Esame corso Informatica Biotecnologia

Docente: Matteo Re

1 Requisiti software:

Lo svolgimento di questo progettod'esame richiede l'utilizzodi R. Le funzioni necessarie per la sua realizzazione sono contenute in diverselibrerie pubbliche che dovrete installare prima di iniziare a lavorare al progetto. Consiglio a tutti di utilizzare unaversionedi R recente. In condizioni ideali dovreste utilizzare la stessa versione di R che avete installato nei lab oratori di informatica e che avete utilizzato durante il corso. Es sa e'disp onibile per il download all'indirizzo:
http://rm.mirror.garr.it/mirrors/CRAN/bin/windows/base/R- 2.15.1- win.exe

Oltre ad R e' necessario installare alcune librerie per la manip olazione ed analisi di strutture molecolari. Alcune di esse sono disp onibili p er il download dal rep ository pubblico CRAN (The Comprehensive R Archive Network). Potete installarle utilizzanzo il comando *instal l.packages()* direttamente dall'interno dell'ambiente R. Le librerie CRANda installare sono:

- **bionetdata** (contiene diverse re ti biologiche e farmacologiche)
- **NetPreProc** (contiene diverse funzioni per la normalizzazione dei dati strutturati in forma di rete)
- **PerfMeas** (contiene diverse funzioni p er la quantificazione delle p erformance ottenute in esperimenti di classificazione e ranking)

Prima di poter installare queste librerie (in particolare NetPrePro c e PerfMeas) dovrete installare altri package contenuti nel repositoy pubblico *Bioconductor*, http://www.bioconductor.org/ . I package Bioconductor da installare sono i seguenti:

- **graph**(http://www.bioconductor.org/packages/release/bioc/html/graph.html)
- **limma**(http://www.bioconductor.org/packages/release/bioc/html/limma.html)
- **RBGL**(http://www.bioconductor.org/packages/release/bioc/html/RBGL.html)

Potete trovare molte informazioni riguardanti questi package Bio conductor ai link che vi ho indicato.
ATTENZIONE: per la realizzazione de l progetto d'esame non è previsto l'utilizzo diretto da parte vostra delle funzioni contenute nei package Bio conductor. Il loro ruolo e' quello di fornire funzioni necessarie per il corretto funzionamento dell'implementazione delle funzioni dei kernel Random Walk (RW) e Random Walk con restart (RWR) che utilizzerete per realizzare gli esperimenti di ranking.

Per rispondere ad alcune domande della prima parte del progetto dovrete anche installare alcune librerie specifiche per la manipolazione di strutture molecolari in R. Queste librerie sono le stesse che avete installato in classe e che avete utilizzato durante le lezioni:

- **rcdk**
- **ChemmineR**

Unitamente al tema di esame riceverete anche un file sorgente R (*RWR.R*) contenente diverse funzioni che dovrete utilizzare per la realizzazione degli esperimenti.

Ad ognuno dei gruppi che si iscriveranno per sostenere l'esame nella sessione verrà comunicata mediante email una seconda categoria terapeutica da utilizzare, insieme alle penicilline, per la realizzazione del progetto d'esame.

2 Realizzazione progetto:

Questa sezione del tema d'esame e' divisa in due parti. La prima e' dedicata all'esplorazione dei dati contenuti nellalibreria *bionetdata* coinvolti nella realizzazione del progetto: DD.chem.data(matrice disimilarita'pairwise tra farmaci appartenenti ad un set di molecole approvate dalla Food and Drugs Administration (FDA) americana), e DrugBank.Cat (matrice binaria delle categorie terap eutiche dei farmaci considerati ottenuta dalla banca dati pubblica DrugBank 3.0, http://www.drugbank.ca/).
La seconda parte è dedicata alla realizzazione di un semplice esperimento di drug repositioning da realizzarsi mediante l'analisi dei dati descritti in precedenza. Come punto di partenza provate a visualizzare tutte le categorie terapeutiche presenti nella matrice delle etichette (DrugBank.Cat). Quante sono? Verificate, inoltre, il numero dei farmaci contenuti nel dataset. Riportate entrambe le informazioni nella relazione.

2.1 Esplorazione dataset (categorie terapeutiche e similarità strutturale)

Domanda 1:
La matrice delle etichette (DrugB ank.Cat)`e una matrice in cui le righe corrisp ondono alle molecole e le colonne alle categorie. Un valore 1 all'incrocio tra la decima riga e la quinta colonna indica che il decimo farmaco appartiene alla categoria associata alla colonna 5. In caso contrario avremmo trovato un valore 0. Calcolare il numero di farmaci presenti in ogni categoria utilizzando un'unica sitruzione R.

Domanda 2:
Quante sono le penicilline presenti nel dataset? Rispondere utilizzando un'unica istruzione R.

Domanda 3:
Estrarre gli identificativi DrugBank della categoria "Penicillins".

Domanda 4:
Estrarre utilizzando lafunzione *which* gli indici(numeri di riga) dei farmaci appartenenti alla categoria"Penicillins". Che tipo di variabile utilizzereste per salvare questa informazione? Motivare la risposta.

Domanda 5:
Scegliere una delle molecole della categoria terapeutica "Penicillins". Scaricare i files SDF e SMILES dei farmaci approvati presenti nella banca dati DrugBank disponibili nella homepage del corso. Utilizzando le informazioni presenti in questi file calcolare con R formula e peso molecolare del farmaco scelto. Disegnare inoltre la molecola utilizzando R (ed includere l'immagine nella relazione).

Domanda 6:
La matrice DD.chem.data contiene i valori dei coefficienti di Tanimoto ottenuti dal confronto a coppie tra tutti i farmaci presenti nel dataset (in questa matrice sia le righe che le colonne rappresentano dei farmaci). Calcolare la similarità media (coeff. Tanimoto) tra tutti i farmaci del dataset.

Domanda 7:
Creare unamatriceTanimo toche contenga solo i valori di similarità pairwise tra i farmaci della categoria "Penicillins".

Domanda 8:
Calcolare la similarità media (coeff. Tanimoto) tra i farmaci presenti nella matrice ottenuta rispondendo alla domanda 7 (in questo esempio le penicilline). Confrontando questa similarità media nella categoria considerata con la similarità media in tutto il dataset (che avete calcolato rispondendo alla domanda 6) cosa potete dire sulla categoria terapeutica considerata?

Rispondere nuovamente alle domande 2, 4, 7 e 8 utilizzando i farmaci della seconda categoria terapeutica assegnata al gruppo.

2.2 Esperimenti di drug repositioning mediante ranking di similarità tra farmaci appartenenti a categorie terapeutiche note

L'obiettivo di questa parte del progetto d'esame è di dimostrare che è possibile predire l'appartenenza di un dato farmaco ad una categoria terapeutica mediante l'utilizzo di metodi di apprendimento automatico.
ATTENZIONE : Le performance per gli esp erimenti di ranking che dovrete realizzare dovranno essere calcolate in termini di Area sotto la curva ROC (*AUC*).

Test di drug rep ositioning basato su ranking di farmaci:
In questo test dovrete valutare l'impatto della cross validazione sulla stima delle performance di ranking sottoforma di area sotto la curvaROC (AUC). Per farlo utilizzate un Random Walk con Restart (RWR). Dovrete procedere in due passi in cui effettuerete il test rispettivamente senza e con la cross validazione (5 o 10 folds).
Per eseguire il test *senza* la cross validazione la funzione da utilizzare (che avrete a disposizione dopo aver caricato il file *RWR.R*)è RWR(m,indpos), i cui argomenti sono una matrice (m) di adiacenza ed un vettore (indpos) contenente gli indici degli elementi positivi (ossia i farmaci appartenenti alla categoria terapeutica considerata).
Una volta eseguito il test dovrete calcolare l'area sotto la curva ROC utilizzando la funzione AUC.single(v,etichette) i cui argomenti sono un vettore delle probabilità di appartenenza alla categoria terapeutica considerata calcolati da RWR ed un vettore binario contenente le vere etichette (1 o 0) p er la categoria terapeutica considerata.
Il test dovrà essere effettuato sia sulla categoria "Penicillins" che sulla seconda categoria terapeutica (quella che vi è stata comunicata contestualmente alla consegna del tema p er il progetto dell'esame). Riportate nella relazionei risultati di performance (in termini di AUC) per i ranking di entrambe le categorie terapeutiche.
Nella seconda fase di questo test rip etete l'esperimento di ranking su entrambi i set di farmaci utilizzando la funzione **RW.cv**(m,indp os,k=5, fun=RWR) in cui m ed indpos sono le stesse variabili che avete utilizzato durante l'utilizzo della funzione RWR(), k=numero è il numero dei fold per la cross validazione e fun è un parametro che indica il tipo di funzione da utilizzare (usate fun=RWR per scegliere il Random Walk con Restart). **ATTENZIONE:** Utilizzate 5 o 10 fold per i test che prevedono l'utilizzo della cross validazione.
Dopo aver eseguito i test quantificate le performance utilizzando la funzione AUC.s ingle(), come avete fatto in precedenza. Riportate i valori di AUC ottenuti nella relazione.

Domanda 9:
I risultati di performance che avete ottenuto con e senza cross validazione sono diversi? In caso abbiate risposto si, i risultati sono diversi per entrambe le categorie terapeutiche?

Domanda 10:
Quale dei risultati è più attendibile? Quello senza cross validazione o quello ottenuto mediante cross validazione? Motivate la risposta.

Domanda 11:
Scegliete (in accordo alla risposta della domanda precedente) i risultati più affidabili. Nel caso in cui le performance di ranking ottenute per le due classi farmacologiche siano differenti cercate di motivare la differenza in AUC osservata tenendo conto:

- della natura dei dati utilizzati (similarità molecolare, coeff. Tanimoto)
- della natura delle diverse categorie terapeutiche dei farmaci che avete analizzato

Domanda OPZIONALE:
Scrivere un *programma* che utilizzi un ciclo a vostra scelta per scrivere sullo schermo:

- nome della categoria terapeutica (nomi colonne matrice etichette)
- numero di farmaci
- performance di ranking (RWR, 10 o 10 folds CV) sottoforma di AUC

per ogni colonna della matrice delle etichette (DrugBank.Cat).

3 Note aggiuntive riguardanti la preparazione e l'invio della relazione

La relazione riguardante il progetto d'esame dovrà essere inviata al mio indirizzo email(quello che avete utilizzato durante il corso per spedirmi gli esercizi di programmazione in R).
La relazione dovrà contenere *nome, cognome e numero di matricola*, in caso contrario non potrà essere valutata ai fini del superamento dell'esame. La relazione dovrà essere scritta in due parti: la prima contenente le risposte alle domande e la seconda contenente *tutto* il codice R che utilizzerete per rispondere alle domande (indicando in quale specifica domandaè stato utilizzato).

4 In caso di necessità

Nel caso in cui durante la realizzazione del progetto d'esame dovessero sorgere dubbi inerenti all'interpretazione delle domande, potrete inviarmi delle email. In questo caso cercate di essere il più chiari possibile in modo da permettermi di interpretare correttamente i vostri dubbi. Risponderò in modo da evitare incidenti dovuti ad errata interpretazione delle domande ma, **in nessun caso**, invierò la soluzione del problema sotto forma di codice R. Tenete anche presente che la maggioranza delle domande che trovate in questo tema d'esame può essere risolta cercando la soluzione nelle slide del corso che sono (e rimarranno) disponibili sul sito web del corso.