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 di diversi film fornite degli utenti. Netflix ha fornito le valutazioni di 17.770 titoli di film da 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 nel training set; tutti gli utenti e i film presenti nel test set erano già presenti nel training set.

Netflix ha valutato le prestazioni in base alla radice quadrata dell'errore quadratico medio (*Root Mean Square Error*) per il test set e ha offerto un premio di $ 1.000.000 al primo classificato che ha migliorato le prestazioni del loro sistema attuale 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. I primi 14 di questi film sono stati valutati da tutti gli utenti; i restanti 85 contengono valori mancanti. L'obiettivo è la valutazione che ogni utente ha dato al film **Miss Detective** (*Miss Congeniality*); vi viene anche fornita la data in cui è stata effettuata ciascuna valutazione.

Il compito è prevedere la valutazione di questo film da parte di altri $m = 2931$ utenti nel test set. Come per il training set, tutti gli utenti del test set hanno valutato i primi 14 film, mentre 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 della valutazione per *Miss Congeniality*. Come per la competizione Netflix, le prestazioni saranno misurate in base alla 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'

Il regolamento

1. Bisogna consegnare entro la scadenza prevista (per i gruppi, quella indicata, per i lavori individuali, almeno una settimana prima della della data di esame) un **UNICO** file in formato **.PDF** (non sono ammessi altri tipi di file) contenente le risposte alle domande (e il codice utilizzato). Il file deve essere nominato nel seguente modo: [MATRICOLA]_HW0.pdf (e.g. 2575_HW0.pdf) per i lavori individuali e [NOME DEL GRUPPO]_HW0.pdf per i lavori di gruppo. Il file dovrà essere caricato sulla pagina MOODLE in corrispondenza all'HOMEWORK0. Per i lavori di gruppo, **tutti** i componenti del gruppo devono caricare lo stesso file. Sarà possibile effettuare **una sola** sottomissione finale (vi verrà chiesta conferma, non sono ammesse consegne via e-mail). Per tutti gli studenti, il mancato rispetto della scadenza prevista corrisponde ad un punteggio di 0.
2. Per rispondere alle domande, è ammesso l'utilizzo di qualsiasi linguaggio di programmazione. Tuttavia il codice utilizzato deve essere riportato e deve risultare **RIPRODUCIBILE**. Troverete nella pagina MOODLE un esempio in Rmarkdown con la risposta alla Domanda0.
3. La valutazione si baserà sulla correttezza, chiarezza e precisione delle risposte fornite e del codice utilizzato. Non è previsto un limite di pagine per il file da consegnare, ma verrà premiata la capacità di sintesi, ovvero una struttura argomentativa ben articolata, codice elegante e leggibile, con le conclusioni che rispondono in modo specifico e puntuale alla domanda iniziale. E' possibile utilizzare fonti (libri, Internet, persone e così via) ma è richiesto di citarle nel testo. L'uso di fonti senza citarle si traduce in un voto nullo.
4. Il docente si riserva la possibilità di chiedere a qualunque studente di spiegare le risposte fornite e/o il codice utilizzato. Per i lavori individuali, questa spiegazione (se richiesta) avverrà il giorno della prova scritta. Il punteggio ottenuto scade alla fine dell'Anno Accademico 2020/21. 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.