Holdout Randomization Tests

Black Box Variable Selection

Haoran Zhang

Columbia University

Sep. 26th 2018

Outline

Introduction

- Black box models
  - deep neural networks
  - random forests
  - ...
- Fields such as biology and chemistry
- Variable selection

$$H_0 : X_j \perp\!\!\!\perp Y | X_{-j} \,. \tag{1}$$

Examples

What gene expression and mutation features affect cancer cell line sensitivity to the drug PLX4720?
What molecular features controls the perceived fragrant intensity of molecules?

Reference

S. Basu, K. Kumbier, J. B. Brown, and B. Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, page 201711236, 2018.

T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test. *arXiv preprint arXiv:1807.05405*, 2018.

E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series*, 2018.

J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *International Conference on Machine Learning*, 2018.

F. Liang, Q. Li, and L. Zhou. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, pages 1–18, 2018.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

R. Sen, K. Shanmugam, H. Asnani, A. Rahimzamani, and S. Kannan. Mimic and classify: A meta-algorithm for conditional independence testing. *arXiv preprint arXiv:1806.09708*, 2018.

W. Tansey, V. Veitch, **H. Zhang**, R. Rabadan, and D. Blei. Holdout randomization tests: Principled and easy black box variable selection. *preprint*, 2018.

State of the field

- model-specific
  - Iterative Random Forests (Basu et al., 2018)
  - Bayesian neural networks (Liang et al., 2018)
  - ...
- focus on interpretation of the model (make simplifying independence assumptions between covariates)
  - SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017)
  - Learning to Explain (L2X) (Chen et al., 2018)
  - ...
- similar approaches
  - Mimic and Classify (Sen et al., 2018)
  - Knockoffs (Candes et al., 2018)
  - Conditional Permutation Tests (Berrett et al., 2018)

Setup and Notation

- Dataset $\mathcal{D}^* = \{(X_i, Y_i)\}_{i=1}^{n^*}$ of n^* samples drawn i.i.d. from $P(X, Y)$
- Predictive model $F_{\hat{\theta}}(X_i) \rightarrow \hat{Y}_i$
- The data is split into train and test sets, D and D' of n and n' samples, respectively
- D is used to fit $\hat{\theta}$ by minimizing loss function $\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(X_i, Y_i, F_\theta)$
- D' is held out for evaluating the generalization error $\mathcal{T}(\hat{\theta}) = \sum_{i=1}^{n'} g(X_i, Y_i, F_{\hat{\theta}})$

Holdout Randomization Test

1: **procedure** HRT(training data $\mathcal{D}$, test data $\mathcal{D}'$, model F, training loss $\mathcal{L}$, generalization error T, null sample size K)
2: Fit $\hat{\theta}$ by optimizing $\mathcal{L}(\mathcal{D}, F, \theta)$.
3: Compute the generalization loss on held out data, $t \leftarrow T(F_{\hat{\theta}}, \mathcal{D}')$.
4: **for** $k \leftarrow 1, \ldots, K$ **do**
5: Sample $\widetilde{X}'_j \sim P(X'_j | X'_{-j})$.
6: Create a new dataset $\widetilde{\mathcal{D}}'$ by replacing X'_j in $\mathcal{D}'$ with $\widetilde{X}'_j$.
7: Compute the generalization loss on the randomized heldout data, $\tilde{t}^{(k)} \leftarrow T(F_{\hat{\theta}}, \widetilde{\mathcal{D}}')$.
8: **return** A (one-sided) p-value,

$$p_j = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{I}\left[t \geq \tilde{t}^{(k)}\right]\right)$$

(Tansey et al., 2018)

Cross-Validation Randomization Test

procedure CVRT(training data split into M folds: $\mathcal{D} = \{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(M)}\}$, model F, training loss $\mathcal{L}$, generalization loss T, null sample size K)

  $t \leftarrow 0$

  for $m \leftarrow 1, \ldots, M$ **do**

    Fit $\hat{\theta}^{(m)}$ by optimizing $\mathcal{L}(\mathcal{D}^{(-m)}, F, \theta)$.

    Add the fold generalization loss, $t \leftarrow t + (1/M)\, T(F_{\hat{\theta}^{(m)}}, \mathcal{D}^{(m)})$.

  for $k \leftarrow 1, \ldots, K$ **do**

    Sample $\widetilde{X}_j \sim P(X_j | X_{-j})$.

    Create a new dataset $\widetilde{\mathcal{D}}$ by replacing X_j in $\mathcal{D}$ with $\widetilde{X}_j$.

    $\tilde{t}^{(k)} \leftarrow 0$

    for $m \leftarrow 1, \ldots, M$ **do**

      Add the fold generalization loss on the randomized data,

$$\tilde{t}^{(k)} \leftarrow \tilde{t}^{(k)} + (1/M)\, T(F_{\hat{\theta}^{(m)}}, \widetilde{\mathcal{D}}^{(m)})\,.$$

  return A (one-sided) p-value,

$$p_j = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{I}\left[t \geq \tilde{t}^{(k)}\right]\right)$$

Fast Approximation

1: **procedure** FASS(training data $\mathcal{D}$, test data $\mathcal{D}'$, model F, training loss $\mathcal{L}$, generalization loss T, sample-wise loss T')

2: Fit $\hat{\theta}$ by optimizing $\mathcal{L}(\mathcal{D}, F, \theta)$.

3: Compute the generalization loss on held out data, $t \leftarrow T(F_{\hat{\theta}}, \mathcal{D}')$.

4: Uniformly choose K thresholds $\mathbf{l}$ such that $\forall l_k \in \mathbf{l}, 0 \leq l_k \leq t$.

5: **for** $D'_i \in D'$ **do**

6: **for** $l_k \in \mathbf{l}$ **do**

7: Find the collection of maximal intervals (r, s) for x'_j in D'_i.

$$\mathcal{I}_{i,k} = \{(r, s) | r \leq x'_j \leq s, T'(F_{\hat{\theta}}, D'_i) \leq l_k\}.$$

8: Compute the probability that x'_j satisfies $T'(F_{\hat{\theta}}, D'_i) \leq l_j$,

$$P_{i,k} \leftarrow \sum_{(r,s) \in \mathcal{I}_{i,k}} \int_r^s Pr(x'_j) dx'_j.$$

9: Compute the difference in probability by a unit change in threshold l, i.e. from l_{k-1} to l_k, $Q_{i,k} \leftarrow P_{i,k} - P_{i,k-1}$

10: **return** A (one-sided) p-value,

$$p_j = \sum_{k=1}^{K} \sum_{u \leq k} \left(Q_{1,u} \sum_{v \leq k-u} \left(Q_{2,v} \ldots \sum_{w \leq k-u-v-\ldots} Q_{n,w} \right) \right)$$

Or Even Faster...

procedure FASSER(training data $\mathcal{D}$, test data $\mathcal{D}'$ with n' samples, model F, training loss $\mathcal{L}$, generalization loss T, sample-wise loss T', null sample size K, number of grids M, number of draws N)

  Fit $\hat{\theta}$ by optimizing $\mathcal{L}(\mathcal{D}, F, \theta)$.

  Compute the generalization loss on held out data, $t \leftarrow T(F_{\hat{\theta}}, \mathcal{D}')$.

  for $k \leftarrow 1, \ldots, K$ **do**

    Resample $\widetilde{X}_j'^{(k)} \sim P(X_j' | X_{-j}')$, where $\widetilde{X}_j'^{(k)} = \{\widetilde{X}_{j,1}'^{(k)}, \ldots \widetilde{X}_{j,n'}'^{(k)}\}$

  Generate a $n' \times M$ matrix $\widetilde{\mathbf{X}}$ where each row $\widetilde{\mathbf{X}}_{i,\cdot}$ are M evenly spaced grid points on the range $(\mu \pm 5\sigma)$ of $\{\widetilde{X}_{j,i}'^{(1)}, \ldots, \widetilde{X}_{j,i}'^{(K)}\}$

  for $m \leftarrow 1, \ldots M$ **do**

    Create new dataset $\widetilde{\mathcal{D}}'^{(m)}$ by substituting $\widetilde{X}_j'$ with $\widetilde{\mathbf{X}}_m$

    Calculate error array $\mathbf{t}_m \leftarrow T'(F_{\hat{\theta}}, \widetilde{\mathcal{D}}'^{(m)})$

  Generate $n' \times M$ t-matrix $\mathbf{T}$ by combining all $\mathbf{t}_m$'s

  Generate $n' \times M$ proposal probability matrix $\mathbf{Q}$ for $\widetilde{\mathbf{X}}$

  for $d \leftarrow 1 \ldots N$ **do**

    Sample $\widetilde{\mathbf{X}}_j$ by $\mathbf{Q}$

    Compute generalization error on the heldout data $\tilde{t}^{(n)}$

  return A (one-sided) p-value,

$$p_j = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} \mathbb{I}\left[t \geq \tilde{t}^{(n)}\right]\right)$$

Benchmarks

$$y = \sum_{j=0}^{N} [w_{4j}x_{4j} + w_{4j+1}x_{4j+1} + \tanh(w_{4j+2}x_{4j+2} + w_{4j+3}x_{4j+3})] + \sigma\epsilon\,, \tag{2}$$

where $\sigma = 0.5$ and $\epsilon \sim \mathcal{N}(0, 1)$. (Liang et al., 2018)

When $N = 0$, i.e. we have 4 variables, representing true signals, and 496 null variables,
HRTs had nearly 100% power and 0% FDR.

What about $N = 9$...

Benchmarks

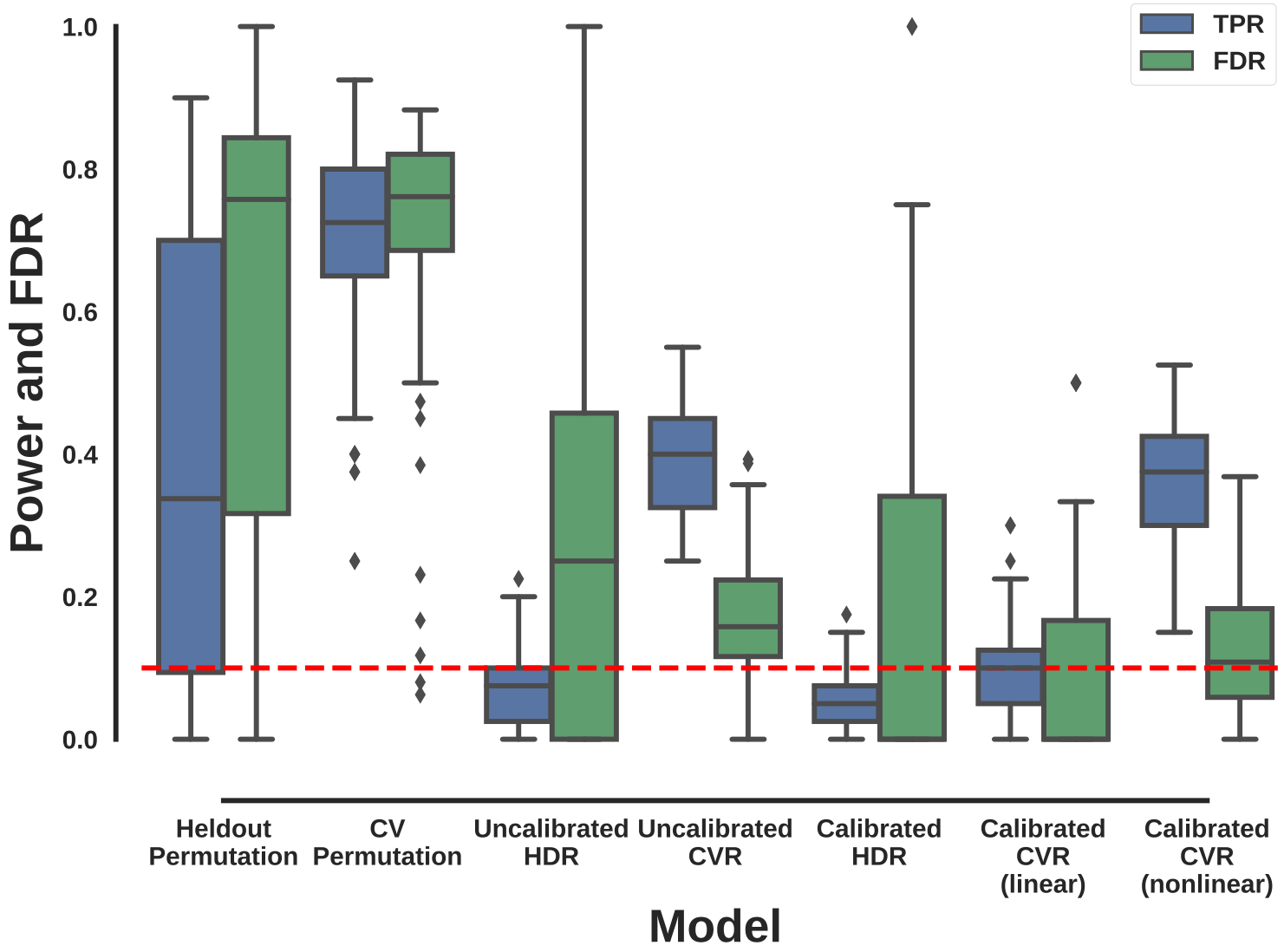

Figure 1: Power and FDR results for each model configuration on the benchmark simulation.

Review of the examples

What gene expression and mutation features affect cancer cell line sensitivity to the drug PLX4720?

What molecular features controls the perceived fragrant intensity of molecules?

(a) Elastic net for cancer drug response

Genomic Feature	**Coefficient weight**	**Est. *p*-value**
BRAF Mut	-0.1566	$\leq 10^{-6}$
RXRG	-0.0950	0.9867
HIP1 Mut	-0.0552	0.0019
GAPDHP36	-0.0503	0.7541
GRM7	-0.0493	1.0000
CTB-33O18.3	0.0421	0.3670
AC005324.7	-0.0396	0.0014
RP13-685P2.7	0.0365	0.9979
ZNF565	0.0356	0.1079
PRR7-AS1	-0.0337	0.6065

(b) Random forests for olfactory perception

Molecular Feature	**Importance score**	**Est. *p*-value**
B03[C-S]	0.0329	$\leq 10^{-6}$
F03[C-S]	0.0129	$\leq 10^{-6}$
LLS 01	0.0109	0.6481
SpAbs B(s)	0.0067	0.8994
SpMax8 Bh(s)	0.0067	0.9725
O-057	0.0066	0.0122
EXPaws	0.0054	0.9960
SP04	0.0047	0.9937
Cyclopentene	0.0045	0.9814
ATS2s	0.0044	0.9031

Figure 2: Two examples of predictive modeling with heuristic post-hoc variable selection in scientific studies.

Future

- for the field
  - Computational expensive in some model (e.g. empirical Bayes)
  - Selecting an individual covariate might not be meaningful in some applications (e.g. CV, NLP)
- for myself
  - Graduate School (hopefully at Columbia)