DeepPINK: reproducible feature selection in deep neural networks

Yang Young Lu*†, Yingying Fan*◊, Jinchi Lv◊, and William Stafford Noble†‡

Department of Genome Sciences, University of Washington†, Department of Computer Science and Engineering, University of Washington‡, Data Sciences and Operations Department, University of Southern California◊, Contributed equally*

NIPS

Neural Information Processing Systems

Abstract

Deep learning has become increasingly popular in both supervised and unsupervised machine learning thanks to its outstanding empirical performance. However, because of their intrinsic complexity, most deep learning methods are largely treated as black box tools with little interpretability. Even though recent attempts have been made to facilitate the interpretability of deep neural networks (DNNs), existing methods are susceptible to noise and lack of robustness. Therefore, scientists are justifiably cautious about the reproducibility of the discoveries, which is often related to the interpretability of the underlying statistical models.

We describe a method to increase the interpretability and reproducibility of DNNs by incorporating the idea of feature selection with controlled error rate. By designing a new DNN architecture and integrating it with the recently proposed knockoffs framework, we perform feature selection with a controlled error rate, while maintaining high power. This new method, DeepPINK (Deep feature selection using Paired-Input Nonlinear Knockoffs), is applied to both simulated and real data sets to demonstrate its empirical utility.

Question: feature selection with controlled error rate

The problem of feature selection:
Given n i.i.d. observations $(\mathbf{x}_i, Y_i)$, $i = 1, \cdots, n$, with $\mathbf{x}_i \in \mathbb{R}^p$ the feature vector and Y_i the response, select a feature subset $\widehat{S} \subset \{1, \cdots, p\}$ such that the features in the complement $\widehat{S}^c$ are conditionally independent of the response Y given $\widehat{S}$.

Evaluate feature selection performance by FDR:
Assume that $\mathcal{S}_0 \subset \{1, \cdots, p\}$ are truly relevant to the response Y. The goal is to identify features in $\mathcal{S}_0$ with a controlled false discovery rate (FDR). For the selected feature subset $\widehat{S}$, the FDR is defined as

$$\text{FDR} = \mathbb{E}[\text{FDP}] \text{ with FDP} = \frac{|\widehat{S} \cap \mathcal{S}_0^c|}{|\widehat{S}|},$$

where $|\cdot|$ stands for the cardinality of a set.

The knockoffs framework [1, 2]

The definition of knockoffs:
For random features $\mathbf{x} = (X_1, \cdots, X_p)^T$, the knockoffs $\tilde{\mathbf{x}} = (\tilde{X}_1, \cdots, \tilde{X}_p)^T$ of $\mathbf{x}$ satisfy the following two properties:

$$(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}(\mathcal{S})} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}}) \text{ for any subset } \mathcal{S} \subset \{1, \cdots, p\} \tag{1a}$$

$$\tilde{\mathbf{x}} \perp\!\!\!\perp Y | \mathbf{x} \tag{1b}$$

where $\text{swap}(\mathcal{S})$ means swapping X_j and $\tilde{X}_j$ for each $j \in \mathcal{S}$ and $\overset{d}{=}$ denotes equal in distribution. Also, $\tilde{\mathbf{x}}$ is independent of response Y given feature $\mathbf{x}$.

The construction of knockoffs:
If $\mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ the covariance matrix, the knockoffs can be constructed as:

$$\tilde{\mathbf{x}} | \mathbf{x} \sim N\big(\mathbf{x} - \text{diag}\{\mathbf{s}\}\mathbf{\Sigma}^{-1}\mathbf{x}, 2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\mathbf{\Sigma}^{-1}\text{diag}\{\mathbf{s}\}\big). \tag{2}$$

where $\text{diag}\{\mathbf{s}\}$ is a diagonal matrix with all components being $\mathbf{s}$. The original features and the model-X knockoff features have the following joint distribution

$$(\mathbf{x}, \tilde{\mathbf{x}}) \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\}\\ \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} & \mathbf{\Sigma}\end{pmatrix}\right). \tag{3}$$

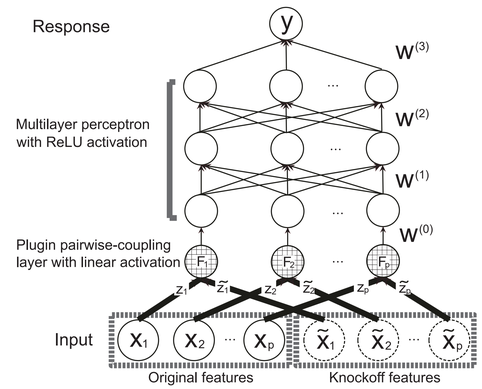
The feature importance score and the knockoff statistics:
Let Z_j and $\tilde{Z}_j$ be the feature importance score for the jth feature X_j and its knockoff $\tilde{X}_j$. Note that these scores are model-dependent, e.g. coefficients in LASSO regression.
Define the knockoff statistics as: $W_j = g_j(Z_j, \tilde{Z}_j)$, where $g_j(\cdot, \cdot)$ is an antisymmetric function (i.e. $g_j(Z_j, \tilde{Z}_j) = -g_j(\tilde{Z}_j, Z_j)$). A simple example is $W_j = Z_j - \tilde{Z}_j$. Important features should have large knockoff statistics whereas unimportant ones have small magnitudes symmetric around 0.

Feature selection by the knockoff statistics:
Sort $|W_j|$'s in decreasing order and select features whose W_j's exceed some threshold T, defined as

$$T_+ = \min\left\{t \in \mathcal{W}, \frac{1 + |\{j : W_j \le -t\}|}{1 \vee |\{j : W_j \ge t\}|} \le q\right\}, \tag{4}$$

where $\mathcal{W} = \big\{|W_j| : 1 \le j \le p\big\} \setminus \{0\}$ is the set of unique nonzero values attained by $|W_j|$'s and $q \in (0, 1)$ is the desired FDR level specified by the user.

Knockoff inference for DNNs



The feature importance score in DNNs
Consider two factors:

- The relative importance between X_j and $\tilde{X}_j$, $\mathbf{z} = (z_1, \cdots, z_p)^T$ and $\tilde{\mathbf{z}} = (\tilde{z}_1, \cdots, \tilde{z}_p)^T$
- The importance of the jth feature among all p features, $\mathbf{w} = W^{(0)} \odot (W^{(1)}W^{(2)}W^{(3)})$

Define Z_j and $\tilde{Z}_j$ as $Z_j = z_j \times \mathbf{w}_j$ and $\tilde{Z}_j = \tilde{z}_j \times \mathbf{w}_j$.
Define the knockoff statistic as $W_j = Z_j^2 - \tilde{Z}_j^2$.

Simulation studies on linear and non-linear models

Simulation settings: $n = 1000$, $p = \{50, 100, 200, 400, 600, 800, 1000, 1500, 2000, 2500, 3000\}$, FDR level $q = 0.2$.
Other competing algorithms: MLP, DeepLIFT, Random Forest, Support Vector Regression (SVR).
Gaussian linear regression models: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

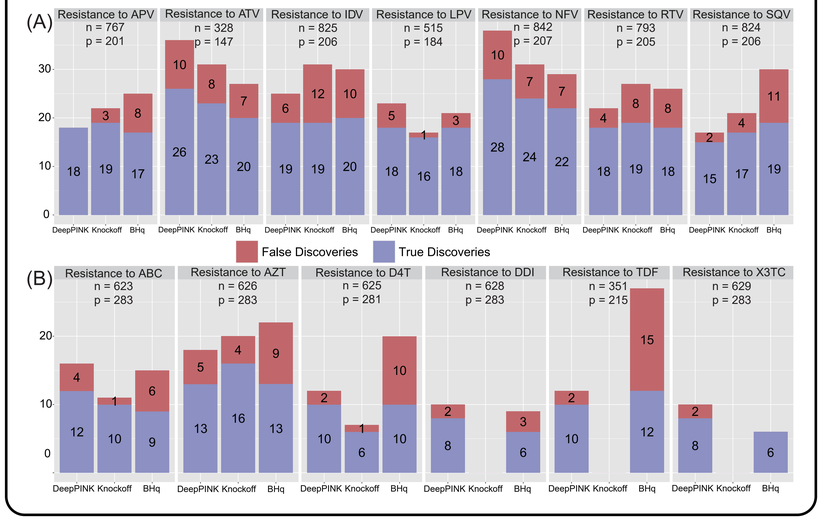
	DeepPINK		MLP		DeepLIFT		RF		SVR	
	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
p=50	0.046	1	0.15	1	0.16	1	0.005	0.45	0.18	1
p=100	0.047	1	0.048	1	0.16	1	0.016	0.61	0.22	1
p=200	0.042	0.99	0.11	1	0.24	0.96	0.013	0.54	0.21	1
p=400	0.022	0.97	0.29	0.95	0.034	0.5	0.017	0.53	0.22	1
p=600	0.031	0.95	0.17	0.8	0.003	0.26	0.023	0.56	0.19	1
p=800	0.048	0.95	0.037	0.62	0	0.17	0.022	0.61	0.22	0.98
p=1000	0.023	0.97	0.007	0.4	0	0.12	0.029	0.59	0.15	0.67
p=1500	0.007	1	0.002	0.41	0.001	0.32	0.045	0.58	0.064	0.043
p=2000	0.026	0.99	0.023	0.4	0.015	0.37	0.033	0.65	0.04	0.002
p=2500	0.029	0.97	0.21	0.5	0.088	0.58	0.034	0.62	0.02	0.005
p=3000	0.046	0.97	0.11	0.43	0.069	0.46	0.05	0.65	0.05	0

Single-Index models: $Y_i = g(\mathbf{x}_i^T\boldsymbol{\beta}) + \varepsilon_i$, for $\quad i = 1, \cdots, n$, where $g(x) = x^3/2$

	DeepPINK		MLP		DeepLIFT		RF		SVR	
	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
p=50	0.13	0.98	0.17	0.89	0.24	0.9	0	0	0.18	0.81
p=100	0.08	1	0.056	0.26	0.13	0.47	0.025	0.045	0.094	0.26
p=200	0.042	1	0	0	0.034	0.067	0.02	0.045	0.061	0.05
p=400	0.022	1	0	0	0.039	0.069	0.033	0.05	0.083	0.01
p=600	0.046	1	0.014	0.013	0.068	0.16	0.11	0.095	0	0
p=800	0.082	1	0.016	0.068	0.16	0.24	0.061	0.12	0	0
p=1000	0.065	1	0.037	0.16	0.013	0.33	0.081	0.17	0	0
p=1500	0.065	1	0.068	0.25	0.13	0.44	0.098	0.17	0	0
p=2000	0.098	1	0.063	0.35	0.1	0.56	0.046	0.14	0	0
p=2500	0.067	1	0.042	0.35	0.32	0.47	0.11	0.18	0	0
p=3000	0.051	1	0.046	0.31	0.14	0.44	0.087	0.17	0	0

Real application to HIV-1 data

Task: Identify mutations associated with drug resistance in HIV-1 [3].
Data: Two drug classes (PIs and NRTIs), Y is the log-transformed drug resistance level, X is the absence/presence of mutations. FDR level $q = 0.2$.
Other competing algorithms: The original fixed-X knockoff filter, the Benjamini-Hochberg procedure.
Evaluation: Compare against the treatment-selected mutations (gold standard).



Real application to gut microbiome data

Task: Identify the important nutrient intake and bacteria genera associated with body-mass index (BMI) [4].
Data: $n = 98$ volunteers, $p_1 = 214$ micronutrients and $p_2 = 87$ bacteria genera. FDR level $q = 0.2$.
Evaluation: Literature evidence.

	Nutrient intake	Bacteria genera	
	Micronutrient	Phylum	Genus
1	Linoleic	Firmicutes	Clostridium
2	Dairy Protein	Firmicutes	Acidaminococcus
3	Choline, Phosphatidylcholine	Firmicutes	Allisonella
4	Choline, Phosphatidylcholine w/o suppl.	Firmicutes	Megamonas
5	Omega 6	Firmicutes	Megasphaera
6	Phenylalanine, Aspartame	Firmicutes	Mitsuokella
7	Aspartic Acid, Aspartame	Firmicutes	Holdemania
8	Theaflavin 3-gallate, flavan-3-ol(2)	Proteobacteria	Sutterella

References

[1] E. Candès, *et al*. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, to appear, 2018.

[2] R. Barber, *et al*. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[3] S. Rhee, *et al*. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

[4] J. Chen, *et al*. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1), 2013.