Homework Assignments

Information Retrieval 1 [2016/2017]

Module 1: Evaluation

Deadline: Wednesday, January 18th, midnight (23:59)
Collaboration: This is a team-based assignment. Form teams of maximum 2 people.
Submit: An IPython Notebook with the necessary (a) **implementation**, (b) **explanations**, (c) **comments**, and (d) **analysis**. Code quality, informative comments, detailed explanations of what each block in the notebook does and convincing analysis of the results will be considered when grading.
Submit your work through Blackboard. **Both students** in the team need to submit their work.
Filename: <student 1 id>–<student 2 id>–hw1.ipynb

The homework will cover the following three topics covered in Lecture 1, 2, and 3:

- Evaluation measures;
- Interleaving;
- Click models.

Commercial search engines use both offline and online approach in evaluating a new search algorithm: they first use an offline test collection to compare the production algorithm (P) with the new experimental algorithm (E); if *E* statistically significantly outperforms *P* with respect to the evaluation measure of their interest, the two algorithms are then compared online through an interleaving experiment.

For the purpose of this homework we will assume that the evaluation measures of interest are:

1. Binary evaluation measures
    a. Precision at rank k,
    b. Recall at rank k,
    c. Average Precision,
2. Multi-graded evaluation measures
    a. Discounted Cumulative Gain at rank k (DCG@k),
    b. Normalized Discounted Cumulative Gain at rank k (nDCG@k),
    c. Rank Biased Precision with persistence parameter $\theta=0.8$ (RBP), and
    d. Expected Reciprocal Rank (ERR).

Further, for the purpose of this homework we will assume that the interleaving algorithms of interest are:

1. Team-Draft Interleaving (Joachims. "Evaluating retrieval performance using clickthrough data". Text Mining 2003.),
2. Balanced Interleaving (Radlinski, Kurup, and Joachims. "How does clickthrough data reflect retrieval quality?" CIKM 2008.), and
3. Probabilistic Interleaving (Hofmann, Whiteson, and de Rijke. "A probabilistic method for inferring preferences from clicks." CIKM 2011.).

In an interleaving experiment the ranked results of *P* and *E* (against a user query) are interleaved in a single ranked list which is presented to a user. The user then clicks on the results and the algorithm that receives most of the clicks wins the comparison. The experiment is repeated for a number of times (impressions) and the total wins for *P* and *E* are computed.

A Sign/Binomial Test is then run to examine whether if the difference in wins between the two algorithms is statistically significant (or due to chance). Alternatively one can calculate the proportion of times the *E* wins and test whether this proportion, *p*, is greater than $p_0=0.5$. This is called an 1-sample 1-sided proportion test.

One of the key questions however is **whether offline evaluation and online evaluation outcomes agree with each other**. In this homework you will determine the degree of agreement between offline evaluation measures and interleaving outcomes, by the means of simulations. A similar analysis using actual online data can be found at Chapelle et al. "Large-Scale Validation and Analysis of Interleaved Search Evaluation".

[Based on Lecture 1]

Step 1: Simulate Rankings of Relevance for *E* and *P (5 points)*

In the first step you will generate pairs of rankings of relevance, for the production *P* and experimental *E*, respectively, for a hypothetical query **q**. Assume a 3-graded relevance, i.e. {N, R, HR}. Construct all possible *P* and *E* ranking pairs of length 5, ~~for~~ ~~which~~ ~~*E*~~ ~~outperforms~~ ~~*P*~~. <= (10/1, 11:06: Postpone the discarding to Step 3)

Example:
P: {N N N N N}
E: {N N N N R}
…
P: {HR HR HR HR R}
E: {HR HR HR HR HR}

(Note 1: If you do not have enough computational power, sample 1000 pair uniformly at random to show your work.)

Step 2: Implement Evaluation Measures *(15 points)*

Implement 1 binary and 2 multi-graded evaluation measures out of the 7 measures mentioned above.

(Note 2: Some of the aforementioned measures require the total number of relevant and highly relevant documents in the entire collection – pay extra attention on how to find this)

Step 3: Calculate the *measure (5 points)*

For the three measures and all *P* and *E* ranking pairs constructed above calculate the difference: Δmeasure = measure_E-measure_P. Consider only those pairs for which *E* outperforms *P*.

[Based on Lecture 2]

Step 4: Implement Interleaving (*15 points*)

Implement 2 interleaving algorithms: (1) Team-Draft Interleaving OR Balanced Interleaving, AND (2), Probabilistic Interleaving. The interleaving algorithms should (a) given two rankings of relevance interleave them into a single ranking, and (b) given the users clicks on the interleaved ranking assign credit to the algorithms that produced the rankings.

(Note 4: Make your implementation as generic as possible such that it could work outside this assignment.)

(Note 5: Note here that as opposed to a normal interleaving experiment where rankings consists of urls or docids, in our case the rankings consist of relevance labels. Hence in this case (a) you will assume that E and P return different documents, (b) the interleaved ranking will also be a ranking of labels.)

[Based on Lecture 3]

Step 5: Implement User Clicks Simulation (*25 points*)

Having interleaved all the ranking pairs an online experiment could be ran. However, given that we do not have any users (and the entire homework is a big simulation) we will simulate user clicks.

We have considered a number of click models including:

1. Random Click Model (RCM)
2. Position-Based Model (PBM)
3. Simple Dependent Click Model (SDCM)
4. Simple Dynamic Bayesian Network (SDBN)

Consider two different click models, (a) the Random Click Model (RCM), and (b) one out of the remaining 3 aforementioned models. The parameters of some of these models can be estimated using the Maximum Likelihood Estimation (MLE) method, while others require using the Expectation-Maximization (EM) method. Implement the two models so that (a) there is a method that learns the parameters of the model given a set of training data, (b) there is a method that predicts the click probability given a ranked list of relevance labels, (c) there is a method that decides - stochastically - whether a document is clicked based on these probabilities.

Having implemented the ~~three~~ two click models, estimate the model parameters using the Yandex Click Log [file].

(Note 6: Do not learn the attractiveness parameter a_{uq}.)

Step 6: Simulate Interleaving Experiment (*10 points*)

Having implemented the click models, it is time to run the simulated experiment.

For each of interleaved ranking run N simulations for each one of the click models implemented and measure the proportion *p* of wins for E.
(Note 7: Some of the models above include an attractiveness parameter a_{uq}. Use the relevance label to assign this parameter by setting a_{uq} for a document u in the ranked list accordingly. (See Click Models for Web Search)

Step 7: Results and Analysis (*25 points*)
Compare the results of the offline experiments (i.e. the values of the *measure*) with the results of the online experiment (i.e. proportion of wins), analyze them and reach your conclusions regarding their agreement.

- Use easy to read and comprehend visuals to demonstrate the results;
- Analyze the results on the basis of
    - the evaluation measure used,
    - the interleaving method used,
    - the click model used.
- Report and ground your conclusions.

(Note 8: This is the place where you need to demonstrate your deeper understanding of what you have implemented so far; hence the large number of points assigned. Make sure you clearly do that so that the examiner of your work can grade it accordingly.)

Yandex Click Log File:
The dataset includes user sessions extracted from Yandex logs, with queries, URL rankings and clicks. To allay privacy concerns the user data is fully anonymized. So, only meaningless numeric IDs of queries, sessions, and URLs are released. The queries are grouped only by sessions and no user IDs are provided. The dataset consists of several parts. Logs represent a set of rows, where each row represents one of the possible user actions: query or click.

In the case of a Query:
SessionID TimePassed TypeOfAction QueryID RegionID ListOfURLs

In the case of a Click:
SessionID TimePassed TypeOfAction URLID

SessionID - the unique identifier of the user session.
TimePassed - the time elapsed since the beginning of the current session in standard time units.
TypeOfAction - type of user action. This may be either a query (Q), or a click (C).
QueryID - the unique identifier of the request.
RegionID - the unique identifier of the country from which a given query. This identifier may take four values.
URLID - the unique identifier of the document.
ListOfURLs - the list of documents from left to right as they have been shown to users on the page extradition Yandex (top to bottom).