10732 Structured Approaches to distribution shift + the label shift story

Refresher: standard prediction setting

Source data $\{(x_i, y_i)\}_{i=1}^{n} \overset{iid}{\sim} P(X, Y)$

Deployment/target data $\{(x_j, y_j)\}_{j=1}^{m} \overset{iid}{\sim} P(X, Y)$

(observe only $\{x_j\}_1^m$)

Because $P_S = P_+$, licensed to evaluate on held out set of source data (exchangeable w deployment data).

Here, fitness of model on target dist is directly observed, no identification problems. Licensed to try (almost) anything, see what works

What can we do when $P_{src} \neq P_{tgt}$?

Some lenses:

- theoretical models
  - distributional robustness over uncertainty sets
  - identification under structure
- Deep learning playbook
  - benchmarks + tricks
- Look out at the real world

Themes:

- All assumptions are too strong
- Problems ill-posed absent assumptions
- Lack a distribution over distributions
- Effective sample size
- What role do theory, benchmarks, etc. play?

Structure of a distribution Shift Problem:

- Set of environments / domains
- Underlying structure shared across environments
- Observability - what variables seen where?
- Manipulation rules:
  - what can change?
  - in what way?
  - by how much?
  - in- vs. out-of-support?
- Statistical / inferential capabilities
  - what relationships can be learned?
  - how effectively?
    - by what measure?
- Objective(s)

Steps:
- Identification
- Estimation / Learning

Note - we are interested in "practical" identification strategies involving only those quantities we have the power to estimate, approximated to the degrees actually attainable

Identification Tips:

- Worry about identification before worrying about practical procedures
- First focus on finite data domains w/ discrete/tabular data.
- Assume direct access to joint dist over observables

(If we can't handle this, we're in trouble)

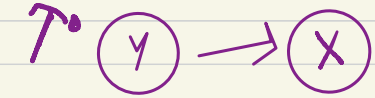
The Move: Leverage blackbox predictors as nonparametric hammers

- Fashion an "id problem" for DL, where fitness can be assessed on appropriate holdout data
- Show that if net has some reasonable property, then we can work in push-forward space.
- Use net + structure to identify/estimate target parameters.

Label Shift

Source: x, y visible

Target: x visible



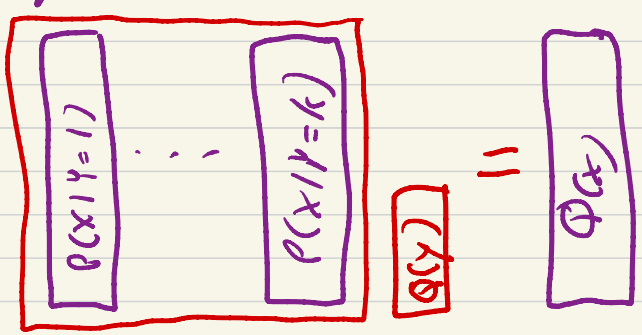
Motivations:
- disease propagation
- general class balance problems

Side note: Test item effect (Zhu et al 2010)

Formally
- $P(X|y) = Q(X|Y)$
- $\text{Supp}(Q(y)) \subseteq \text{Supp}(P(Y))$

Identification:

Everything we see:

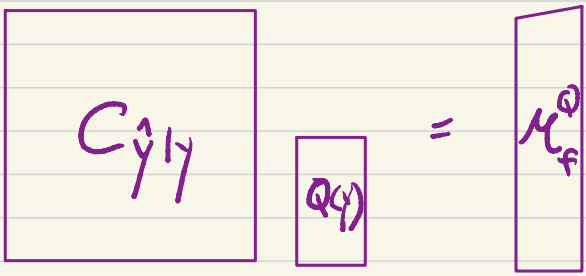


Linear system, solution identified whenever class-conditionals are linearly independent.

$$Q(x) = \sum_y Q(x, y) = \sum_y Q(y) Q(x|y) = \sum_y Q(y) P(x|y)$$

BBSE:

- Train neural net f on source data
- Consider mean classifier output



Detection: $f(x)$ is sufficient to detect label shift

Estimation: $\hat{q}(y) = \hat{C}_{\hat{y}|y}\, \mu_f^Q$

$\hat{w}(y) = \hat{C}_{\hat{y},y}\, \mu_f^Q$

Error bounds: whp, for some constant C,

$$\|\hat{w} - w\|_2^2 \leq \frac{C}{\sigma_{min}^2}\left(\frac{\|w\|\log n}{n} + \frac{k \log m}{m}\right)$$

(SAERENS 2000)

Other estimation ideas using Black Box:

EM / MLLS

1. Start off as though $\hat{P}_t(Y) = \hat{P}_S(Y) \triangleq \hat{Q}(Y)$

For $n = 1 \ldots$

$$\hat{P}_t^n(y_j|x_i) = \frac{\hat{P}_S(y_j|x_i)\frac{\hat{P}_t^n(y_j)}{\hat{P}_S(y_j)}}{\sum_k \hat{P}_S(y_k|x_i)\frac{\hat{P}_t^n(y_j)}{\hat{P}_S(y_k)}}$$

$$\hat{P}_t^{n+1} = \frac{1}{N}\sum_{i=1}^{N} \hat{P}_t^n(y_j|x_i)$$

- Exactly the MLE of $P_T(y)$ when $\hat{P}_S(y|x)$ is the true source conditional, Else, when is this consistent?

  Turns out, requires some invertible CM + canonical calibration (Garg 2020)

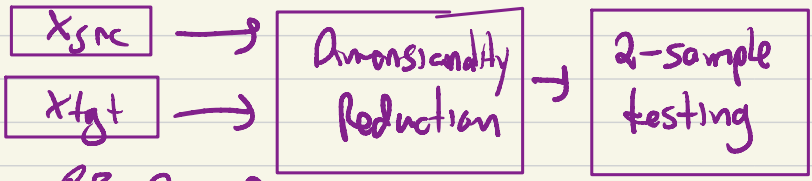
  In practice, works well w BCTS (Alexandri 2019) (temp scaling w per-class parameters)

Error decomp: $\|\hat{w}_t - w\| \leq \|w_t - \hat{w}_t\| + \|w_t - w^*\|$

finite sample, calibration

(Rabanser 2018)

Failing loudly: pipeline for detecting shift



+ BBSD effective broadly, even w LS violated

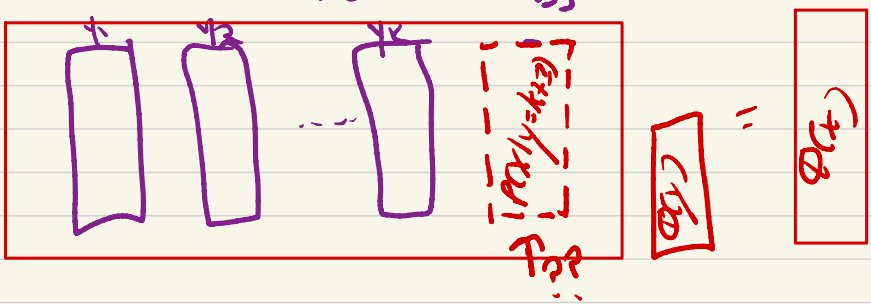
− too strict, there's always a tiny shift, but when should we care?

IW-ERM w DL (Byrd 2019)

$$\min_{\theta} \hat{\sum_{i=1}} w_i \, \ell(f_\theta(x_i), y_i)$$

Problem: what role do weights play when training loss for all examples $\to 0$?

Extending Label Shift to Handle a novel class



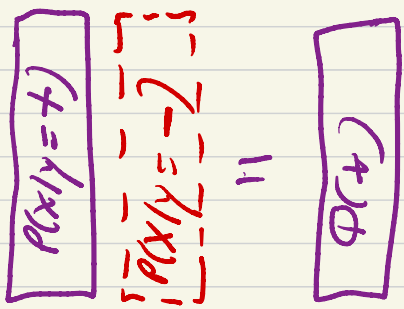
No source data for est. $P(x|y=k+1)$.

Who is to say $P(x|y=k+1) \neq Q(x)$?

Perhaps $Q(y=k+1) = 1$!!

Requires addtl identifiability conditions.

Focus on base case: $k=1$



don't need to know which x!

Assumption: Separable Sub-Support

$\exists x \in \mathcal{X}$ $P(x|y=+) > 0$ & $P(x|Y=-) = 0$

Steps:

1. Mixture proportion estimation of $Q(Y=+) \triangleq \alpha$
2. Learn a PvN classifier

Previous approaches

Elkan / Noto provide estimators
key idea is to relate PvN quantities to PvU "non-traditional classifiers"

No theory / analysis.

Other recent approaches act in classifier output space but need domain heuristic: no conditions for consistency.

Sugiyama introduced unbiased loss for learning the classifier:

$$\hat{R} = \alpha R_P^+ + R_U^- - \alpha R_P^-$$

Identification: $\alpha^* = \max_\alpha \alpha$

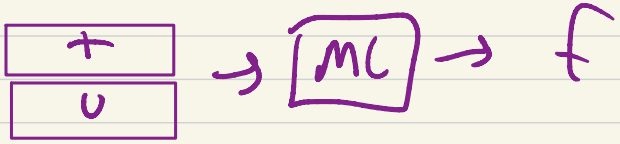
$$\text{s.t. } \forall x \quad \alpha P(X|Y=+) \leq Q(X)$$

Intuition

higher α: impossible
lower α: violates irreducibility / separable subsystem

α Estimation Strategy:

Domain discrimination

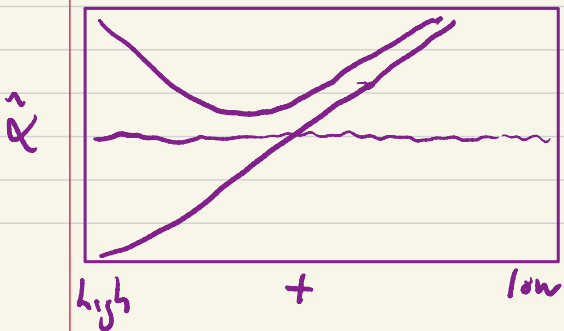


Insight: α determined by the "most positive" points (those in separable sub-support)

- Use f to map all points to scalar $z_i = f(x_i)$
  - Set threshold t
  - Using holdout + & U data estimate

$$\hat{\alpha} = \frac{P_u(z > t)}{P_+(z > t)} \geq \alpha$$

Bias-Variance tradeoff



Choose t that gives lowest upper bound on α

Bound

$$\|\hat{\alpha} - \alpha\| \leq \frac{c_1}{q_p(c^*)}\left(\sqrt{\frac{\log(t/\delta)}{n_u}} + \sqrt{\frac{\text{"1"}}{n_p}}\right)$$

Obtaining the optimal classifier:

A uPU loss

B NNuPu loss

Proposal CUIR (for separable data)

Sort data, discard highest loss α fraction among unlabeled at each iteration.

Consistent when data separable avoids overfitting mislabeled data

Need more innovation for non-sep data

Further empirical improvements w iterated $\hat{\alpha}$, $\hat{f}$ estimation (TED)^N.

Back to Open Set Label Shift (Garg, 2022)

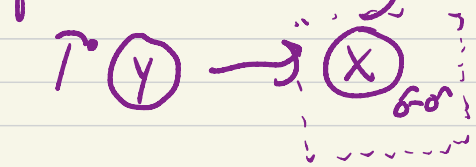
- Reduction to k PU problems
- Require each prev seen class has a separable subsupport

  OR negatives entirely separated.
- Tricks for coping with overestimation of α

Algorithm:

1. Train source classifier f
2. Estimate each $\hat{\alpha}_i$ via PU via BBE, using $f(y=i|x)$ as score
3. Normalize to sum to 1 (correcting abit for overestimation)
4. Reweight source data via $\hat{\alpha}'$
5. Estimate frac belonging to novel class through one final application of BBE, treating the reweighted src as one giant + class.

Unsupervised Learning under Latent Label Shift



Data observed across many domains
(# domains ≥ # labels)

Given this structure (shift due to label shift),
when are underlying labels recoverable?
(up to permutation)

Motivation:
- Butterflies, caterpillars

Principle:
- items that shift together group together

Identification
- Finite data, isomorphism to topic modeling, apply Arora
  - ↑ separable sub-support = anchor
- Continuous data + black box
  - ↑ product $\Phi(d|x)$