Fidelity-Weighted Learning

Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, Bernhard Scholkopf

1University of Amsterdam, 2MPI for Intelligent Systems, 3Google Brain

Learning from Samples of Variable Quality

- For a large class of problems, we can reduce un/semi-supervised learning to learning from weak labels.
- A weak annotator can be:
  - a heuristic or a rule-based model
  - a weaker classifier
    - e.g. trained on non-expert crowd-sourced data (to which we don’t have access)
  - a biased classifier
    - e.g. trained on data from different domains

All samples are equal, but some samples are more equal than others.

Fidelity-Weighted Learning

Step 1
Pre-train the student on weak dataset using weak labels generated by the weak annotator.

The main goals of this step are:
- To learn a task-dependent representation of the data:
  - Learning a representation function \( f(x) \)
- To pretrain the student.

Step 2
Train the teacher on the strong data represented in terms of the student representation, and then use the teacher to generate a soft dataset consisting of (sample, predicted label, confidence) for all data samples.

The GP is trained on the strong data to learn the posterior mean (used to generate soft labels) and posterior covariance (which represents label uncertainty).

Soft labels: a new set of (hopefully less weak) labels for which we know the certainty.

Step 3
Fine-tune the weights of the student by training on the soft dataset generated by the weak annotator.

We update the parameters of the student by training on the soft dataset using SGD:

\[
\theta' = \theta - \sum_{i=1}^{n_s} \frac{1}{n_s} \nabla J(y_i, \hat{y}_i, \alpha_i, \beta_i)
\]

where:
- \(\theta\) is the (soft) label
- \(\hat{y}_i\) is the predicted label
- \(\alpha_i\) is the sample confidence/fidelity
- \(\beta_i\) is the sample certainty.

Discussions and Analysis

Handling The Bias-variance Trade-off
- We considered three cases in the toy problem:
  - Having 10 observations from the true function ("Toy Data")
  - Having only 5 observations from the true function ("Toy Data")
  - Having \((x)=x+1\) as the weak function, which is an extremely bad approximator of the true function ("Toy Data =")

Sensitivity Of FWL To The Quality Of The Weak Annotator
- In the ranking task, besides BM25 we used three other weak annotators:
  - Vector space model with binary term occurrence (BTO) weighting schema
  - Vector space model with TF-IDF weighting schema
  - BM25-HM3 that uses BM3 as the pseudo-relevance feedback method on top of BM25 (better than BM25).

Main Results

Toy problem:
- Let \(f(x)\) be the true function from which a small set of observations is provided. (These observations might be noisy, in the same way that labels obtained from a human labeler could be noisy.)
- A weak annotator function \(\hat{f}(x) = f(x) + \text{noise}(x)\) is provided, as an approximation to the true function.
- The task is to obtain a good estimate of the true function given the small set of strong observations and the weak annotator function.

Training student on 50 examples from the weak function.
Fitting teacher based on 10 observations from the true function.
Fine-tuning the student based on soft labels derived from the teacher.

Two real world tasks:

Document Ranking

Sentiment Classification

A Good Teacher Is Better Than Many Observations
- We looked at the rate of learning for the student as the amount of training data is varied:
  - The student learns faster when there is a teacher!
  - A LUPI (learning using privileged information) perspective to FWL.

From Modifying The Learning Rate To Weighted Sampling
- The confidence score can be interpreted as sample quality/fidelity, or alternatively the likelihood of improving the student model’s performance by including that sample in its training set.
- Baseline: Rather than using this score as the multiplicative factor in the learning rate, we can use it to bias sampling procedure of mini-batches so that the frequency of training samples are proportional to the confidence score of their labels.
- FWL has more chance to explore the input space, while it controls the effect of updates on the parameters for samples based on their merit.

inspired by George Orwell, Animal Farm, 1945