Assignment 7 **Team: Technopreneurs**
Data Science Project Date: 01 April 2015

Problem Statement

Are all responses equally important in a Crowd-sourcing experiment? How to evaluate the expertise of users for a task?

Literature Survey

While using crowd to solve problems variety of platforms like Amazon Mechanical Turk are used. Many users with varying level of expertise in various fields solves the tasks. How do we decide that which responses should be given more importance?

Generative model of Labels, Abilities, and Difficulties (GLAD) [1] was the system developed for this purpose in the field of image tagging. It was based on ”standard probabilistic inference on a model of the labeling process”. With some computation cost it could also handle a million of parameters.

CrowdSense [2] was a similar approach to solve the problem of sampling subset of labelers based on some criterion to generate a weighted combination of votes to generate the approximate opinion of crowd.

Yan et. al. [3] and Donmez et. al. [4] also discuss this problem and try to estimate the expertise level of the annotators/labelers.

Experiment Proposed

Create two kinds of tasks depending on our knowledge of the answers.

1. Tasks for which we know the answers.

2. Tasks for which we don’t know the answers.

For the first case, we can take some images and asks users to label them. Guessing movie release year based on its poster and guessing country name based on flag can be used for this experiment. It will ensure that the expertise of same user on two different tasks is evaluated simultaneously.

For the second case, we can ask users to predict which party will get more votes in presidential elections in 2016 and how many goals will be scored in a match of MLS 2015. We can do the analysis of these results and match them afterwards with the actual result. Hence each user expertise will also be evaluated on two more kinds of areas.

So by asking each user these 4 questions we can collect data and evaluate their expertise on 4 different fields. And for each task, give different weights to their answers while generating the final answer.

Data Analysis

1. **Solution when some tasks are already solved:**

   While collecting data for a task include some tasks for which we know the correct answer. Lets call these tasks Expertise Evaluation Tasks (EET). These EET will also have varying level of difficulties. Based on answers of these EETs we can evaluate the expertise of the users. For example, if some user completes 2 very difficult tasks correctly, we can give more weights to his responses as compared to that of of some user who could not complete any of those 2 tasks.

2. **Solution when tasks are completely new:** Of all the solutions that we have studied, the one implemented in CrowdSense [2] seems to be the best to implement. It solves this problem by predicting the quality of a labeler by calculating the level of agreement of its annotations with that of the whole crowd.

References

[1] J. Whitehill, P. Ruvolo, T. fan Wu, J. Bergsma, and J. Movellan, "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 2035-2043.

[2] Seyda Ertekin, Haym Hirsh, and Cynthia Rudin. *Approximating the Wisdom of the Crowd.* Computational Social Science and the Wisdom of Crowds, Second Workshop on. 2011.

[3] Y. Yan, R. Rosales, G. Fung, M. W. Schmidt, G. H. Valadez, L. Bogoni, L. Moy, and J. G. Dy, "Modeling annotator expertise: Learning when everybody knows a bit of something," *Journal of Machine Learning Research - Proceedings Track (JMLR)*, vol. 9, pp. 932-939, 2010.

[4] P. Donmez, J. G. Carbonell, and J. Schneider, Efficiently learning the accuracy of labeling sources for selective sampling, in *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009, pp. 259-268.