Can higher payments make human image classification more accurate?

Ryan Delgado – W241 Essay 2

Introduction

Crowdsourced labor has become a popular solution for creating training datasets for machine learning models. Online platforms provide an easy, relatively inexpensive method for data scientists and researchers to outsource the labeling of raw datasets, as opposed to hiring contractors to label the data. It's become a popular method for creating datasets, as machine learning researchers and practitioners have been using web-based crowdsourced platforms for generating training datasets for almost a decade now.

Amazon Mechanical Turk (AMT) is one such crowdsourced platform for publishing Human Intelligence Tasks (HITs) to thousands of potential workers. Tasks are generally small, such as labeling or tagging images, taking surveys, or transcribing audio. The workflow is simple – a Requestor posts an HIT containing raw data and a question about the data (e.g. what is it?), and Workers complete the HIT for a small reward. The payment for an HIT can range from a few cents for a task that takes less than a minute, to as much as $20 for an hour and a half long task. AMT seems to be the dominant online crowdsourced labor platform used for building online datasets.

Data Quality is often a concern, however, when building training datasets with AMT. Given that training a supervised machine learning model on data assumes that it's correct in the first place, incorrectly labeled observations will hurt the model's accuracy on unseen data. Many Requestors report poor quality results if provisions aren't taken to filter out spammers or Workers with little knowledge of the subject. Requesters have several common options to improve quality: implement a "qualifier" task that Workers must pass before beginning the actual HIT, allowing workers to complete the HIT once, and denying Workers with a poor reputation from previous HITs. However, there's little literature on one of the more obvious knobs Requestors can adjust: paying Workers more. Is it possible that offering larger monetary rewards motivates Workers to do a better job? This experiment will seek to answer this question with a mock supervised learning task – labeling breeds of dogs. It will try different levels of cash rewards for the HITs to explore nonlinearities in the relationships between payouts and quality.

Background & Motivation

Supervised learning is a machine learning task where a function is inferred from labeled training data. This function can then be used to predict the outcome of new, unlabeled data. Classification is a particular kind of supervised learning where a model predicts one of multiple outcomes or classes given input information. Classic examples include spam classification, speech transcription, and image tagging.

Having quality data is critical to the success of supervised learning approaches, as the model will not be able to learn the correct representation of a cat, for example, if pictures of balloons are incorrectly labeled as cats. The training dataset must also be sufficiently large in order to give the model enough information about what a cat looks like. Collecting a large, correctly-labeled training dataset is often the most costly and time-consuming part of the machine learning process. Practitioners often turn to online crowdsourced mediums for labeling training data.

AMT and other similar platforms are a natural choice for building large training datasets. With a workforce size in the hundreds of thousands, it can be an inexpensive and quick way to build a large training dataset. However, some Requestors report inconsistencies in quality of the resulting data. Sometimes the HIT is completed by spammers, who thoughtlessly and quickly complete the task in order to get fast payment. Additionally, many of the HITs involve some expertise in specialized subjects that not all Workers have. Opening such HITs to the public inevitably results in subject novices attempting the task and producing poor quality data.

Experiment Setup

The supervised learning task will be framed as a classification problem where the model is classifying the breed of a dog based on an image. Since this is a pilot experiment, it will only focus on 8 different dog breeds. In each HIT, the Worker will be shown example images for each breed, and then 40 images of dogs that will need to be classified with multiple choice answers.

A dataset of 200 images of dogs will be compiled from images from the Web. The research team will individually label the breed of each picture and agree on each observation. The dataset will be split into four 50-question multiple choice tests. These tests will be built using SurveyMonkey, and will be split into a qualifying pre-test and an actual test. The tests will be of equal difficulty. Funds from each researcher will be pooled into one AMT account and used for the cash reward during the experiment. There will be four levels of cash reward in the experiment: $0.10, $0.20, $0.40, and $0.80. We're experimenting at multiple levels instead of just two levels in order to explore non-linear treatment effects of marginal increases in cash rewards.

Experimental Risks

There are a few risks that we could encounter when conducting this experiment:

- Self-selection bias related to the different dollar values
- Clustering effects related to dog ownership rates in different countries
- Violations in non-interference if a worker participates in the study at multiple levels
- Subjectivity in determining the dog breeds.

One risk in this experiment is selection bias. Perhaps the HITs with lower rewards are more likely to attract spammers and lower quality workers. If we don't have the same approximate skill level and drive for quality work across the reward levels, it will bias the experiment and lead us to overestimate the

treatment effect. We can mitigate this risk by first giving Workers a short, 8 question qualifier test that Workers must get 5 out of 8 correct on in order to proceed to the paid HIT.

We could also encounter clustering effects when making the test open to workers across countries and at varying levels of dog ownership. Not all workers are from countries were dogs are standard household pets. India is one such example, where it's relatively uncommon to even see different breeds of dogs, let alone have one as a household pet. We could account for this effect by either 1. Restricting out sample to US workers who have not owned dogs before or 2. Including questions in the test about which country the worker is from and if they've owned a dog or not and factoring these into our analysis. Option 2 is preferable, as we want to use a representative sample of the population in this experiment, and approximately a third of all AMT workers are from India. Additionally, clustered standard errors will be used in the regression to make our inferences robust to clustering effects.

The non-interference assumption is also at risk if the same worker participating in the trial at multiple dollar value levels. This may result in the worker improving their skill at discerning between dog breeds as the continue taking tests at increasing reward levels. In a sense, this would violate this key assumption in our experiment, as one worker will become multiple "subjects" in the study if they complete more than one HIT. This interference would lead us to overestimate the treatment effects of higher monetary rewards. We can eliminate this effect by restricting the HIT of later tests to only workers that haven't taken the previous test(s). Additionally, the experiment will not proceed in the order of the dollar values, we'll do $0.10, $0.80, $0.20, and $0.40. This additional step will guarantee that the monetary reward isn't correlated with the ordering of the experiment.

The final risk we could encounter is subjectivity in breed judgement. Some breeds, such as greyhounds and whippets, look similar to other breeds and it will be difficult to determine the breed of the dog. We can counteract this by choosing a subset of dog breeds that are unquestionably distinct from each other, such as pugs, great danes, daschunds, and chihuahas. The research team will also decide on a set of pictures of pure bred dogs that unquestionably belong to a particular breed. These measures should minimize this risk.

Statistical Analysis

Given that both the outcome variable, accuracy score, and treatment variable, money, are both continuous (though reward is discretized for simplicity), the experiment results will be analyzed with a regression. Since country of residence and dog ownership could be additional variables that contribute to accuracy score, we will also include those in the regression:

$$Score_i = \beta_0 + \beta_1 Reward_i + \beta_2 India_i + \beta_3 OtherNonUSCountry_i + \beta_4 DogOwner_i + \varepsilon_i$$

Hypothesis tests will be performed on the beta coefficients at the 95% confidence level to determine if there's a relationship between the regressors and outcome variable.

Experiment Process

The experiment will proceed as follows:

I. Preparation

a. Collect and label 200 pictures of pure-bred daschunds, chihuahuas, great danes, pugs, greyhounds, golden retrievers, rottweilers, and pit bulls.
b. Create 4 SurveyMonkey multiple choice tests that contain 10 "qualifier" questions and 40 real questions. Each test should be equally difficult as the others, and should include an initial question about which country the worker lives in.

II. Set up

a. Create, but do not publish, the tests in Mechanical Turk.
b. Configuring the tests in such a way that prevents users that have taken previous tests from taking a new test. Also configure the tests to be available to all workers on AMT.

III. Experiment

a. Publish the Conduct the first experiment using $0.10 as a reward.
b. Publish the second HIT using $0.80 as the reward.
c. Publish the third HIT using $0.20 as the reward.
d. Publish the last HIT using $0.40 as the reward.

IV. Analysis

a. Regress the accuracy score of each participant on the reward amount, country of residence, and dog ownership.
b. Collect & report results

Future Directions

This experiment will shed light on whether or not AMT requestors can use payment as an additional knob to turn increase quality of results from HITs. The experiment process could also provide a blueprint for future training set building endeavors to determine the optimal reward for their Mechanical Turk HITs. It could be a "pilot" study that researchers conduct before the actual dataset building commences in order to determine the optimal "hyperparameters" for maximizing data quality. Future research could look into wider ranging or more granular monetary rewards, or could experiment with HITs that require different levels of expertise like language translation.