Formal Hypothesis Test for Distribution Shift via Random Forest

Heqiao Ruan, Lucas Mentch

Department of Statistics, University of Pittsburgh

Overview

- For an already deployed model, we will not know how the next batch of data differs from the existing data, so understand the **distribution shift** would be one of the most important topics for **machine learning practitioners**.
- We propose a computationally efficient permutation test **PermOOB** to detect **distribution shift** for both regression and binary classification tasks.
- The proposed permutation test for distribution shift can **extend to arbitrary machine learning model** with a concretely defined validation set.

Motivation

- Data coming in batches, $\mathbf{D}_{exist} = (\mathbf{X}_{exist}, Y_{exist})$ and $\mathbf{D}_{new} = (\mathbf{X}_{new}, Y_{new})$. Types of shifts can vary:
  - Covariate Shift: $P(\mathbf{X}_{exist}) \neq P(\mathbf{X}_{new})$ where $P(Y_{exist}|\mathbf{X}_{exist}) = P(Y_{new}|\mathbf{X}_{new})$, only covariates distribution change where the conditional dependence remain the same.
  - Concept Drift: $P(Y_{exist}|\mathbf{X}_{exist}) \neq P(Y_{new}|\mathbf{X}_{new})$ where $P(\mathbf{X}_{exist}) = P(\mathbf{X}_{new})$, only conditional dependence change where the covariate distribution remain the same
  - Distribution Shift: At least one of $P(Y|\mathbf{X})$ and $P(\mathbf{X})$ changes or both changes across existing batch of data and new batch of data.
- In practice, the distribution shift is prevalent - we will not know exactly which of the covariate shift and concept drift exist or both exist.
- **Existing methods [2, 3, 4, 5] only focus on certain type of shift. We aim to develop hypothesis testing procedure efficient in most of scenarios.**

Hypothesis Testing Procedure

- Random forest is among one of the most popular machine learning algorithms. Out-of-bag(OOB) error provides a readily available validation set.
- **The intuition of the test**: We fit a model on the current batch of data and perform prediction on the next batch of data. We aim to compare the OOB error(MSE) and the Prediction Error(MSE) on observation level.

Algorithm 1 Permutation Test for Distribution Shift(**PermOOB**)

BEGIN We start from the existing batch of data $D_{exist} = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_{n_{exist}}, Y_{n_{exist}})\}$ and new batch of data $D_{new} = \{(\mathbf{X}_{n_{exist}+1}, Y_{n_{exist}+1}), ..., (\mathbf{X}_{n_{exist}+n_{new}}, Y_{n_{exist}+n_{new}})\}$ each $\mathbf{X}_j = (x_1, ..., x_p)$ consists of p covariates $\forall j = 1, ., n_{exist} + n_{new}$. We conduct the hypothesis testing procedure $H_0 : E[MSE_{OOB}(\mathbf{X}; Y)] = E[MSE_{new}(\mathbf{X}; Y)]$ *vs.* $H_a : E[MSE_{OOB}(\mathbf{X}; Y)] < E[MSE_{new}(\mathbf{X}; Y)]$.

S1 Standardize the covariates both in the existing batch of data $\mathbf{X}_{exist}$ and in the new batch data $\mathbf{X}_{new}$. Then train a random forest model on D_{exist} with T trees denoted as $h_1, ..., h_T$.

S2-1 For $i = 1, 2, ..., n_{exist}$, get the average prediction error on the OOB sample for that observation, denote as $MSE_{i,oob}$, calculated as $MSE_{i,oob} = (Y_i - \frac{1}{n_{oob,i}} \sum_{j \in oob} h_j(\mathbf{X}_i))^2$ with T trees $h_1, ..., h_T$.

S3 For $i = 1, 2, ..., n_{new}$, get the average prediction error on the predicted sample, denote as $MSE_{i,new}$, calculated as $MSE_{i,new} = (Y_i - \frac{1}{T} \sum_{j=1}^{T} h_j(\mathbf{X}_i))^2$.

S4 Conduct a one-sided permutation test based on $L1 = \{MSE_{i,oob}, i = 1, 2, ..., n_{exist}\}$ and $L2 = \{MSE_{i,new}, i = 1, 2, 3, ..., n_{new}\}$. Denote the original difference as $d_0 = mean(L_1) - mean(L_2)$ Then a permutation test is conducted with B(set as 5000 by default) times, denote the difference as $d_i = L_{\pi_i,2} - L_{\pi_i,1}$, record the empirical cdf of $\{d_1, ..., d_B\}$ as L_π. Calculate the p-value as p-value, $p = 1 - \hat{L}_\pi(d_0)$.

S5 Reject the H_0 if $p < \alpha(\alpha$ is the significance level) and otherwise not reject.

END

Simulation Setting

Model Name	Data Generating Process
Mars Model	$Y = 50 * sin(\pi X_1 X_2) + 5 * (X_3 - 0.05)^2 + 5 * X_4 + 5 * X_5$, $X_i \sim N(0, 1)$
Linear Model	$Y = X\beta + \epsilon$, $X \sim N(0, \Sigma_{(i,j)} = \rho^{\lvert i-j \rvert})$, $SNR = \frac{\beta^T \Sigma \beta}{Var(\epsilon)}$
Concept Drift1	$Y = (X_5 + X_6 + X_7 + X_8) * (1 - T * \lambda) + X_1 + X_2 + X_3 + X_4$, T is indicator of data batch
Concept Drift2	$Y = (X_7 + X_8) * (1 + 5 * T * \lambda) + X_1 + ... + X_6$, T is indicator of data batch
Mean Shift Mean Shift	Existing Batch Data $Mean(X_1, X_2, X_3, X_4, X_5) \sim (1 + \delta, 1 + 2\delta, 1 + 3\delta, 1 + 4\delta, 1 + 5\delta)$ New Batch Data $Mean(X_1, X_2, X_3, X_4, X_5) \sim (1 + 5\delta, 1 + 4\delta, 1 + 3\delta, 1 + 2\delta, 1 + \delta)$

Table 1. List of Simulation settings, each of the setting will impose varying number of random noise features, noise levels $var(\epsilon) = \sigma^2$ and dependence level among features ρ.

Simulation Results

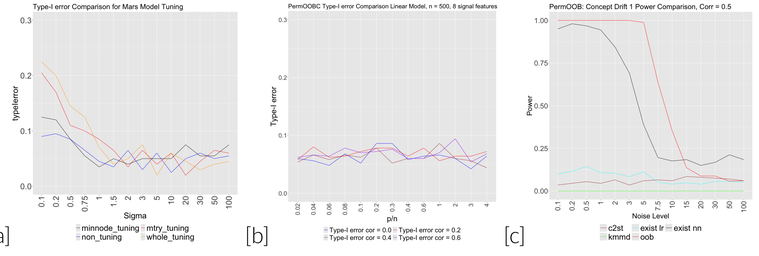

Figure 1. (a) Type-I error for tuning comparison: Regression Setting (b) Type-I error Comparison: Binary Classification Setting (c) Power Comparison for Concept Drift

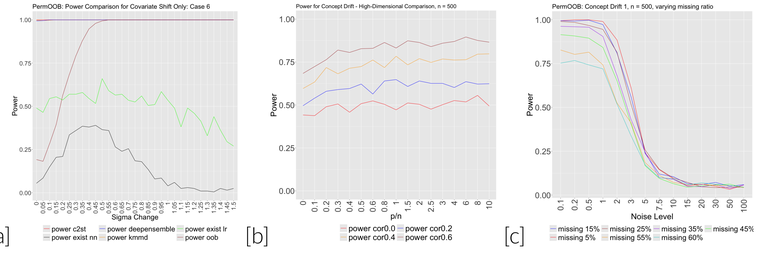

Figure 2. (a) Power Comparison for detecting covariate shift (b) Power for Linear Model in high-dimensional data (c) Power comparison with missing value in New batch of data

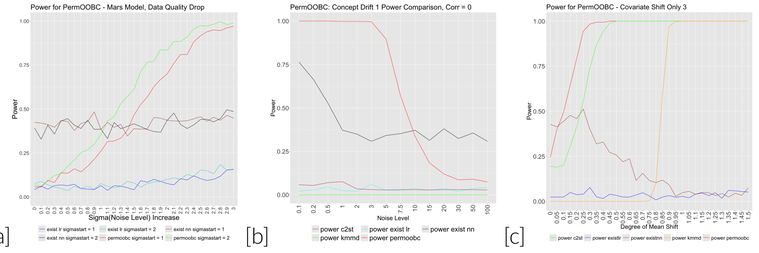

Figure 3. (a) Power Comparison for Data Quality Drop (b) Power Comparison for Concept Drift in Binary Response (c) Power Comparison for detecting Covariate Shift in Binary Response

Real Data Application

We evaluate the performance of permutation tests on a variety of UCI datasets where we use features to split the data into training set and testing set. To select the features to split onto, we regress response on all predictors and select those features with the lowest p-value. Then split based on the median value of that feature. **By construction, distribution shift exists.**

Dataset - Feature	PermOOB	Existing-NN	Existing-LL	CFPerm
AToxic - V1	**1.000**	0.370	0.02	0.125
AToxic - V2	**1.000**	0.265	0.035	0.310
AToxic - V4	**1.000**	0.468	0.026	0.595
Autoprice - V7	**1.000**	0.445	0.000	0.175
Autoprice - V8	**1.000**	0.500	0.063	0.220
Autoprice - V10	**1.000**	0.410	0.038	0.295
Boston - rm	**1.000**	0.483	0.115	0.565
Boston - nox	**1.000**	0.469	0.027	0.110
Boston - crim	**1.000**	0.37	0.020	0.275
StudentPerf - higher	**1.000**	0.493	**1.000**	0.475
StudentPerf - famsup	**1.000**	0.225	0.007	0.110
ethanol - X	**0.622**	0.580	0.440	0.555
airfoil - X5	**1.000**	**1.000**	0.868	0.985
airfoil - X4	**1.000**	0.996	0.764	0.885
airfoil - X2	0.066	**0.068**	0.046	0.060

Table 2. Comparison for power of detecting the covariate shift on more **UCI datasets** - made to split the dataset based on the features that contributes most significantly to the response

PermOOB achieve higher power than existing methods in most cases.

Theoretical Guarantee

Asymptotic Validity for Permutation Test

Theorem 1: Given existing batch of data and new batch of data

$$D_{exist} = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_{n_{exist}}, Y_{n_{exist}})\}\ , D_{new} = \{(\mathbf{X}_{n_{exist}+1}, Y_{n_{exist}+1}), .., (\mathbf{X}_{n_{exist}+n_{new}}, Y_{n_{exist}+n_{new}})\}$$

The random forest fitted on the existing batch of data with B decision trees predicted value on X is denoted as $U_B(X)$, the out-of-bag MSE for the existing batch of data denoted as $MSE_{oob,1}, ..., MSE_{oob,n_{exist}}$ and the predicted MSE for the new batch of data denoted as $MSE_{new,1}, ..., MSE_{new,n_{new}}$. Suppose the following assumptions are satisfied (1) Trees in the random forest are asymptotically exchangeable (2) Prediction value for out-of-bag sample $U_{oob}(\mathbf{X}_{exist,i}), U_{oob}(\mathbf{X}_{exist,j})$ are asymptotically independent for $i \neq j \in \{1, .., n_{exist}\}$ and also for the predicted value in the testing data $U_B(\mathbf{X}_{new,i}), U_B(\mathbf{X}_{new,j})$ are also asymptotically independent for $i \neq j \in \{1, ..., n_{new}\}$. (3) The variance for the out-of-bag MSE and predicted MSE are both consistent where $Var(MSE_{oob}) \xrightarrow{P} \sigma^2_{oob}$ and $Var(MSE_{new}) \xrightarrow{P} \sigma^2_{new}$. Then consider a test of null hypothesis

$$H_0 : E[MSE_{oob}(\mathbf{X}; Y)|\mathbf{X}, Y] = E[MSE_{new}(\mathbf{X}; Y)|\mathbf{X}, Y]$$

vs.

$$H_a : E[MSE_{oob}(\mathbf{X}; Y)|\mathbf{X}, Y] < E[MSE_{new}(\mathbf{X}; Y)|\mathbf{X}, Y]$$

using the statistic $T_{n_{exist},n_{new}} = \frac{1}{n_{new}} \sum_{i=1}^{n_{new}} MSE_{new,i} - \frac{1}{n_{exist}} \sum_{i=1}^{n_{exist}} MSE_{oob,i}$. Then under null hypothesis H_0, based on Theorem 2.2 in [1], the permutation distribution of $\sqrt{n_{exist} + n_{new}} T_{n_{exist},n_{new}}$ converges to a normal distribution with mean 0 and variance

$$V_{n_{exist},n_{new}} = \frac{n_{exist} + n_{new}}{n_{exist}} \sigma^2_{oob} + \frac{n_{exist} + n_{new}}{n_{new}} \sigma^2_{new}$$

as $n \to \infty$ which is also the variance of the unconditional distribution of $\sqrt{n_{exist} + n_{new}} T_{n_{exist},n_{new}}$ Thus the permutation test attains the asymptotic type-I error control.

For Binary Response, Brier Score $BS_i = (Y_i - \hat{RF}(\mathbf{X}_i))^2$, $Y_i \in \{0, 1\}$. Then the test become

$$H_0 : E[BS_{OOB}(\mathbf{X}; Y)|\mathbf{X}, Y] = E[BS_{new}(\mathbf{X}; Y)] \; vs. \; H_a : E[BS_{OOB}(\mathbf{X}; Y)|\mathbf{X}; Y] < E[BS_{new}(\mathbf{X}; Y)|\mathbf{X}; Y]$$

Then the Lyapunov CLT is established followed by the asymptotic type-I error control.

Conclusion & Practical Considerations

Conclusion

- The test is easy to implement and fast in computation, its performance is among one of the best in almost all of the circumstances. The test is robust under high-dimensional data, moderate missing data and heavy tail noise.
- We demonstrate that the test can be extended comfortably extended to binary classification problem with solid theoretical guarantee.
- The good fit of the random forest model will ensure the desired power for the test.

Practical Considerations

- Practically speaking, the proposed permutation test can serve as a pre-screening procedure for **any already deployed Machine Learning Models** - If **permOOB** not reject H_0 then we will make a concrete argument that there's no distribution shift among batches of data.
- The test give implications about extrapolation: as long as there exist subset of features $\mathbf{X}_S \in \mathbf{X}$ where the conditional dependence $Y|\mathbf{X}_S$ changes significantly along those directions -$P(Y|\mathbf{X}_S)$ changes indicate $P(Y|\mathbf{X})$ changes. There will be distribution shift.
- Extend beyond the scope of random forest - as long as one have a concrete definition of the validation set, the permutation test can be extended to **any machine learning model.**
- The test can be comfortably extended out of the scope of tabular data - large language model evaluation and image classifier. The evaluation metrics would be adjusted.

References

[1] Tim Coleman, Wei Peng, and Lucas Mentch.
Scalable and efficient hypothesis testing with random forests.
Journal of Machine Learning Research, 23(170):1–35, 2022.

[2] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola.
A kernel two-sample test.
The Journal of Machine Learning Research, 13(1):723–773, 2012.

[3] Xiaoyu Hu and Jing Lei.
A two-sample conditional distribution test using conformal prediction and weighted rank sum.
Journal of the American Statistical Association, 119(546):1136–1154, 2024.

[4] David Lopez-Paz and Maxime Oquab.
Revisiting classifier two-sample tests.
arXiv preprint arXiv:1610.06545, 2016.

[5] Jian Yan and Xianyang Zhang.
A nonparametric two-sample conditional distribution test.
arXiv preprint arXiv:2210.08149, 2022.

Presenting Author: Heqiao Ruan Email: [1] :her76@pitt.edu, [2] :lkm31@pitt.edu