Problema 1

Si ringraziano Giles J. Hooker e Saharon Rosset per aver condiviso questi dati.

La descrizione del problema

Il **Netflix Prize** è stata una competizione il cui scopo era quello di prevedere le valutazioni (*rating*) di diversi film fornite dagli utenti. Netflix ha fornito le valutazioni di 17.770 titoli di film da parte di 480.189 utenti, insieme alla data di ciascuna valutazione. Il compito era quello di prevedere le valutazioni per 282.000 combinazioni di utente-film-data che non erano presenti nel training set.

Netflix ha misurato la bontà delle previsioni con la radice quadrata dell'errore quadratico medio (*Root Mean Square Error*) e ha offerto un premio di $ 1.000.000 al primo classificato (che ha migliorato la bontà delle previsioni del loro approccio di oltre il 10%). Il premio è stato vinto nel 2009. I dettagli del Premio Netflix sono disponibili presso www.netflixprize.com

La competizione

Poiché la competizione Netflix prevedeva un dataset molto grande e un problema non-standard, la nostra competizione semplificherà notevolmente il problema.

Il training set fornisce le valutazioni di $n = 10000$ utenti per 99 film, insieme alle date in cui sono state effettuate le valutazioni. Si noti che 14 di questi film sono stati valutati da tutti gli n utenti, mentre i restanti 85 film contengono valori mancanti.

L'obiettivo è prevedere la valutazione da parte dei $m = 2931$ utenti del test set per il film **Miss Detective** (titolo originale: *Miss Congeniality*); vi viene anche fornita la data in cui è stata effettuata ciascuna valutazione. Come per il training set, tutti gli m utenti del test set hanno valutato 14 film, mentre per i restanti 85 ci sono valori mancanti. Il test set fornisce le stesse informazioni del training set: le date e le valutazioni di questi 99 film insieme alla data delle valutazioni per *Miss Congeniality*.

Come per la competizione Netflix, la bontà delle previsioni verrà valutata con la radice quadrata dell'errore quadratico medio (RMSE) sul test set.

I dati

I dati per la competizione sono disponibili nell'archivio del corso in formato file di test delimitato da tabulazioni, e comprendono:

- Train_ratings_all.dat: il training set contiene le valutazioni che gli utenti hanno assegnato a ciascuno dei 99 film
- Test_ratings_all.dat: come sopra per il test set
- Train_dates_all.dat: per il training set, le date in cui sono state effettuate le valutazioni di cui sopra.
- Test_dates_all.dat: come sopra per il test set
- Train_y_rating.dat: le valutazioni che gli utenti del training set hanno assegnato a *Miss Congeniality*
- Train_y_date.dat: per il training set, le date in cui gli utenti del hanno valutato *Miss Congeniality*
- Test_y_date.dat: Stesse informazioni per il set di test
- Movie_titles.txt: i titoli e le date di uscita per i 99 film, indicati nello stesso ordine delle colonne dei dati descritti sopra

Alcune osservazioni:

- Le valutazioni sono numeri interi da 1 a 5. Un valore di 0 indica un valore mancante
- Per comodità, le date sono fornite come numero di giorni trascorsi a partire dal primo Gennaio 2017 fino ad una certa data. L'etichetta che identifica i valori mancanti per le date è '0000'

Utilizzo di **dati esterni**:

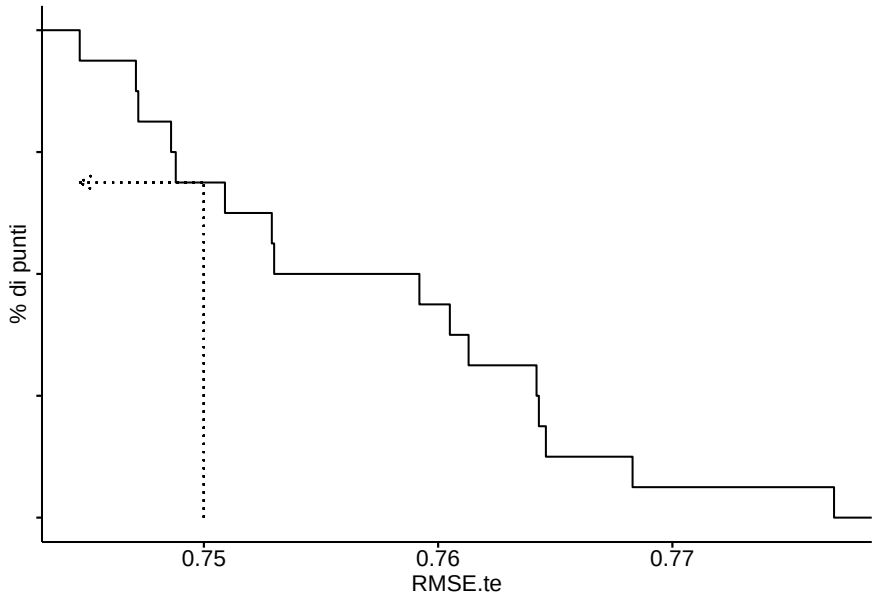
- E' ammissibile utilizzare dati provenienti da fonti esterne a condizione di **pubblicare** sulla pagina MOODLE il dataset con relativa descrizione a beneficio degli altri partecipanti alla competizione.

La valutazione delle previsioni

Le previsioni finali $\hat{y}_1^*, \ldots, \hat{y}_m^*$ verranno valutate in termini di RMSE sul test set:

$$\text{RMSE}_{\text{Te}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i^* - \hat{y}_i^*)^2} \tag{1}$$

I punti attribuiti verranno assegnati calcolando il **quantile** del vostro punteggio rispetto alla distribuzione dei punteggi della classifica finale, come illustrato dal seguente grafico:



Ad esempio, un punteggio finale di $\text{RMSE}_{\text{Te}} = 0.75$ corrisponde a 68.75% dei punti. Si noti che la distribuzione dei punteggi finali verrà sempre aggiornata con le ultime sottomissioni (quella visualizzata corrisponde ai punteggi dell'anno scorso, che saranno comunque inclusi nel calcolo).

In ogni caso, se risultasse necessario, mi riservo il diritto di cambiare la regola di valutazione (il che significa che la vostra % di punti potrebbe essere calcolata in modo diverso).

Solo per i partecipanti in gruppi, durante la competizione sarà possibile sottomettere le proprie previsioni sulla piattaforma **BeeViva** (utilizzando il pulsante "Submission" di http://www.bee-viva.com/competitions/miss_c). La classifica (*leaderboard*) parziale di questa piattaforma vi mostrerà il vostro **punteggio parziale**, ovvero il RMSE su un sottoinsieme casuale $S \subset \{1, \ldots, m\}$ di 1000 osservazioni del test set (che resterà sempre lo stesso):

$$\text{RMSE}_{\text{Te.parziale}} = \sqrt{\frac{1}{1000}\sum_{i \in S}(y_i^* - \hat{y}_i^*)^2} \tag{2}$$

Alla fine della competizione, la classifica mostrerà il **punteggio finale**, che è l'RMSE (1) per tutte le $m = 2931$ osservazioni del test set. Il punteggio finale determinerà il vincitore finale (**l'ultima sottomissione prima della scadenza sarà considerata quella definitiva**). Questo approccio è utilizzato per prevenire ai partecipanti di ottimizzare il proprio punteggio fornito da (2). Tutti i partecipanti in gruppi devono iscriversi alla piattaforma e sottomettere la **stessa** previsione finale (ma non necessariamente le stesse previsioni intermedie). Nel corso della competizione, ci potrebbero essere malfunzionamenti della piattaforma BeeViva: siete pregati di segnalarli nel Forum dedicato (non via e-mail).

Per tutti gli altri studenti che lavorano individualmente, le previsioni dovranno essere consegnate almeno una settimana prima della data di esame, caricandole sulla pagina MOODLE secondo il formato richiesto. Sarà possibile consegnare le previsioni **una volta sola** per A.A.

La relazione e la presentazione

Oltre alla previsione finale, bisognerà produrre una **relazione** da consegnare entro la scadenza prevista (per i gruppi, quella indicata, per i lavori individuali, almeno una settimana prima della della data di esame). La relazione deve essere un **UNICO** file in formato **.PDF** (non sono ammessi altri tipi di file) contenente

1. la descrizione dell'analisi
2. il codice utilizzato per ottenere la previsione finale. E' ammesso l'utilizzo di qualsiasi linguaggio di programmazione, tuttavia il codice utilizzato per produrre le previsioni finali deve risultare **RIPRODUCIBILE**. Riportate solo il codice **strettamente necessario** (evitando di produrre grafici intermedi, modelli non utilizzati, etc.)

Il file deve essere nominato nel seguente modo: [MATRICOLA]_HW1.pdf (e.g. 2575_HW1.pdf) per i lavori individuali e [NOME DEL GRUPPO]_HW1.pdf per i lavori di gruppo. Il file dovrà essere caricato sulla pagina MOODLE in corrispondenza all'HOMEWORK1.

La descrizione dell'analisi può includere

- Le idee chiave dell'analisi
- La pre-elaborazione (*pre-processing*) dei dati (trasformazioni di variabili, etc.)
- Come sono stati gestiti i valori mancanti (*missing values*)
- La creazione di nuove variabili (*feature engineering*)
- La selezione delle variabili (*feature selection*)
- L'utilizzo di eventuali dati esterni
- I modelli considerati e quello utilizzato per la previsione finale, e se effettuata, la regolazione del modello (*model tuning*) e la stima dell'errore di previsione (e.g. con convalida incrociata)
- Le fonti utilizzate (libri, Internet, etc.). L'uso di fonti senza citarle si traduce in un voto nullo.
- Etc.

Non è previsto un limite di pagine per il file da consegnare, ma verrà valutata la capacità di sintesi (del testo e del codice).

Infine, **gli studenti che partecipano in gruppo** dovranno preparare una **breve presentazione** del loro lavoro (circa 10 minuti per gruppo). La presentazione si svolgerà nell'ultima lezione del corso (o in una o più date specificate dal docente).

La valutazione complessiva

Il punteggio che ciascun studente ha ottenuto dalle previsioni finali potrà subire aggiustamenti sulla base

- della relazione
- della presentazione (se prevista)
- del form di valutazione sul lavoro di gruppo (se previsto)

o su altre considerazioni da parte del docente.

Il regolamento

1. Il docente si riserva la possibilità di chiedere a qualunque studente di spiegare la relazione e/o il codice utilizzato. Per i lavori individuali, questa spiegazione (se richiesta) avverrà il giorno della prova scritta
2. Il punteggio ottenuto scade alla fine dell'Anno Accademico 2020/21
3. Tutti gli studenti sono tenuti ad aderire ad un codice di condotta, che vieta il plagio, la falsificazione, l'assistenza non autorizzata, imbrogli e altri atti gravi di disonestà accademica. Comportamenti non corretti possono essere soggetti a provvedimenti disciplinari come da Art. 35 e 36 del Regolamento Didattico di Ateneo.