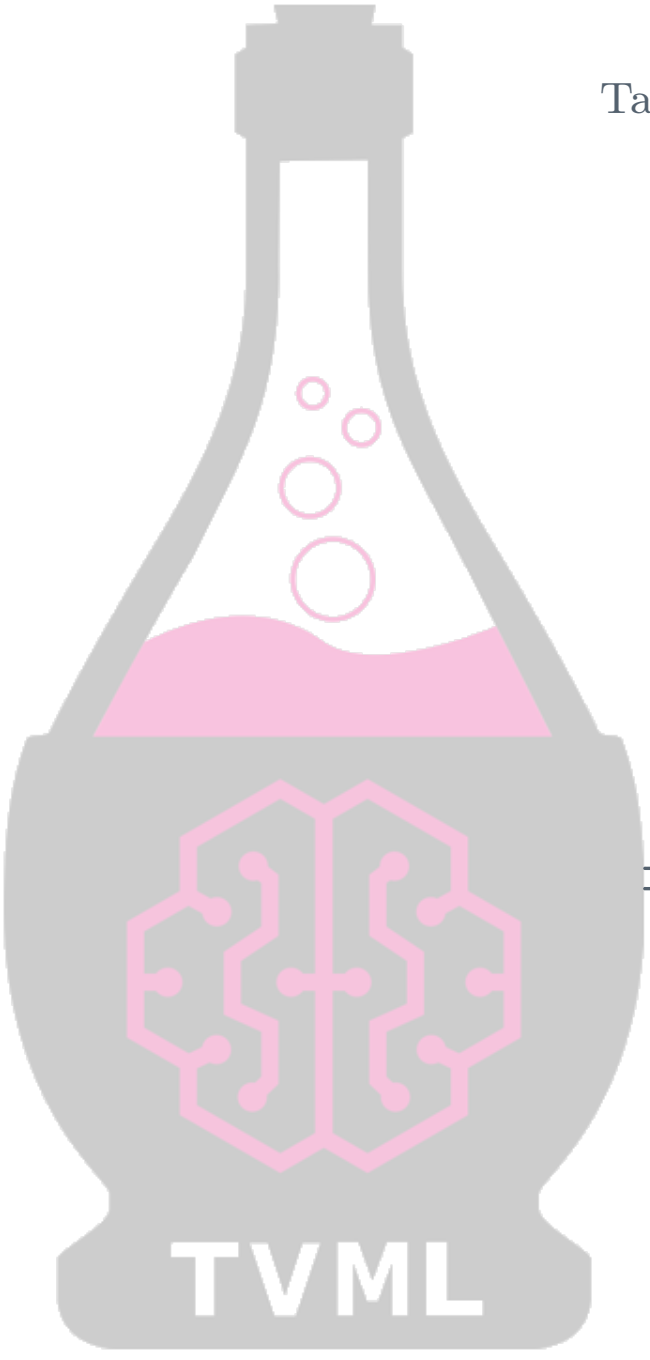



Tarallucci, Vino e Machine Learning
"Il paper della buonanotte"™

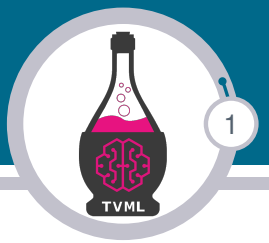
# Case Study

## sviluppo di un Clamidia-Detector

Fabio Mardero
fabio.mardero@gmail.com
github.com/fmardero

9 maggio 2019

# Indice

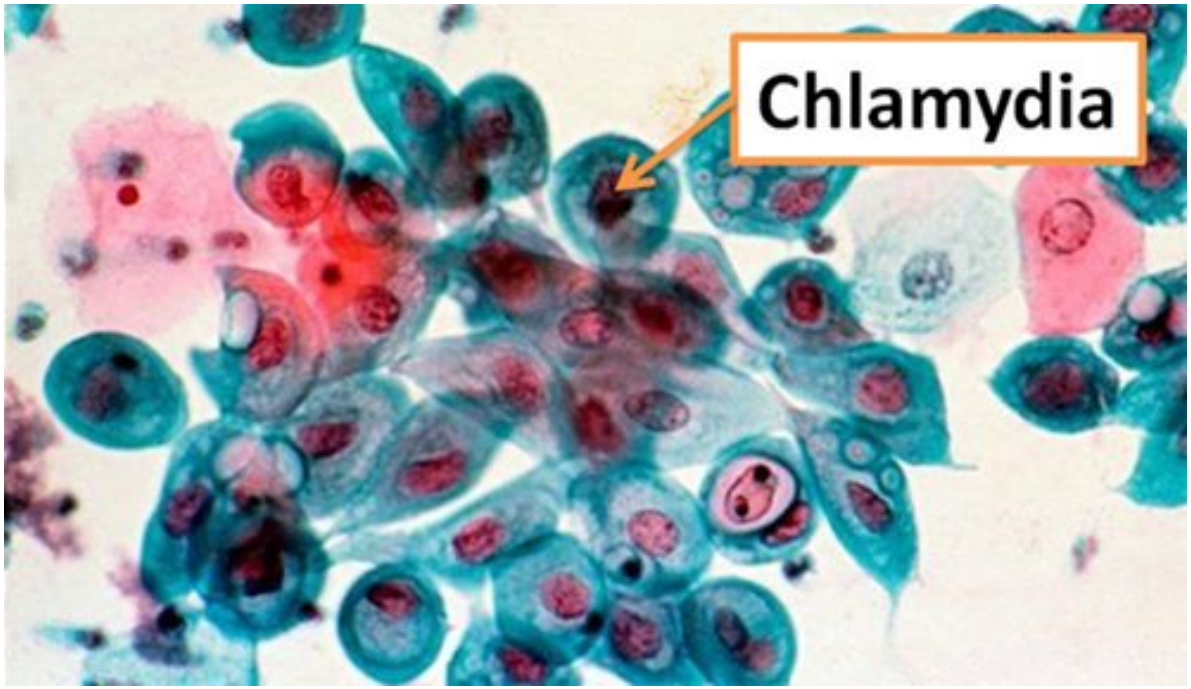


# Il progetto

# Il problema



La clamidia è un'infezione batterica venerea e generalmente asintomatica. Il metodo più attendibile per identificare la malattia consiste nell'analizzare un campione di sangue della persona.



# L'obiettivo



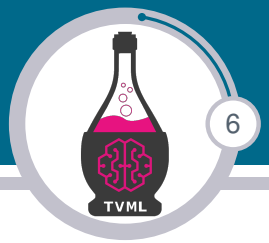
Una clinica medica commissiona al gruppo TVML di identificare, sulla base dell'immagine di un campione di sangue, se la persona è infetta o meno da clamidia. Al momento di lancio del progetto, l'equipe medica ha raccolto duemila immagini ($\mathcal{D}_0 = \{2e3\}$), classificandole:

P: "clamidia"
N: "non clamidia"

**NOTA BENE:** in quest'analisi si trascura la difficoltà reale del problema.

# Un primo modello

# Un primo modello





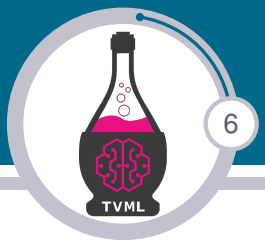

- Che modello di machine learning conviene utilizzare?
- Possibili architetture?

# Un primo modello



- Che modello di machine learning conviene utilizzare?
- Possibili architetture?
- Tipo di allenamento? Tecniche?

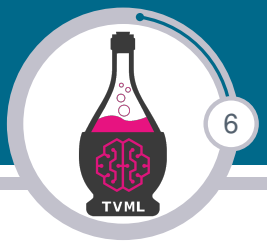
# Un primo modello






- Che modello di machine learning conviene utilizzare?
- Possibili architetture?
- Tipo di allenamento? Tecniche?
- Proporzioni train&valid&test?

# Un primo modello

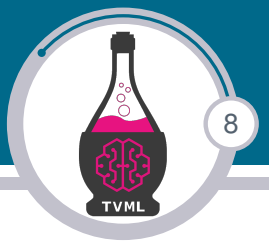





- Che modello di machine learning conviene utilizzare?
- Possibili architetture?
- Tipo di allenamento? Tecniche?
- Proporzioni train&valid&test?
- Metriche di allenamento e di supervisionamento degli errori?

# Lavori pregressi

# Lavori pregressi: Dati simili






Su Kaggle sono disponibili cinquanta mila immagini $\mathcal{D}_1 = \{5e4\}$ relative a campioni di sangue infetti da una di 100 malattie. Ogni dato è opportunamente etichettato con l'infezione corrispondente. Le immagini appaiono simili a quelle in $\mathcal{D}_0$.

- Posso usare questi dati?

# Lavori pregressi: Dati simili



Su Kaggle sono disponibili cinquanta mila immagini $\mathcal{D}_1 = \{5e4\}$ relative a campioni di sangue infetti da una di 100 malattie. Ogni dato è opportunamente etichettato con l'infezione corrispondente. Le immagini appaiono simili a quelle in $\mathcal{D}_0$.

- Posso usare questi dati?
- Proporzioni train&valid&test?

# Lavori pregressi: Modello

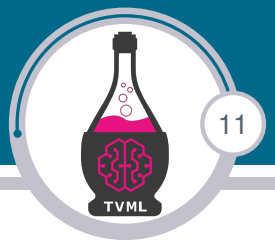


Tra i kernel della competizione è possibile scaricare un modello allenato su $\mathcal{D}_1$ che fornisce previsioni estremamente accurate rispetto alle 100 classi.

- Posso utilizzare il modello?
- Come?

# Tentativi di training

# Previsioni accuratissime






**TENTATIVO 1**

Il modello di clamidia-prediction è allenato per la prima volta. Si ottiene un'accuratezza del 98% sul test set.

- Il risultato è attendibile?

# Previsioni accuratissime



**TENTATIVO 1**

Il modello di clamidia-prediction è allenato per la prima volta. Si ottiene un'accuratezza del 98% sul test set.

- Il risultato è attendibile?
- Eventuali controlli?

# Previsioni accuratissime



**TENTATIVO 1**

Il modello di clamidia-prediction è allenato per la prima volta. Si ottiene un'accuratezza del 98% sul test set.

- Il risultato è attendibile?
- Eventuali controlli?
- Metodi di bilanciamento delle classi?

# Migliorare il modello



**TENTATIVO 2**

Allenato nuovamente, il modello fornisce previsioni con un'accuratezza del 65% sul test set.

- Risulta possibile migliorare le previsioni senza acquisire nuovi dati?

# Migliorare il modello



**TENTATIVO 2**

Allenato nuovamente, il modello fornisce previsioni con un'accuratezza del 65% sul test set.

- Risulta possibile migliorare le previsioni senza acquisire nuovi dati?
- In quali casi le tecniche di data augumentation potrebbero peggiorare le previsioni?

# Unlabeled Data

# Dati non classificati



Grazie alla data augumentation, l'accuratezza del modello sul test set è pari a 80%. Nel frattempo la clinica ha collezionato $\mathcal{D}_2 = \{1e4\}$ diecimila immagini di campioni di sangue di persone che si sono sottoposte al test per la clamidia. A seguito di un errore informatico, l'esito delle valutazioni sui campioni è stato perso.

- Si possono utilizzare comunque questi dati?
- Come?

# Valutazione dei risultati

# Confusion Matrix



Con la nuova fase di training, l'accuratezza del modello sul test set è pari a 85%. Di seguito la relativa confusion matrix.

| | | TRUE P | TRUE N |
|---|---|---|---|
| PRED | P | 40% | 1% |
| | N | 14% | 45% |

- I risultati sono soddisfacenti? Posso mettere in produzione il modello?

# Confusion Matrix



Con la nuova fase di training, l'accuratezza del modello sul test set è pari a 85%. Di seguito la relativa confusion matrix.

| | | TRUE P | TRUE N |
|---|---|---|---|
| PRED | P | 40% | 1% |
| | N | 14% | 45% |

- I risultati sono soddisfacenti? Posso mettere in produzione il modello?
- Soluzioni all'eventuale problema?

# Confusion Matrix



Risolto il problema si ottiene la seguente confusion matrix.

| | | TRUE P | TRUE N |
|---|---|---|---|
| PRED | P | 40% | 12% |
| | N | 3% | 45% |

- Qual è l'accuratezza del modello?

# Confusion Matrix

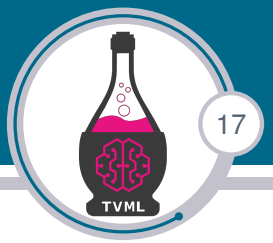


Risolto il problema si ottiene la seguente confusion matrix.

| | | TRUE P | TRUE N |
|---|---|---|---|
| PRED | P | 40% | 12% |
| | N | 3% | 45% |

- Qual è l'accuratezza del modello?
- Qual è la precisione del modello?

# Confusion Matrix



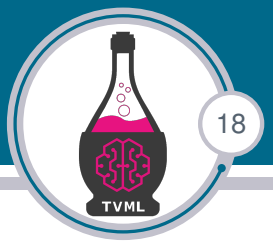



Risolto il problema si ottiene la seguente confusion matrix.

| | | TRUE P | TRUE N |
|---|---|---|---|
| PRED | P | 40% | 12% |
| | N | 3% | 45% |

- Qual è l'accuratezza del modello?
- Qual è la precisione del modello?
- Qual è il recall del modello?

# Attendibilità delle previsioni

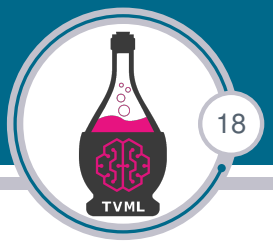





Il modello è preciso al 93%. Un medico, in condizioni normali, restituisce diagnosi con una precisione del 89%.

- Il modello ha ottenuto buone performance?

# Attendibilità delle previsioni

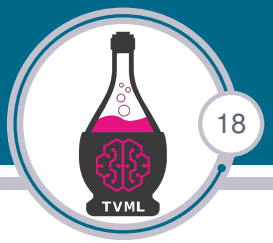





Il modello è preciso al 93%. Un medico, in condizioni normali, restituisce diagnosi con una precisione del 89%.

- Il modello ha ottenuto buone performance?
- I risultati sono attendibili?

# Attendibilità delle previsioni

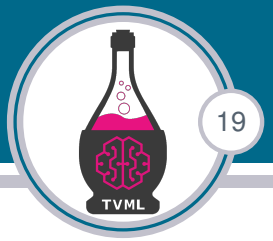





Il modello è preciso al 93%. Un medico, in condizioni normali, restituisce diagnosi con una precisione del 89%.

- Il modello ha ottenuto buone performance?
- I risultati sono attendibili?
- Com'è possibile ottimizzare il problema di labeling?

# Apprendimento del modello






La clinica medica rimane stupita da una precisione così elevata per un computer. Vuole capire cosa il modello valuta per fornire una data previsione.

- Tecniche?

Grazie dell’attenzione!