Nuclear Discrepancy for Single-Shot Batch Active Learning

Tom J Viering, Jesse H Krijthe, Marco Loog

ECML 2019 Optimization and Learning Theory

Code online:
https://github.com/tomviering/NuclearDiscrepancy

UNIVERSITY OF COPENHAGEN

TU Delft

Outline

- Motivation for AL, setting
- Domain adaptation bounds for AL
  - MMD, Discrepancy, Nuclear Discrepancy
- Theoretical results
- Experiments

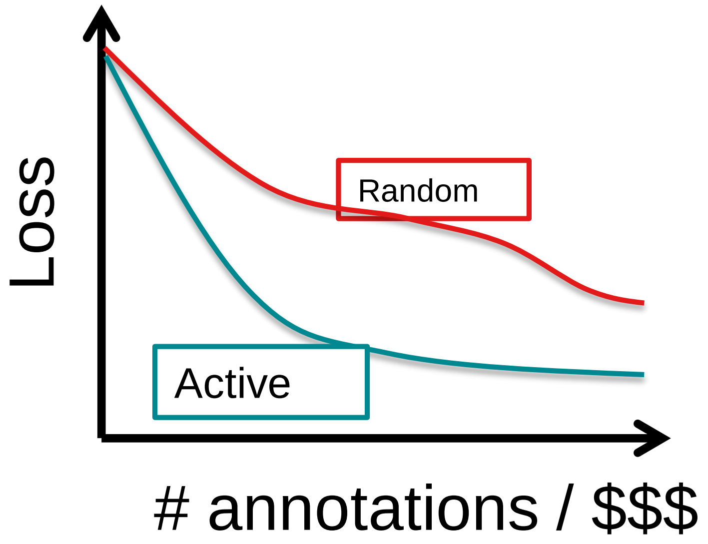
Setting

- A lot of unlabeled data (unannotated recordings)
- Few labeled data (annotated recordings)

- Labeling: expensive, time consuming, difficult

Active Learning

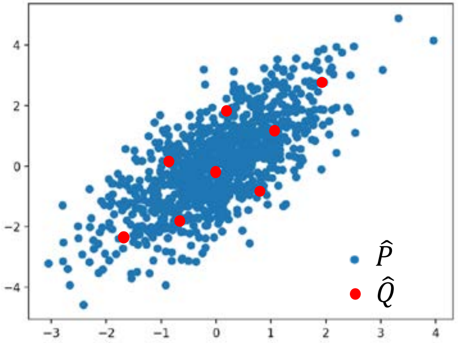
- Algorithm (active learner) selects what data to annotate
- Model can learn faster from ‘smart’ selection of data



Single-shot Batch AL procedure

Input: label budget n, unlabeled data $\hat{P}$

1. Active learner (AL) chooses n samples $\hat{Q} \in \hat{P}$ such that $div(\hat{Q}, \hat{P})$ minimized
2. Request labels for $\hat{Q}$
3. Train KRR model on $\hat{Q}$
4. Evaluate on unseen test set



- Note: AL never sees labels.
  - Selects 'representative' samples

Generalization bounds

- Squared loss L, binary classification
- Model: h, kernel ridge regression model, $h \in H$ (RKHS)
- Unknown:
  - Distribution P over input space x,
  - Deterministic labeling function f, $y = f(x)$

Domain Adaptation Bounds for AL

Empirical Sample	Active Learning	Domain Adaptation
$\hat{Q}$	Labelled	Source
$\hat{P}$	Unlabelled	Target

- Domain adaptation bounds:
  - $L_P(h,f) \leq L_{\hat{Q}}(h,f) + div(\hat{Q},\hat{P}) + C + \eta$

Model misspecification (ignore)

Loss on distribution (risk)

Minimized by training (empirical risk)

Divergence measure Minimized by Active Learning

Complexity term (ignore)

What Divergence to use?

- From Domain Adaptation:
  - MMD [Huang 2007], also used for AL by Chattopadhyay et. al. (2012)
  - Discrepancy [Cortes, Mohri 2011]

- Research questions:
  - How do the MMD and Disc. compare?
  - Why one or the other better for AL?

TUDelft

Recap: how to get to the Discrepancy?

- Quantity to bound: $|L_{\hat{P}}(h, f) - L_{\hat{Q}}(h, f)|$
- Assume $f \in H$ (realizeable)
- Consider worst case over h, f:
- $\max_{f,h \in H} |L_{\hat{P}}(h, f) - L_{\hat{Q}}(h, f)| = disc(\hat{Q}, \hat{P})$
- Depends on model class, loss function

TUDelft

Compare to MMD

- Quantity to bound: $|L_{\hat{P}}(h, f) - L_{\hat{Q}}(h, f)|$
- $MMD(\hat{P}, \hat{Q}) = \max_{l \in H'} \frac{1}{n_{\hat{P}}} \sum_{x \in \hat{P}} l(x) - \frac{1}{n_{\hat{Q}}} \sum_{x \in \hat{Q}} l(x)$
  - H' is usually heuristically chosen as RKHS of a RBF kernel

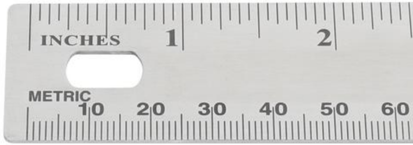
- Idea: use $l(x) \approx \left(h(x) - f(x)\right)^2$ to relate both
- This analysis suggests how to choose H':
  - MMD and Disc now compareable!

TUDelft

Compare Disc and MMD

- Assume worst case for f, h
- $disc(\hat{Q}, \hat{P}) \leq MMD(\hat{Q}, \hat{P})$
- Disc provides tighter bound!
  - Disc provides better AL?
  - Empirically: MMD beats Disc. Why?

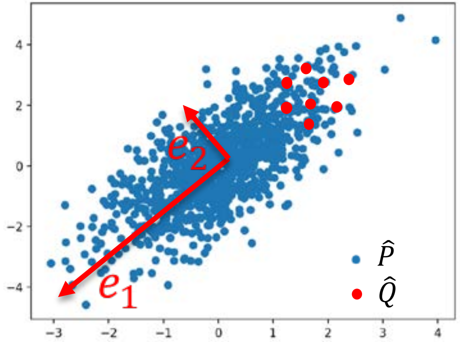
MMD (coarse)



Discrepancy (fine)

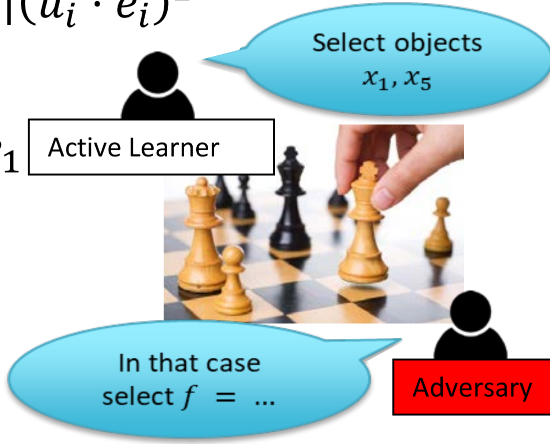
Disc is too pessimistic

- $u = h - f$
- $M = \frac{1}{n_{\hat{P}}} X_{\hat{P}}^T X_{\hat{P}} - \frac{1}{n_{\hat{Q}}} X_{\hat{Q}}^T X_{\hat{Q}}$
- Let $e_1, e_2, \ldots, e_d$ be orthonormal eigenvectors
- Eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_d|$



- Then $|L_{\hat{P}}(h, f) - L_{\hat{Q}}(h, f)| = |u^T M u| \leq \sum_i^d |\lambda_i| (u_i \cdot e_i)^2$

- Disc assumes worst case for f, h, then $u \propto e_1$
  - Assumes u in very specific direction
  - $Disc(\hat{P}, \hat{Q}) \propto |\lambda_1|$
  - Our choice $\hat{Q}$ determines f, very pessimistic
  - Disc. doesn't 'spread' well



TU Delft

Nuclear Discrepancy

- Assume $u \sim p(u)$ and create probabilistic bound (holds in expectation)
  - $p(u)$ should be symmetric
  - Should be independent of our choice $\hat{Q}$

- Choose $\mathrm{p}(u)$ uniform on sphere centered at origin **[optimistic case]**
  - Optimal strategy: minimize Nuclear Discrepancy (proposed)
  - $ND\left(\hat{P}, \hat{Q}\right) = \sum_i^d |\lambda_i|$ (all directions are equally important)
  - In this case, $ND\left(\hat{P}, \hat{Q}\right) \leq MMD\left(\hat{P}, \hat{Q}\right) \leq Disc(\hat{P}, \hat{Q})$
  - Our bound is tightest under this assumption

TU Delft

Experimental setup

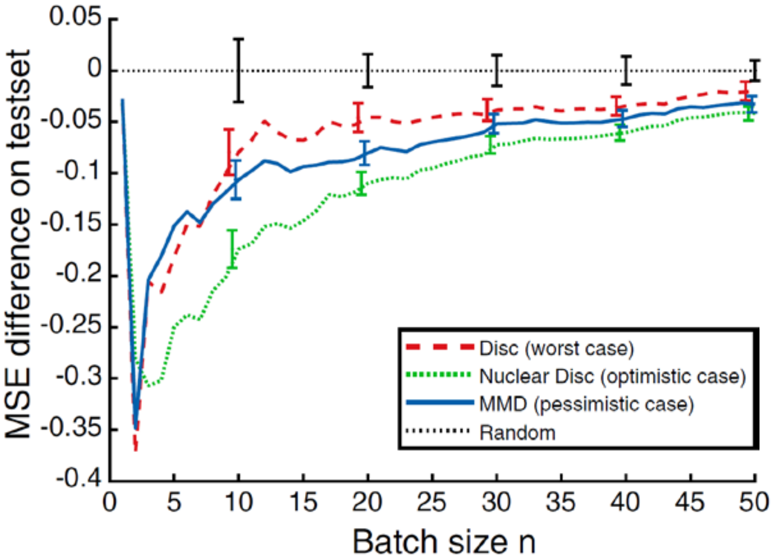
- Preprocess to remove model misspecification
  - Train KRR on whole dataset, use predictions as new targets
  - Assumption $\eta = 0$ satisfied. Bounds compareable.
  - Good hyperparameters make sure this is a reasonable approximation of the original binary label.
- Optimize $div(\hat{Q}, \hat{P})$ greedily
  - Discrepancy, MMD, Nuclear Discrepancy

TU Delft

Experimental setup

- Budget = 1,2,3,…,50.
- Repeat 100 times
  - New trn/tst splits
- Evaluate on 15 datasets
- Performance in MSE
- Area Under Learning Curve
  - summarizes performance for multiple budgets (standard in AL)
- Significance test using paired t-test (p = 0.05)

Learning Curve
MNIST 5vs8



TU Delft

Results

Table 2 Area Under the mean squared error Learning Curve (AULC) for the strategies in the realizable setting, averaged over 100 runs

Dataset	Random	Discrepancy	MMD	Nuclear Discrepancy
vehicles	11.1 (2.2)	**8.0 (1.0)**	**7.9 (0.9)**	**7.9 (0.9)**
heart	3.5 (0.8)	2.3 (0.3)	2.2 (0.3)	**2.1 (0.3)**
sonar	13.9 (1.7)	12.5 (1.2)	11.9 (1.1)	**11.3 (1.2)**
thyroid	6.8 (1.5)	5.2 (0.9)	**5.1 (0.9)**	**5.0 (1.0)**
ringnorm	13.2 (1.2)	12.7 (0.8)	10.0 (0.3)	**9.4 (0.3)**
ionosphere	7.0 (1.3)	5.6 (0.8)	5.0 (0.8)	**4.6 (0.6)**
diabetes	1.7 (0.4)	**1.2 (0.1)**	**1.2 (0.1)**	**1.2 (0.1)**
twonorm	6.4 (1.2)	4.1 (0.4)	3.7 (0.4)	**3.3 (0.3)**
banana	7.5 (0.9)	5.0 (0.4)	**4.8 (0.3)**	**4.8 (0.3)**
german	1.4 (0.3)	1.2 (0.1)	1.1 (0.1)	**1.0 (0.1)**
splice	10.8 (1.3)	9.9 (0.8)	9.9 (0.9)	**9.0 (0.9)**
breast	3.4 (0.9)	2.1 (0.2)	2.1 (0.2)	**2.0 (0.2)**
mnist 3vs5	29.5 (4.3)	26.9 (2.3)	25.0 (2.1)	**23.8 (1.7)**
mnist 7vs9	13.2 (2.5)	10.9 (1.4)	10.0 (1.0)	**8.9 (0.7)**
mnist 5vs8	30.1 (3.4)	26.9 (2.7)	26.1 (2.3)	**24.5 (2.1)**

Bold indicates the best result, or results that are not significantly worse than the best result, according to a paired t-test ($p = 0.05$). Parenthesis indicate standard deviation

TU Delft

Results

Assumption Bound	Worst-case	Optimistic Case $p(u)$	Performance
Disc	Tightest	Loosest	Worst
MMD	Medium	Medium	Medium
ND **(ours)**	Loosest	Tightest	Best

TUDelft

Discussion / Future work

- With no preprocessing:
  - Bounds not directly comparable ($\eta_{MMD} \neq \eta_{Disc}$)
  - Similar trend observed
  - Needs further investigation

- Now MSE, what about accuracy? Multiple rounds?

- Our results support the ideas of
  - Germain et. al. (2013): probabilistic bounds for DA
  - Cortes et. al. (2019): more refined worst-case analysis for Discrepancy for DA

Conclusion

- Tighter bounds ≠ improved performance
- Assumptions of bounds: at least as important!

- Bonus theoretical results for MMD:
  - Interpretation as probabilistic generalization bound
  - Squared kernel $K_{MMD}(x_i, x_j) = K_{model}(x_i, x_j)^2$ is a natural choice for H' in case of regression

TU Delft

Thanks!

Tom J Viering, Jesse H Krijthe, Marco Loog

TU Delft

Nuclear Discrepancy for Single-Shot Batch Active Learning
Code online: https://github.com/tomviering/NuclearDiscrepancy