November 6 Updates

Simon Dovan Nguyen

November 7, 2024

Contents

1 Meeting Updates

1.1 Logistics:

1. **Metrics:**
   (a) AUC vs. RMSE:
   - RMSE:
     - A lot of other people use RMSE
     - Wu, Lin, and Huang (2018) uses RMSE and then uses AUC := $\frac{RMSE}{CC}$
   - AUC (Surzhikova and Proppe, 2024)
     - AUC := $\frac{R^2}{RMSE}$ (I think, not clear from paper)
     - States RMSE values at a specific training set size is not a suitable performance comparison between different sample selection methods, as these values are fluctuating on a smaller scale.
     - Instead, AUC shows the overall progression of that method's performance
     - Wu, Lin, and Huang (2018) also uses a similar AUC := $\frac{RMSE}{CorrelationCoefficient}$

   (b) Error Calculation:
   - The standard (but **very expensive**) way to calculate the recommendation metric is to calculate **expected** error reduction for each and every candidate observation (Roy and McCallum, 2001).
   - This is very expensive.
   - There is a precedent in Nguyen and Smeulders (2004) that suggests to query candidates with largest contribution to **current** error.
   - We are currently doing what Nguyen and Smeulders (2004) are doing.

   (c) Ambiguity Recommendation Metric
   - Do we want to measure **variance** or current/expected **error**?
   - We are currently meaasuring **current** error:

$$\sum_{m\in\mathcal{M}}\sum_{k}\left[(\hat{y}_k(x) - y(x)\right]^2$$

   - And then matching based on covariates
   - Variance:
     i. Burbridge, Rolwand, and King 2007 suggest this metric (basically the variance)

$$\sum_{k}\left[(\hat{y}_k(x) - \bar{y}(x)\right]^2$$

     ii. I forgot is this on the training or the candidate data). If it's on the training, then we calculate this metric then match. If it's on the candidate, we just suggest the one with the largest metric.

     iii. We can adapt this to include Rashomon-weighting easily:

$$\sum_{m\in\mathcal{M}}\sum_{k}\left[(\hat{y}_{k,m}(x) - \bar{y}_m(x)\right]^2$$

   - Clarify objective – ambiguity or error?

2. **Distance:** Euclidean distance is commonly used (Surzhikova and Proppe, 2024).

1.2 New Algorithms

1. GSy: Greedy Sampling on the Output Wu, Lin, and Huang (2018):
   (a) Algorithm Steps:
      i. Construct initial training set.
      ii. Build initial predictive model $F(X)$.
      iii. Apply model on training set $F(X_{\text{tr}})$.
      iv. Apply model on candidate set $F(X_{\text{cand}})$.
      v. Calculate distances between all predicted training and candidate observations:
         A. $d^y_{nm} = ||F(X_{\text{cand}_m}) - y_{\text{true, training}}||\ \forall m, n.$
      vi. Compute the shortest distance from $F(X_{\text{cand}_m})$ to $y_{\text{true, training}}$:
         A. $d^y_n = \min_m d^y_{nm}\ \forall n = k+1, \ldots, N.$
         B. Identify the nearest neighbor candidate for each training observation.
      vii. Select candidate with the largest d^y_n:
         A. Label candidate with the furthest nearest neighbor.
2. Difference in Approach:
   (a) Current approach calculates the nearest neighbor only for the most uncertain or incorrect observation.

(b) GSy computes nearest neighbors for all observations in output space.

1.3 Data Generating Process (DGP):

1. Continual new data generation vs. single dataset with random start.
   (a) Wait think about this one a bit more
   (b) That is, the current process is
      i. Generate new data set
      ii. Random initial training data
      iii. Run 3 methods with the same data set and same initial training data set
   (c) Two sources of randomness: DGP and initial training set.
2. Data complexity:
   (a) Quotes from Castro, Willet, Nowak (2005):
      - Function complexity spatially homogenous random sampling better (near-minimax optimal)
      - Simple uniform sampling scheme is naturally matched to the homogeneous "distribution"
      - Functions with spatially non-uniform complexity highly concentrated in small subsets of the domain → active learning better
   (b) Intuition (from Kentaro):
      - Imagine you're trying to estimate a line. It doesn't matter where you collect your data - the slope is the same everywhere.
      - However, if you're trying to plot a more complicated function (eg. parabola/quadratic/etc.), where you collect your data matters
      - This is because the slope changes at different points in the estimand function
      - That's why Holder-smooth functions (spatially homogeneous complexity) are better with random sampling: Given $C \in \mathbb{R}^{\geq 0}$ and $\alpha \in \mathbb{R}^{>0}$,

$$\begin{aligned} |f(x) - f(y)| &\leq C||x-y||^{\alpha} && \forall x, y \in dom(f) \\ \frac{|f(x) - f(y)|}{||x-y||^{\alpha}} &\leq C && \forall x, y \in dom(f) \\ \text{"slope"} &\leq C \end{aligned}$$

      - That is, a Holder continuous function is limited in how fast it can change
   (c) Consider the DGP model by Burbridge, Rolwand, and King 2007, (page 213). Note the data is not discretized (how will this affect things?).
3. Application: Datasets (UCI):
   (a) Concrete
   (b) Housing
   (c) Wine
   (d) Yacht
   (e) Diamonds
   (f) NCI ALMANAC
   (g) More on this later.

1.4 Other thoughts

1. *Noise:*
   - Active Learning (AL) outperforms Random when the model is correctly specified and there is no output noise Burbridge, Rolwand, and King 2007.
   - Rashomon Ratios are larger (and potentially more informative) when there is noise (Semenova et al., 2023).
   - Question: Are AL and Rashomon opposed, or could Rashomon aid AL?
2. Statistical Testing: Wilcoxon Signed Rank Test then Dunn's procedure for multiple hypothesis correction Wu, Lin, and Huang (2018).
3. Change wording/notation: Delta metric as "ambiguity metric".

2 Post-Meeting to-do

1. **Algorithm** (Wu, Lin, Huang (2018)):
   - Implement GSx
   - Implement GSy
   - Implement iGS
2. **Data Generating Process**:
   - No need for useless covariate.
   - Implement DGP from Castro, Willet, Nowak (2008), Figure 2.
   - Implement DGP from Burbridge, Rolwand, and King 2007, (page 213).
3. **Ambiguity Recommendation Metric**
   - Implement:
     (a) Current Error Ambiguity Metric (CEAM)
       - Calculate CEAM for the test set.
       - Match test and candidate set based on this metric.

       $$\frac{1}{m}\sum_{m\in\mathcal{M}}\left(\sum_k\left[\hat{y}_k(x)-y(x)\right]^2\right)\cdot P(m\in\mathcal{M})$$

     (b) Variance Ambiguity Metric (VAM) on the **test set**
       - Calculate VAM for the test set.
       - Match test and candidate set based on this metric.

       $$\frac{1}{m}\sum_{m\in\mathcal{M}}\left(\sum_k\left[\hat{y}_k(x)-\overline{y}(x)\right]^2\right)\cdot P(m\in\mathcal{M})$$

     (c) Variance Ambiguity Metric (VAM) on the **candidate set**
       - Calculate VAM for the candidate set.
       - Query candidate observations directly with this.

       $$\frac{1}{m}\sum_{m\in\mathcal{M}}\left(\sum_k\left[\hat{y}_k(x)-\overline{y}(x)\right]^2\right)\cdot P(m\in\mathcal{M})$$

   - Rashomon Averaging vs. Worst Case
     - Currently, we are averaging across the Rashomon Set.
     - Consider if instead of averaging, we took the worst model in the Rashomon Set.
     - Do we still need the $P(m\in\mathcal{M})$?
     - Compare to averaging.
     - For instance, take CEAM on the test set:

     $$\max_{m\in\mathcal{M}}\left(\sum_k\left[\hat{y}_k(x)-y_k(x)\right]^2\right)\cdot P(m\in\mathcal{M})$$
4. **Other**: Think about NAs recommendation more.