Alma Mater Studiorum - Università di Bologna
Scuola di Scienze

Pipeline per il Machine Learning: Analisi e orchestrazione di workflow in Demand forecasting & Radiology

Luca Genova

Relatore:
Chiar.mo Prof.
Maurizio Gabbrielli

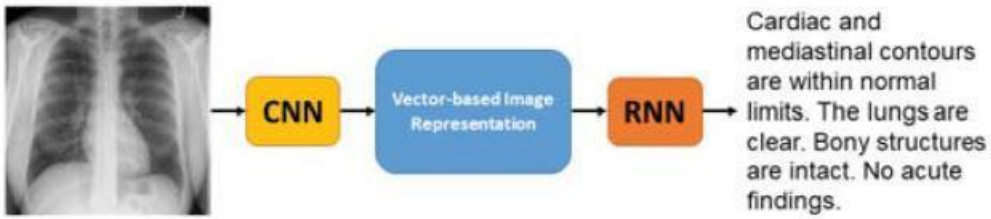
Correlatore:
Dott. Stefano Pio Zingaro
Dott. Saverio Giallorenzo

Corso di Laurea Triennale in Informatica
A.A. 2020/2021 - Sessione II

Introduzione

- Possibilità di automatizzare delle pipeline per il machine learning
- Trovare dei framework integrabili
- Strategia bottom-up analizzando due casi di studio specifici:
  - AI in radiology
  - Demand forecasting

AI in radiology



Workflow:

- Raccolta dei dati
- Cura dei dati
- Creazione del modello
- Formazione del modello
- Analisi e metriche
- Distribuzione

AI in radiology

Raccolta e cura dei dati:

- Quantità e qualità sono fondamentali
- Il problema della disponibilità
- Passaggi applicabili:
  - Etichettatura dei dati
  - Controllo qualità
  - Esclusione dei dati

Creazione e formazione del modello:

- CNN per l'estrazione delle caratteristiche dall'immagine
- RNN (LSTM) per l'estrazione del testo
- Set di allenamento D = (I, y) e il modello viene addestrato al fine di massimizzare un modello probabilistico

AI in radiology - Analisi

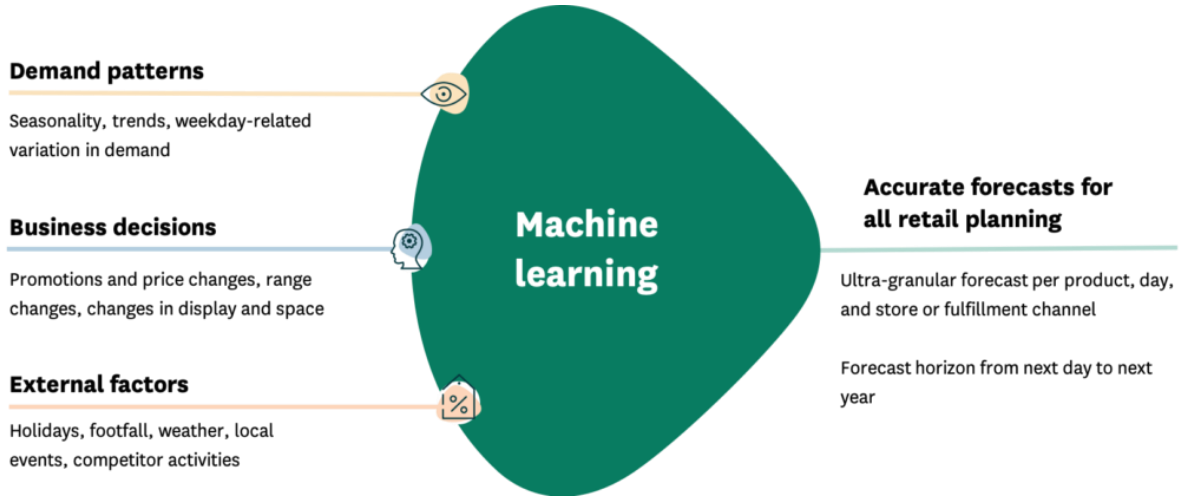
I dati:

- Input: immagini DICOM e/o annotazioni
- Grandissimo lavoro di pre-elaborazione
- Problemi di privacy e disponibilità dei dati: Transfer Learning e/o aumento dei dati
- Strategie di campionamento:
  - Divisione del set in addestramento, convalida e test
  - Convalida incrociata K-fold

Il modello:

- Immagine: sempre una CNN
- Testo: sempre una RNN (LSTM o GRU)

Demand forecasting



Workflow:

- Raccolta dei dati
- Comprensione e pre-elaborazione dei dati
- Costruzione del modello
- Formazione del modello
- Validazione del modello
- Miglioramenti
- Distribuzione

Demand forecasting

Raccolta e pre-elaborazione dei dati:

- Enorme quantità e diversa natura dei dati
- Metodi:
  - Trasformazione dei dati (log o trasformazione di potenza Box-Cox)
  - Destagionalizzazione dei dati
  - Detrending dei dati

I modelli maggiormente trovati:

- Perceptor multistrato (MLP)
- Rete neurale bayesiana (BNN)
- Reti neurali di regressione generalizzata (GRNN)
- Regressione K-Nearest Neighbor (KNN)
- Regressione del vettore di supporto (SVR)
- Rete neurale ricorrente (RNN)
- Rete neurale di memoria a lungo termine (LSTM)

Demand forecasting - Analisi

I dati:

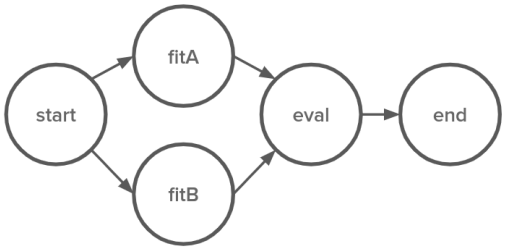
- Input: svariati tipi di dati
- Pretrattamento dei dati anche assente, con possibile riduzione del set di dati
- Estrazione delle caratteristiche con PCA (Principal Component Analysis)
- Stagionalità e clustering dei dati

Il modello è molto variabile:

- Strategia ibrida di metodi paralleli
- Oltre a quelli visti precedentemente: RBF, CART e GP
- Ensamble Learning: Stacking Algorithm

Framework - MetaFlow

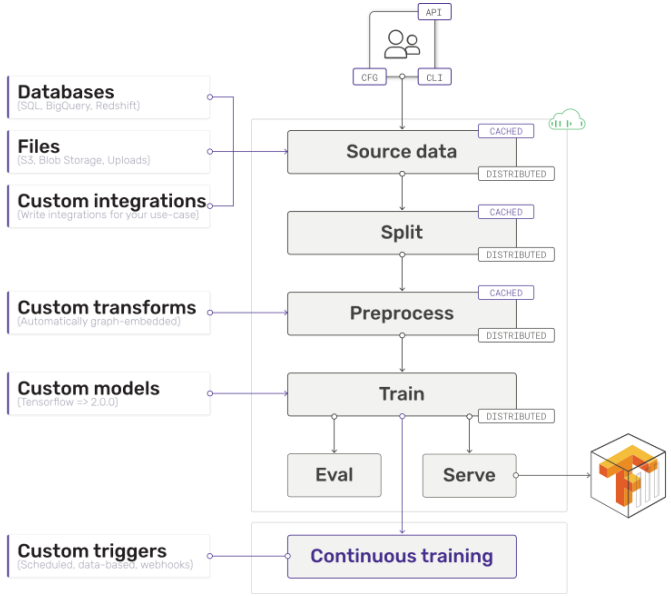
- Vantaggi chiave:
  - Prontezza alla produzione
  - Capace di addestrare due versioni di un modello in parallelo e scegliere quella con il punteggio più alto
  - Riusabilità del codice




```python
class MyFlow(FlowSpec):

  @step
  def start(self):
    self.data = load_data()
    self.next(self.fitA, self.fitB)

  @step
  def fitA(self):
    self.model = fit(self.data, model='A')
    self.next(self.eval)

  @step
  def fitB(self):
    self.model = fit(self.data, model='B')
    self.next(self.eval)

  @step
  def eval(self, inputs):
    self.best = max((i.model.score, i.model)
                    for i in inputs)[1]
    self.next(self.end)

  @step
  def end(self):
    print('done!')
```

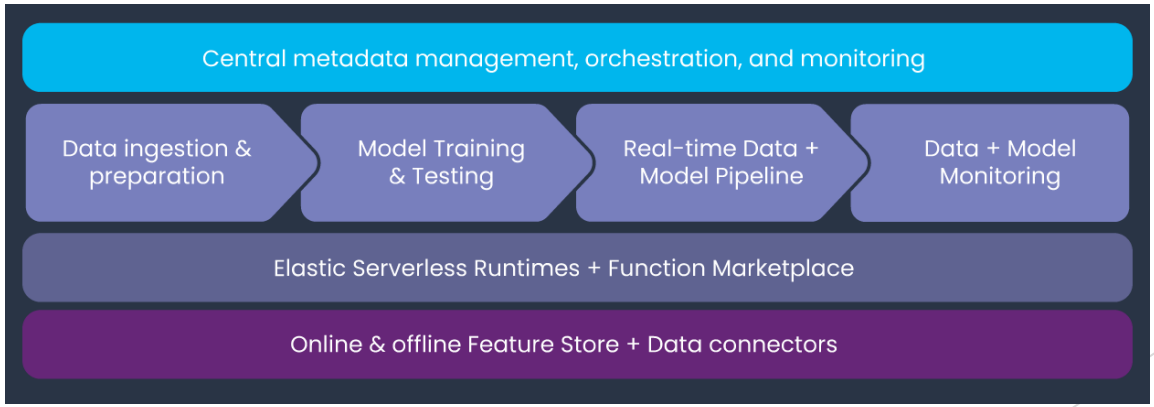
Framework - ZenML

- Vantaggi chiave:
  - Integrazione
  - Pre-elaborazione e riduzione di grandi set di dati
  - Riutilizzo del codice
  - Possibilità di passare rapidamente dall'ambiente locale a quello cloud (es: Kubernetes)



Framework - MLRun

- Vantaggi chiave:
  - Implementazione rapida del codice nelle pipeline di produzione
  - Supporta anche il parallelismo
  - Funziona ovunque: il tuo IDE locale, multi-cloud o on-premise
  - Ha un'interfaccia utente



Conclusioni

- Due focus differenti:
  1. Focus sui dati -> AI in radiology
     - Dati con la stessa struttura
     - Grande lavoro di pretrattamento
     - Poca variabilità nei modelli

  2. Focus sul modello -> Demand forecasting
     - Diversa natura dei dati
     - Pretrattamento anche assente
     - Molta variabilità nei modelli

- Punto di vista cronologico: focus che si sposta sempre di più sull'analisi dei modelli che si utilizzano.