Data Mining

Homework 5

Due: First appello (June 25): 24/6/2014, 23:59
All other appelli: 13/7/2014, 23:59

You must hand in the homeworks electronically and before the due date and time. Check the web page for instructions about collaboration, about being late, and about handing in the homework.

Problem 1. Your band is missing a drummer. To hire one you put an announcement and, after the replies, you schedule n interviews from some drummers who were intereted in joining your band. Assume that the candidate drummers are coming for an interview in a uniformly random order. Your goal is to get the best drummer for your group. Whenever a drummer comes, you want to decide on the spot: if after the interview you decide to hire the drummer, you tell him so and you stop interviewing other candidates. Otherwise, you tell him that you don't want to hire him, and you procede with the next candidate, without being able to change your decision later.

We consider the following strategy. First you interview $\ell < n$ drummers and you reject them all. This will give you an idea of how good the candidates are. After the ℓth candidate you hire the first drummer that is better than all the previous drummers (or the last one, if you cannot find a better one).

We want to see what value of ℓ (as a function of n) maximizes the probability that we hire the best drummer.

1. Let E be the event that we hire the best drummer, and let E_i be the event that the ith drummer is the best and we hire him. Compute $\mathbf{Pr}(E_i)$ and then show that
$$\mathbf{Pr}(E) = \frac{\ell}{n} \sum_{i=\ell+1}^{n} \frac{1}{i-1}.$$
2. Prove the following two relations:
   - $\displaystyle\sum_{j=a}^{b-1} \frac{1}{j} \geq \ln b - \ln a$
   - $\displaystyle\sum_{j=a+1}^{b} \frac{1}{j} \leq \ln b - \ln a$

   (**Hint:** Relate the integral $\int_a^b \frac{1}{x} \mathrm{d}x$ with the above summations.)
3. Use Part 2 to prove that $\sum_{i=\ell+1}^{n} \frac{1}{i-1}$ to show that
$$\frac{\ell}{n}(\ln n - \ln \ell) \leq \mathbf{Pr}(E) \leq \frac{\ell}{n}(\ln(n-1) - \ln(\ell-1)).$$
4. Show that $\ell(\ln n - \ln \ell)/n$ is maximized when $\ell = n/e$, and explain why this means that $\mathbf{Pr}(E) \geq 1/e$ for this choice of ℓ.

Problem 2. As we said in the class (and is written in the notes), the k-means algorithm tries to minimize the value of the objective:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i)^2,$$

where $C_1, \ldots, C_k$ correspond to the k clusters with C_i being the set of points that belong to the ith cluster, and $\boldsymbol{\mu}_i$ is the center of the ith cluster (the average of all the points in C_i).

Even though often the k-means algorithm gives good solutions, it may not converge to the optimal, and it may give bad clusters.

1. One way to initialize the k-means algorithm, is by selecting k random points as centers. Construct an example where the k-means algorithm with this initialization rule, may converge to a bad solution, a solution that may be far from the best solution (the one that minimizes the above expression).

2. Another way to initialize the algorithm, is by selecting as a center a point at random and then keep selecting centers by selecting at each time as a center the point that is as far as possible from all the previously selected centers, till we select k centers. Again, give an example that shows that even using this rule may lead to bad solutions.

Problem 3. In this problem we will see a problem related to classification. Consider the following scenario: You wan to place bets on a series of football games and you can ask for the help of n people who are supposed to be specialists. (For simplicity, let us ignore ties. Assume that you bet on only on one of the teams winning, and you win if the team wins, you lose if the team loses, and in the case of tie you neither win nor lose.)

You don't know who of the specialists is the best, so you think a bit and you come up with the following procedure for deciding, which will discount the opinion of specialists who make mistakes. For each of the n specialists you maintain a weight w_i, for $i = 1, \ldots, n$. Initially $w_i = 1$ for all i. When you need to decide, you ask for the opinions and you sum the weights of those that say *Team A* will win and those that say that *Team B* will win, and you bet based and the majority. Then, for the specialists that were wrong, you half their weight and you keep continue.

In detail, your procedure is the following. Assume that there are ℓ matches.

```
1. For each $i = 1, \ldots, n$, let $w_i = 1$
2. for $j = 1$ to $\ell$
    (a) Ask the $n$ specialist to make a prediction for the $j$th match
    (b) Let $S_A$ ($S_B$) be the set of specialists that say that team $A$ ($B$) will win
    (c) $W_A = \sum_{i \in S_A} w_i$
    (d) $W_B = \sum_{i \in S_B} w_i$
    (e) If $W_A \geq W_B$ then bet for team $A$, else bet for team $B$
    (f) If team $A$ won, then for each $j \in S_B$ let $w_i = w_i/2$
    (g) If team $B$ won, then for each $j \in S_A$ let $w_i = w_i/2$
```

How many mistakes can this procedure make? Assume that the best specialist makes m mistakes.
Prove that this procedure guarantees that you will make at most $O(m + \ln n)$ mistakes.
Hint: Assume that after the ℓ games you have done M mistakes in total, and assume that the total weight of the specialists is W. Show that $W \leq n\left(\frac{3}{4}\right)^M$, and use this fact to show that $M = O(m + \ln n)$.

Problem 4. In class we saw PageRank for measuring the authoritativeness of a node in a network. PageRank is one example of a definition of importance, also known as *centrality*, which is used to capture the relative importance of pages in the web graph. Centrality is a much more general concept however, and is useful broadly in networks. For example, we may be interested in knowing how important a person is in a social network, or how important a road is in a citys transportation infrastructure. As you might expect, there are a number of different ways of defining the centrality of a node in a network, and each might induce a different ranking of the nodes in terms of *importance*. How appropriate any one definition is depends on the application scenario. In this exercise, we will introduce you to the four most common measures of centrality.

Lets start with the definitions. Consider a connected, undirected graph G with n nodes, labeled $1, 2, \ldots, n$.

1. **Degree centrality** Let d_i denote the degree of node i. The degree centrality of node i is defined as
$$C_D(i) = \frac{d_i}{n-1}.$$
This definition of centrality assigns importance to a node proportional to its degree.

2. **Closeness centrality** Let $\ell(i,j)$ denote the distance (length of the shortest path) between nodes i and j. The closeness centrality of node i, denoted $C_C(i)$, is defined as the reciprocal of the average distance between node i and all other nodes, namely,
$$C_C(i) = \frac{n-1}{\sum_{j \neq i} \ell(j,i)}.$$
So a node has high closeness centrality if its total distance to the rest of the nodes is small.

3. **Betweeness centrality** Let $P(j,k)$ denote the number of shortest paths between nodes j and k, and $P_i(j,k)$ denote the number of those shortest paths that pass through i. One can think of $\frac{P_i(j,k)}{P(j,k)}$ as a measure of the importance of i with respect to connecting j and k. Betweenness centrality attempts to capture how important a node is with respect to connecting other nodes in the graph. The betweenness centrality of node i is defined as
$$C_B(i) = \frac{\sum_{j,k: j,k \neq i, j \neq k} \frac{P_i(j,k)}{P(j,k)}}{\binom{n-1}{2}}.$$

4. **PageRank** Suppose you replace each undirected edge $\{i,j\}$ in G by two directed edges (i,j) and (j,i). This gives us a directed, strongly connected graph. On this graph, we define the centrality of a node to be its PageRank using teleportation probability equal to 0 (i.e., $\beta = 1$).

Our goal is to compare the different definitions of centrality.

1. Contrast the four definitions of centrality described above. Specifically, for each pair of definitions, either prove that for any connected, undirected graph G, the two definitions rank the nodes in the same order of importance, or give a counterexample that proves otherwise.

   Hint: To get you started, we tell you that the PageRank centrality turns out to be proportional to the degree centrality $C_D(i)$. Of course, you still have to prove this! (Don't be confused though, what were claiming is that for an *undirected* graph, with teleporting probability equal to 0, PageRank is equivalent to degree centrality. This is certainly not true for the web graph, which is directed.)

2. For each of the centrality measures, describe a specific application setting where, in your opinion, this measure is better than the other measures in capturing the notion of importance that we are interested in the application.