Esame Strutture Dati per la Bioinformatica - 07/03/2018
Appello III
Sessione Invernale
Docente: Marco Pietrosanto

I seguenti esercizi possono essere risolti con l'ausilio del computer ma molti di questi anche con solo un attento ragionamento. L'importante è che tutti i risultati siano riportati sul foglio, qualunque codice rimasto sul computer non potrà essere valutato.

Sarà possibile ottenere 14 punti per sezione (max 42 punti). Qualunque voto uguale o superiore a 32 equivale ad una lode. Un esercizio può fornire anche solo una parte dei suoi punti, a seconda della qualità dello svolgimento.

Ricordo che è possibile usare qualunque fonte di informazioni online (documentazione, stackOverflow etc) ma è proibito comunicare tra di voi o con esterni (nessuna forma di chat è consentita).

A - Python core (14 punti)

A1) Qual è la struttura di una list comprehension? (2 punti)
A2) Scrivi una list comprehension che data una lista composta da stringhe restituisca una lista con il primo elemento di ogni stringa: (2 punti)
`(es. ['Walter', ' space before me', 'Printer', 'youtube']`
`-> ['W', ' ', 'P', 'y']`
A3) *for* _____ *in enumerate(iterable):*
for _____ *in enumerate(zip (iterable1, iterable2)):*
Completa i *for* statements (con nomi delle variabili a piacimento) (3 punti)
A4) Trova, **spiega** l'errore e **descrivi** come lo risolveresti: (3 punti)

```
info = ["meth1 50", "meth2 90"]
mydict = {}
for row in info:
  spl = info.split()
  method=spl[0]
  score=spl[1]

  mydict[method].append(score)
```

A5) Un collega ti dice che sta lavorando con un file di ~50GB. Ti dice che quando prova ad aprirlo in Python si blocca tutto e che sta pensando di abbandonare Python per R, qual è la seconda domanda che gli fate? (2 punti)
A6) Devi scrivere una funzione ma non sai a priori quanti argomenti gli verranno passati, in che modo scrivi la definizione? (2 punti)

B - NumPy, sklearn (14 punti)
B1) Trova e spiega l'errore: (2 punti)

```
import numpy
np.mean( [0, 1, np.nan] )
```

B2) Dato l'array:

```
arr = np.array([12, 10, 0, 20])
```

scrivi un comando che seleziona, con una maschera booleana, gli elementi < 11 (2 punti)

B3) Cosa succede se sommo due **array numpy** fatti così: [1,2,3] + [3,2] ? Come si chiama questa "proprietà" (3 punti)
B4) Quante "regression metrics" puoi trovare nel modulo sklearn.metrics? (2 punti)
B5) Spiega a **cosa serve** e **quando serve** impostare un *seed* per un generatore di numeri random nel modo più sintetico possibile (3 punti)
B6) Trova e spiega l'errore: (2 punti)

```
import numpy
arr = [1,1,1]
numpy.dot(arr,arr.T)
```

C - Pandas (14 punti)
Supponi di avere un Pandas DataFrame in una variabile *df*:

	Col1	Col2	Col 3
Row1	Y	0.4	20
Row2	N	0.9	10
Row3	N	NaN	3

Scrivi un comando che:

C1) Estragga un DataFrame con solo le ultime due colonne (1 punto)

C2) Selezioni seconda riga e colonna 'Col3' (attenzione questo è specifico, selezionate per label) (2 punti)

C3) Rimuova la colonna contenente il NaN (assumendo che il NaN sia in una forma riconoscibile dai metodi di pandas, come ad esempio un np.nan) (2 punti)

C4) Se volessi modificare il DataFrame **senza crearne una copia** da assegnare ad una variabile, ad esempio rimuovendo i NaN, qual è l'opzione da utilizzare e dova va inserita? (2 punti)

C5) Raggruppa gli elementi in Col1 poi media gli elementi in Col3 (per gruppo) (3 punti)

C6) Restituisce gli elementi in ‘Col 3’ che hanno ‘N’ in ‘Col 1’ (3 punti)

C7) Che differenza c'è tra la notazione **df.col** e **df['col']** per selezionare una colonna di un DataFrame? (1 punto)