CHECKLIST SETUP DEFINITIVO PER PROGETTI DI MACHINE LEARNING

by Intelligenzaartificialeitalia.net

STEP 1 -

Inquadra il problema e guarda il quadro generale

Per ogni domanda o richiesta usa la cella sottostante per rispondere

1.1 Definire l'obiettivo in termini di business.

1.2 Come verrà utilizzata la tua soluzione?

1.3 Quali sono le attuali soluzioni/soluzioni alternative (se presenti)?

1.4 Come dovresti inquadrare questo problema (con supervisione/non supervisione, ecc.)?

1.5 Come dovrebbero essere misurate le prestazioni?

1.6 La misura della performance è allineata con l'obiettivo aziendale?

1.7 Quale sarebbe la performance minima necessaria per raggiungere l'obiettivo aziendale?

1.8 Quali sono i problemi comparabili? Puoi riutilizzare l'esperienza o gli strumenti?

1.9 Hai a disposizione conoscenza e competenza per il progetto?

1.10 Come risolveresti il problema manualmente?

1.11 Elenca le ipotesi che tu (o altri) avete fatto finora.

1.12 Elenca qui altre note come LIMITI o difficoltà per arrivare al raggiungimento dell'obiettivo

---

STEP 2 -

Passiamo ai DATI

Usa le celle per scrivere il codice direttamente su questo file

2.1 Elenca i dati di cui hai bisogno e quanti te ne servono.

In [ ]:

2.2 Trova e documenta dove puoi ottenere quei dati.

2.3 Controlla quanto spazio ci vorrà. (alcuni progetti richiedono l'utilizzo di server)

In [ ]:

2.4 Verifica gli obblighi legali e ottieni l'autorizzazione se necessario. (sopratutto ramo Finanziario e Medico)

2.5 Crea uno spazio di lavoro (con spazio di archiviazione sufficiente).

2.6 Carica i Dati

In [ ]:

2.7 Converti i dati in un formato che puoi facilmente manipolare (senza modificare i dati stessi).

In [ ]:

2.8 Garantire che le informazioni sensibili siano cancellate o protette (ad es. rese anonime).

In [ ]:

2.9 Verificare la dimensione e la tipologia dei dati (serie storica, campionaria, geografica, ecc.).

In [ ]:

2.10 Prova un set di test, mettilo da parte e non guardarlo mai.

STEP 3 -

Esploraziona DATI

Usa le celle per scrivere il codice direttamente su questo file

3.1 Crea una copia dei dati per l'esplorazione (campionandola fino a una dimensione gestibile, se necessario).

In [ ]:

3.2 Studia ogni attributo e le sue caratteristiche

- Tipo (categoriale, int/float, bounded/unbounded, text, strutturato, ecc.) % di valori
- mancanti Rumorosità e tipo di rumore (stocastico, outlier, errori di arrotondamento, ecc.)
- Tipo di distribuzione (gaussiana, uniforme, logaritmica, ecc.)

In [ ]:

3.3 Per le attività di apprendimento supervisionato, identificare target e variabili descrittive.

In [ ]:

3.4 Visualizza i DATI

In [ ]:

3.5 Studia le correlazioni tra gli attributi.

In [ ]:

3.6 Identificare dati extra che sarebbero utili.

In [ ]:

3.7 Documenta ciò che hai imparato [IMPORTANTE !]

STEP 4 -

Preparazione dei DATI

- Lavora su copie dei dati (mantieni intatto il dataset originale).
- Scrivi funzioni per tutte le trasformazioni di dati che applichi.
  - In questo modo puoi facilmente preparare i dati la prossima volta che ottieni un nuovo set di dati
  - Quindi puoi applicare queste trasformazioni in progetti futuri

Usa le celle per scrivere il codice direttamente su questo file

4.1 Data cleaning.

- Correggi o rimuovi i valori anomali (opzionale)
- Inserisci i valori mancanti (ad es. con zeri, media, mediana...) o elimina le loro righe (o colonne)

In [ ]:

4.2 Selezione delle feature (opzionale).

- Elimina gli attributi che non forniscono informazioni utili per l'attività Se necessario,
- utilizza una tecnica di riduzione della dimensionalità (PCA, KernelPCA, LLE...)

In [ ]:

4.3 Feature engineering, dove appropriato.

- Discretizza le funzioni continue
- Scomponi le caratteristiche (ad es. categoriale, data/ora, ecc.)
- Aggiungi trasformazioni (ad es. log(x), sqrt(x), x^2, ecc.)
- Aggrega le feature in nuove promettenti campi o variabili derivate

In [ ]:

4.4 Feauture scaling.

- Standardizza o normalizza le feature spiegando i motivi

STEP 5 -

Scelta del Modello

- Se i dati sono enormi, potresti voler campionare set di addestramento più piccoli in modo da poter addestrare molti modelli diversi in un tempo ragionevole (tieni presente che ciò penalizza modelli complessi come grandi reti neurali o foreste casuali).
- Ancora una volta, prova ad automatizzare il più possibile questi passaggi.

Usa le celle per scrivere il codice direttamente su questo file

5.1 Addestra molti modelli rapidi e sporchi di diverse categorie (ad esempio, reg. lineare, Bayes, SVM, Random Forest, reti neurale, ecc.) utilizzando parametri standard.

In [ ]:

5.2 Misurare e confrontare le loro prestazioni.

- Per ogni modello, utilizzare la convalida incrociata N-fold e calcolare la media e la deviazione standard della misura delle prestazioni su N-fold.

In [ ]:

5.3 Analizza le variabili più significative per ciascun algoritmo.

In [ ]:

5.4 Analizza i tipi di errori commessi dai modelli.

- Quali dati avrebbe usato un essere umano per evitare questi errori?

In [ ]:

5.5 Esegui un rapido giro di feature selection e engineering.

In [ ]:

5.6 Ripeti i passaggi dal 5.1 al 5.5 per due o tre volte

5.7 Seleziona i migliori che reputi migliori dettagliandone i motivi

STEP 6 -

Ottimizzazione (Solo per esperti)

- Ti consigliamo di utilizzare quanti più dati possibile per questo passaggio, soprattutto mentre ti muovi verso la fine della messa a punto
- Come sempre automatizza quello che puoi.

Usa le celle per scrivere il codice direttamente su questo file

6.1 Perfeziona gli iperparametri usando la convalida incrociata.

- Tratta le tue scelte di trasformazione dei dati come iperparametri, specialmente quando non sei sicuro di esse (ad esempio, se non sei sicuro se sostituire i valori mancanti con zeri o con il valore mediano, o semplicemente eliminare le righe).
- A meno che non ci siano pochissimi valori di iperparametro da esplorare, preferire la ricerca casuale alla ricerca nella griglia. Se l'addestramento è molto lungo, potresti preferire un approccio di ottimizzazione bayesiano (ad esempio, utilizzando i processi gaussiani a priori).

In [ ]:

6.2 Prova i metodi Ensemble. La combinazione dei tuoi modelli migliori spesso produrrà prestazioni migliori rispetto all'esecuzione individuale.

In [ ]:

6.3 Una volta che sei sicuro del tuo modello finale, misura le sue prestazioni sul set di test per stimare l'errore di generalizzazione.

- Non modificare il tuo modello dopo aver misurato l'errore di generalizzazione: inizieresti semplicemente a sovraadattare il set di test. (a noi data scientist non piace l'overfitting, meglio evitarlo)

In [ ]:

6.4 Salva il modello

STEP 7 -

Presenta e Discuti la tua soluzione

Usa le celle per scrivere le risposte o il codice

7.1 Riporta quì sotto la documentazione che ti ha portato a fare le scelte fatte

7.2 Crea una bella presentazione

- Aiutati con immagini e file multimediali

7.3 Spiega perché la tua soluzione raggiunge l'obiettivo aziendale

7.4 Non dimenticare di presentare punti interessanti che hai notato lungo il percorso.

- Descrivi cosa ha funzionato e cosa no.
- Elenca le tue ipotesi e i limiti del tuo sistema

7.5 Assicurati che i tuoi risultati chiave siano comunicati attraverso belle visualizzazioni o annotazioni facili da ricordare.

STEP 8 -

Si inizia la distribuzione

Usa le celle per scrivere il codice direttamente su questo file

8.1 Prepara la tua soluzione per la produzione (collegamento a input di dati di produzione, scrittura di unit test, ecc.).

8.2 Scrivi un codice di monitoraggio per controllare le prestazioni in tempo reale del tuo sistema a intervalli regolari e attivare avvisi quando andrà in crash.

- Attenzione al lento degrado: i modelli tendono a "marcire" man mano che i dati si evolvono.
- Monitora anche la qualità dei tuoi input (ad esempio, un sensore malfunzionante che invia valori casuali o l'output che diventa obsoleto).

Ciò è particolarmente importante per i sistemi di apprendimento online.

8.3 Riaddestra regolarmente i tuoi modelli su dati aggiornati (automatizza il più possibile).

Grazie a questa checklist oltre 10.000 persone sono riuscite a completare con successo il loro progetto

by Intelligenzaartificialeitalia.net