Unified View of Label Shift Estimation Presentation

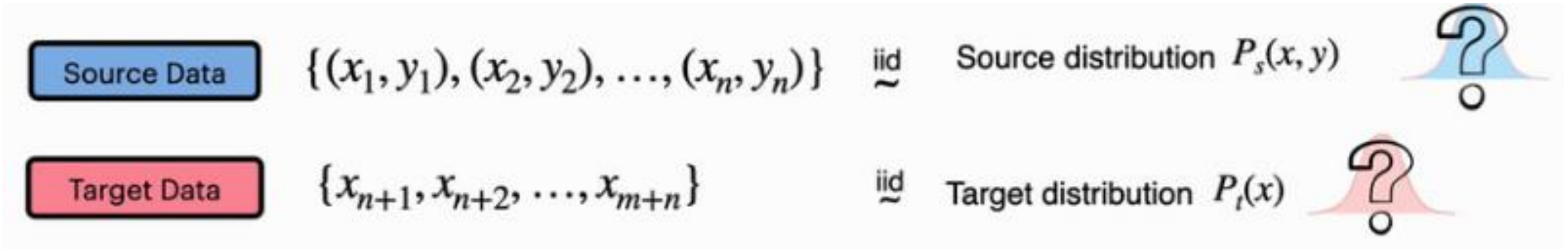
CS 598 Transfer Learning

Daniel Campos

10-13-2021

Agenda

- Background
- Lit Review
- Contributions
- Technical Walkthrough
- Future Work
- Impact



Background/Motivation of Work

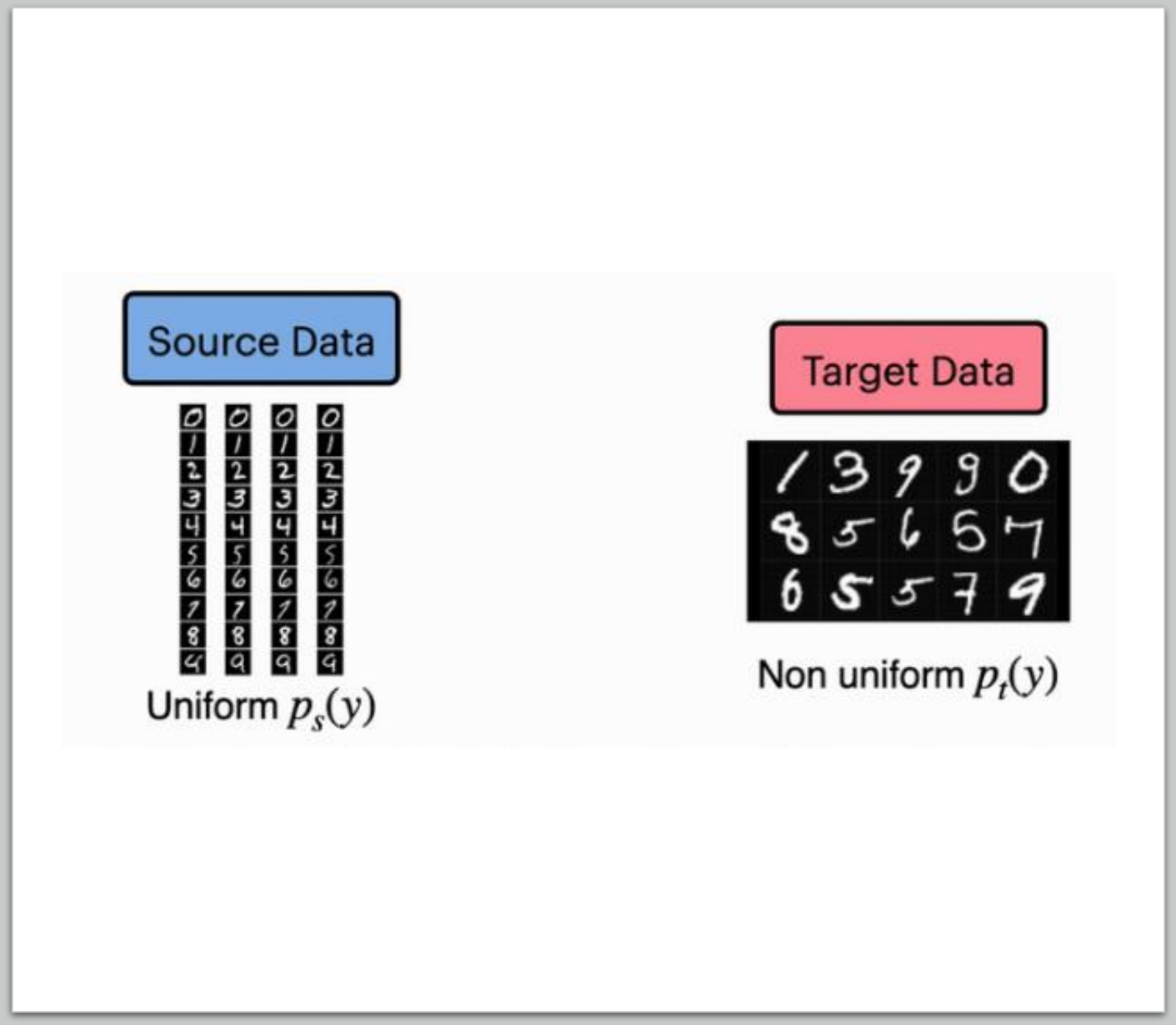
- In machine learning a common paradigm is fitting a model to source data and then inferencing on some unknown target data.
- Assumption here is that any new data matches the distribution of the train data
  - Academic work addresses this by selectic test/dev portions via random sampling from the same distribution which is hopeful at best

Problems

- What if the assumption that the target distribution changes?
  - Temporal Change
    - GPT-3 was trained on web data from 2019. Has no notion of 2020
  - Label Drift
- It is impossible for a classifier to work well on every possible distribution

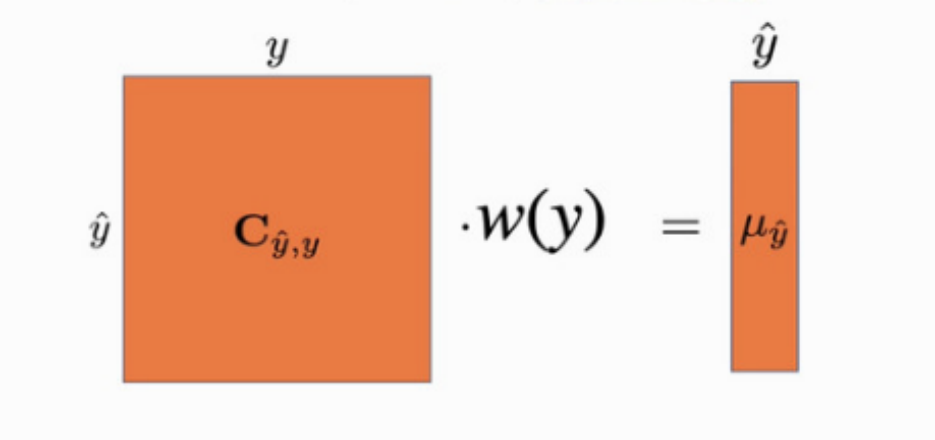
Background

- Assume that p(y) can change but the conditional p(x|y) does not change (aka label shift)
- Important because training data may not be representative of real-world data (rare events may be common in data)



Prior Work

- Black Box Shift Estimation (Saerens et al. ’02) (Lipton et al. ‘18)
  - Leverage Confusion matrix approach
  - Predict the importance of each y by determining w(y) an importance score



Prior Work MLLS

- Maximum Likelihood Label Shift (MLLS) leverages maximum likelihood estimation to predict importance weights (Saerens et al. ’02) (Alexandari, Amr et al. '20)
- Differs from BBSE by using true probabilities of p(y|x) vs the predicted labels.
- No Theoretical guarantees and empirically messy
- Estimates can be inconsistent without tuning

$$\widehat{w}_f := \arg\max_{w \in \mathcal{W}} \frac{1}{m} \sum_{i=1}^{m} \log \left[\hat{f}(x_i) \; w \right]$$

MLLS

- Finds a weight vector by minimizing the KL divergence of the target and the weight vector
- Which is equivalent to

$$\mathrm{KL}(p_t(z), p_w(z)) = \mathbb{E}_t [\log p_t(z)/p_w(z)],$$

$$\tilde{w} := \arg\max_{w \in \mathcal{W}} \mathbb{E}_t \left[\log \sum_{y=1}^{k} p_s(y|z) w_y \right] .$$

- Which can be solved using EM optimization algorithm which can use ground truth(synthetic data) or an estiamtor (neural network)

$$w_f := \arg\max_{w \in \mathcal{W}} \mathcal{L}(w, f) := \arg\max_{w \in \mathcal{W}} \mathbb{E}_t \left[\log f(x)^T w \right] .$$

MLLS Algorithm

Algorithm 1 Maximum Likelihood Label Shift estimation

input : Labeled validation samples from source and unlabeled test samples from target. Trained blackbox model $\hat{f}$, model class $\mathcal{G}$ and loss function l for calibration (for instance, MSE or negative log-likelihood).

1: On validation data minimize the loss l over class $\mathcal{G}$ to obtain $f = g \circ \hat{f}$.
2: Solve the optimization problem (5) using f to get $\widehat{w}$.

output : MLLS estimate $\widehat{w}$

Key Contributions

- Introduce a framework for conecting two differing methods of label shift estimation (MLLS and BBLE) finding them to work better together
- Prove that the loss function of label estimation is not as important as the callibaration of the shift estimation.

MLLS continued and Research Questions

- When f is a simple neural network, it tends to perform poorly but with some post hoc tuning it can outperform prior methods such as BBSE. (Alexandari et al. '19)
- Why is callibaration so important for MLLS?
- Is calibration necessary for consistency of MLLS?
- Why is MLLS with calibration better than BBSE?

Why Calibration works

MMLS with a calibrated predictor is equivalent to performing distributional matching on the latent space Z.

Lemma 2. *If f is calibrated, then the two objectives* (3) *and* (4) *are identical when $\mathcal{Z}$ is chosen as Δ^{k-1} and $p(z|x)$ is defined to be $\delta_{f(x)}$.*

Can calibration produce consistency?

- Consistency is driven by the sample-based estimator.
- Consistency of MLLS relies on the linear indepence of the distribution of all classes of y.

Theorem 1 (Population consistency of MLLS). *If a predictor $f : \mathcal{X} \mapsto \Delta^{k-1}$ is calibrated and the distributions $\{p(f(x)|y) : y = 1, \ldots, k\}$ are strictly linearly independent, then w^* is the unique maximizer of the MLLS objective* (4).

Proposition 1. *For a calibrated predictor f, the following statements are equivalent:*

(1) $\{p(f(x)|y) : y = 1, \ldots, k\}$ *are strictly linearly independent.*

(2) $\mathbb{E}_s\left[f(x)f(x)^T\right]$ *is invertible.*

(3) The soft confusion matrix of f is invertible.

Is MLLS + Calibration better than BBSE

- Label shift estimation is one of two
  - Define a latent space p(z|x)
  - Distribution matching on a new latent space Z
- BBSE designs a latent space p(z|x) using confusion matrix and then distribution matching by solving linear equations
- MLLS does not define how to find a predictor but uses KL divergence minimization to find distributional match.
- If one applies the confusion matrix approach for MLLS aproach for latent space creation the MLLS becomes a special case of BBSE.

Theorem: (Error of MLLS)

If f satisfies a regularity condition, we have

$$\left\| \widehat{w}_f - w^* \right\| \leq \sigma_{f,w_f}^{-1} O_p\left(m^{-1/2}\right) + C \cdot \sigma_{f,w^*}^{-1} \mathscr{E}(f)$$

where $\mathscr{E}(f)$ is the calibration error of the predictor f

Theorectical Analysis of MMLS

- Assuming MLLS is perfectly calibrated since it leverages a sample-based approximation it cannot estimate true shift (finitie sample error) nor is there any guarentee that the predictor is perfectly calibrated on the source distribution (miscallibration error)

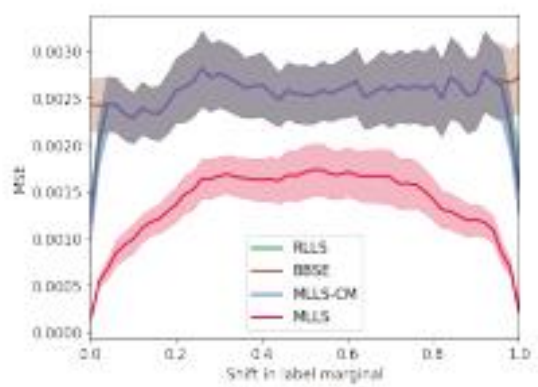
Experiments

- Explore the effect of label shift with MNIST (Digit Recognition) and CIFAR-10(Object reconition).
  - Note that both datasets are simple 10 class classification
- For each run target label distribution is sampled with a concentration parameter which changes the label distribution
- Training data is designed to have a uniform label distribution.
- Predictor is a RestNet-18 (He et al. '16)
- 100 datasets are sampled for each shift parameter and the Mean Squared Error is evaluated with respect to the variance in estimated weights.

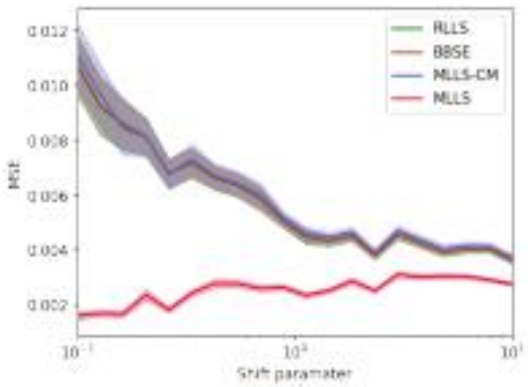
Experiments Described in Plain English

- Train a image classifier using a uniform distribution and then generate artificially skewed test samples.
- Evaluate how well the skewed label distribution can be predicted by comapring the true distribution to the predicted distibution using MSE.
- Experiment on the following variables
  - Degree of shift
  - Sample size
  - Calibrated predictiors
    - MLLS, BBSE, MLLS-CM, RLLS

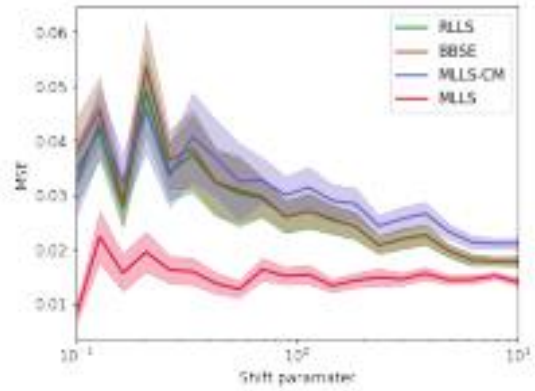
Results



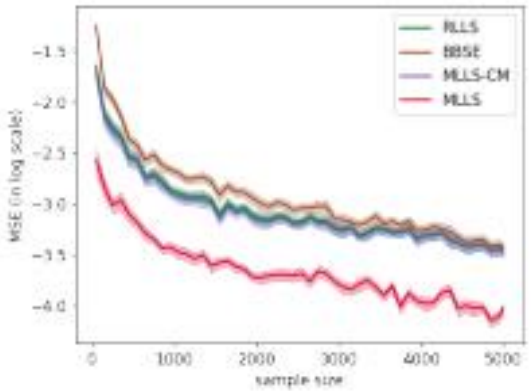
(a) GMM



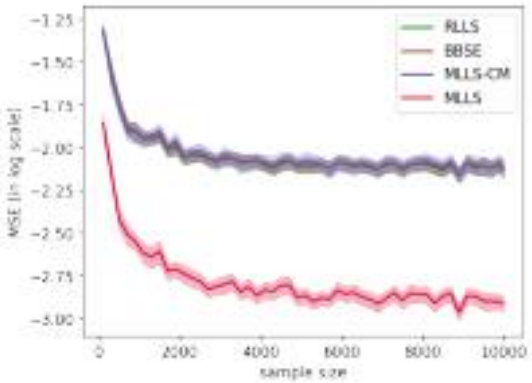
(b) MNIST



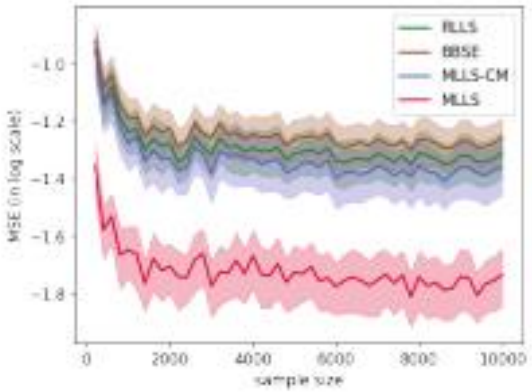
(c) CIFAR-10



(d) GMM



(e) MNIST



(f) CIFAR-10

Future Work

- Provide Theoretical Guarentees and provide guidance on how to calibrate shift estimation

Applications of Research

- **Online Adaptation to Label Distribution Shift (**Wu et al.' 21)
  - **World changes while model is in production so allow it to change online**
- **Coping with Label Shift via Distributionally Robust Optimisation (**Zhang et al. '21)
  - **Models have varying error rates across classes so adversarial inputs may modify the distribution to take advantage of model skills**
- **Active Learning under Label Shift(**Zhao et al.' 21)
  - **Leverage Label shift to inform active learning**

Questions?

References

1. Lipton, Zachary Chase et al. “Detecting and Correcting for Label Shift with Black Box Predictors.” *ArXiv* abs/1802.03916 (2018): n. pag.
2. Saerens, Marco et al. “Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure.” *Neural Computation* 14 (2002): 21-41.
3. Azizzadenesheli, Kamyar et al. “Regularized Learning for Domain Adaptation under Label Shifts.” *ArXiv* abs/1903.09734 (2019): n. pag.
4. Alexandari, Amr et al. “Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation.” *ICML* (2020).
5. Shrikumar, Avanti et al. “Adapting to Label Shift with Bias-Corrected Calibration.” (2019).
6. He, Kaiming et al. “Deep Residual Learning for Image Recognition.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 770-778.
7. Wu, Ruihan et al. “Online Adaptation to Label Distribution Shift.” *ArXiv* abs/2107.04520 (2021): n. pag.
8. Zhang, J. et al. “Coping with Label Shift via Distributionally Robust Optimisation.” *ArXiv* abs/2010.12230 (2021): n. pag.
9. Zhao, Eric et al. “Active Learning under Label Shift.” *AISTATS* (2021).