Appunti Molecular Design

Saul Pierotti

May 17, 2019

Lex.1

Lex.2

Lex.3

Lex.4

Lex.5

Lex.6

Lex.7

Lex.8

Lex.9

Lex.10

Lex.11

Lex.12

Lex.13

- Intelligenza artificiale
- Alan Turing ha dato le basi per lo sviluppo dell'AI
- ENIAC, uno dei primi computer
- Kasparov e Deep Blue
- Perchè oggi vi è l'esplosione di AI?
  - Disponibili grandi quantità di dati
- Gestire l'AI

Lex.14

- Primo software AI sviluppato a Stanford per risolvere gli spettri MS

- All'aumentare della massa il numero di molecole compatibili con un certo spettro MS aumenta esponenzialmente
- Fecero quind spettri MS/MS, che riducevano le ambiguità
- Svilupparono un software capace di riconoscere gruppi chimici dagli spettri, in questo modo affinando i risultati a poche o 1 molecola
  - E' considerato il primo knowledge-based system
  - E' basato sulla conoscenza di esperti nel campo, che hanno scritto il programma
- A Perugia si è riapplicato lo stesso concetto per identificare lipidi in studi di lipidomica
- Per determinare il passaggio tra 2 stati, non è necessario descrivere gli stati stessi
  - E' quello che viene fatto calcolando la strada per un posto
- I sistemi knowledge based sono applicabili solo a sistemi noti, non generano soluzioni nuove
  - Sono molto usati nella ricerca universitaria, meno in quella di nicchia ed applicata di frontiera
- Machine learning è il sistema più usato in AI
  - Estrae un pattern da dati raw
  - La sua efficacia dipende molto dal sistema di coordinate usato
  - Posso applicarlo per ottimizzare la resa di una reazione
    - Devo descrivere i reagenti e le condizioni

Vacanze di pasqua

Lex.15

- La parte difficile di AI è estrarre knowledge dai dati
- Pattern recognition
- I computer di solito lavorano con matrici
- Lavoriamo solo con matrici 2d perchè è possibile ricondurre un a matrice n-dim a 2d
- Analisi di immagine
  - Un'immagine può essere ricondotta ad una matrice
  - L'esperimeto è il pixel, le variabili sono la luminanza dei vari canal Analisi di immagine
    - Un'immagine può essere ricondotta ad una matrice
    - L'esperimeto è il pixel, le variabili sono la luminanza dei vari canali
- MALDI/TOF su tessuti??
- Per ricondurre una matrice a 2d si fa unfolding
  - accodo tutti i dati su una sola dimensione, ottengo un vettore
- AI lavora meglio con dati hard (quantitativi)
  - Quando possibile è meglio convertere dati soft in hard
- I dati continui sono più trattabili matematicamente di quelli discreti
  - Sono derivatizzabili
- Least squares
- Per semplificare conviene linearizzare le relazioni tra dati
- Un modello non linear è la superificie di risposta
- Il modello più semplice è più probabile, overfitting
- Le reti neurali fanno un po' overfitting
- Linear discriminant analisys
  - Prendo una retta casuale e vi proietto tutti i punti, e valuto quanto efficacemente li clusterizza
  - Prendo un'altra retta e faccio la stessa cosa
  - La retta di separazione è perpendicolare a quella di proiezione
  - Se lavoro con uno spazio 3d proietto su dei piani

Lex.16

- La scelta dei descrittori influenza la qualità della discriminazione
- LDA funziona bene con poche classi, fino a 5 circa

- LA PCA è un metodo unbiased che non richiede la conoscenza pregressa della presenza di classi
  - E' un metodo di riduzione della dimensionalità
  - La prima componente è la direzione spaziala che discrimina il maggior numero possibile di punti
  - E' la combinazione lineare delle varie dimensioni
  - Score plot con oggetti ripetto alle componenti principali
  - Loading plot riporta il cos delle componenti principali rispetto alle variabili originali

Lex.17

- slides su chemiome.chm.unipg.it/MolDes19/
- In PCA posso fare autoscaling per rapportare equamente le dimensioni delle variabili
  - Riporta tutte le variabili nel range 0-1 moltiplicando per una costante
- La sterling a corciano produce steroidi
- Projection to latent strcutures (PLS)
  - E' un metodo supervised
  - Cerca di relazionare una matrice di variabili con una matrice di risposte
  - Una volta si faceva multiple regression analysis
    - Funzionava solo con poche x
    - Non permette buchi nella matrice
    - Non è stabile con x correlate tra loro
  - Trovo PC1 nel mondo x e nel mondo y
  - Creo un plot con le 2 PC1 x e y
  - Posso fare la stessa cosa con PC2 ecc, ma di solito perde correlazione
  - In realtà modifica le PC per massimizzarne la relazione
- Per evitare overfitting faccio cross validation, all'aumentare delle variabili l'errore prima diminuisce e poi aumenta perché sto modellando il rumore

Lex.18

Lex.19

- In molti casi le proprietà chimiche di una molecola non sono sufficienti per prevedere le interazione con un suo recettore
  - Volsurf è utile per predire le interazioni con i solventi, non con il recettore
- Le piccole molecole di solito interagiscono in tasche del recettore, le proteine su superfici dello stesso
- FLAP è un software che gestisce queste situazioni
  - Usa comunque i MIF
  - Uso probes dry, OH e ionici per definire le proprietà delle tasche
  - In queste interazioni è importante anche la geometria dell'interazione
  - Testo una banca dati di gruppi chimici sulle tasche
  - Mettendo poi insieme i gruppi creo una molecola che interagisce col recettore
  - Prendo una molecola
    - La faccio interagire con probes chimici e trovo le interazioni
    - Definisco dei punti a massima interazione e più spaziati possibili per i vari probes per semplificare la descrizione
      - Lo fa l'algoritmo
      - Approssimo la forma delle varie zone d'interazione
    - Definisco la sua superfici
    - Unisco le 2 cose creando uno shape con i punti sovrapposti
    - Semplifico anche la descrizione della superficie riducendo il numero di punti
    - Una quadruplet è un'entità geometrica di 4 punti uniti con dei segmenti (6 per unirli tutti)
      - E' definita dalle 6 distanze, dalle feature dei 4 punti (dry, ecc.)
      - Uso quattro punti perchè ho 4 probes (3+1 shape)

        - Un ultimo descrittore mi indica la chiralità
        - Non è chiralità chimica, ma dei punti
        - Può essere positivo o negativo
        - Tutte le quadruplette sono chirali anche se hanno tutti i punti uguali
        - Creo una matrice piana con tutte le quadruplette possibili
        - Con queste descrivo una matrice cubica che ha una piana per ogni conformazione possibile
- Prendo una tasca
    * Faccio la stessa cosa che avevo fatto con la molecola
- Faccio fitting
    * Confronto la tasca e la molecola quadrupletta per quadrupletta
    * Sovrappongo donatori con accettori, non gruppi uguali (!)
    * Il dry invece sta col dry
    * C'è un po' di tolleranza
    * Poso la molecola nella cavità sovrapponendo le quadruplette
    * In una piccola molecola 2-3*10^5 quadruplette, in una proteina 3-5*10^6