Hw2-2 report

Leader: 陳偉 R06944043

Teammate: 周思佳 R07922146，陈熙 R07922151

Private Score | Public Score： 0.32666 | 0.33052

I. Preprocessing & negative sampling:
- Feature 為 user 對於各 food 的[次數, 上一次吃的时间间隔，上上次吃的时间间隔, …]，
- 對於各個 user，將出現距離當前最遠的 food 作為 negative sample，y 值設為 0，距離當前最近的 food 最為 positive sample，y 值設為 1。

II. Model：
- Linear regression，因 most frequency 的結果要好于 Matrix Factorization，猜測 frequency 和時間中含有的信息量很高。因此對 user food pair 的次數，時間間隔作為 feature，線性回歸。

III. Other Experiment：

尝试了 1. MF; 2. most frequency, most recent; 3. 將 MF 的 score 作為 feature; 4. 使用user 和 food feature, 5. 增加 data sparsity, 6. DeepWalk。
- 如同 hw2-1 的作法，對 user food 的 occurrence matrix 進行矩陣分解，結果為 0.14483，顯然 performance 非常低
- most frequency，猜測結果為該 user 歷史食用食物次數的 ranking，結果為 0.28871；most recent，猜測結果為該 user 過去吃過距離現在時間的 ranking，結果為 0.20752，可以看到，frequency 和 time 含有相當高的信息量。
- 將 MF 的 score 作為 Linear regression 的 feature， 結果為 0.26809，這比不用的結果 0.32666 低很多。
- user 和 food feature，不管是加入全部 feature 還是部分 dense 的 feature，都使 performance 降低。
- 增加 data sparsity, 統計了下 occurrence matrix 的sparsity, 發現 sparsity = 3.6 %，顯然低於老師上課講的 10% 的 threshold，於是想到通過刪除掉 一些不常出現的 food，來增加 sparsity。結果表明 performance 沒有上升，猜想助教有對 data 做過處理，剩下的 food 都是頻率較高的，所以刪除會損失 performance。
- DeepWalk，屬於突發奇想，隨便亂試，考慮到 deep walk 學到使相連的 node 之間的 embedding similarity 越高，因此想到把 user 和 food 都當作 node，user 有吃過 food 就作為中間有 path，去訓練，結果performance 很差 (Private | Public：0.00082 | 0.00076)，推測理由是，1. 資料中 user 和 food 都是互相高頻出現的，因此任何 user 和 food 都兩兩有 path，導致產生的 embedding 沒有區分度，計畫未來將 path 設計得有差異性； 2. Matrix 有其結構性，而 graph 沒有，正因為其缺少結構性，才折衷使用 random walk，用非 optimal 的方法去訓練結構性的資料，沒有好好利用其結構的特徵，自然 performance 會相較更差。