UCLA Statistics

Neyman-Pearson Criterion (NPC): A Model Selection Criterion for Asymmetric Binary Classification

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

Model selection for binary classification: a motivating example

Automated disease diagnosis a binary classification problem

- Features/predictors: $\sim$ 20K human gene expression levels
- Response: binary disease status: 0 (diseased) and 1 (healthy)

sample	A1BG	A1CF	A2BP1	A2LD1	A2M
breast cancer tissue	0.2313966	0.6515808	0.4277823	0.8855574	0.6718676
breast cancer tissue	0.2479010	0.6686116	0.4653367	0.8820843	0.7234769
breast cancer tissue	0.2253524	0.7167720	0.3317310	0.7430520	0.8042019
breast cancer tissue	0.1732131	0.7957590	0.5316374	0.8725860	0.7594307
breast cancer tissue	0.2202075	0.6620401	0.3481031	0.7282685	0.7953009
normal tissue	0.2245508	0.7429916	0.5128035	0.8923495	0.7125915
breast cancer tissue	0.2641653	0.7466348	0.5220898	0.9173302	0.8038506
normal tissue	0.2418038	0.7126189	0.5523689	0.8840687	0.7019090
normal tissue	0.2168959	0.6747223	0.4917510	0.8137360	0.6560015
breast cancer tissue	0.2202682	0.7067618	0.5223232	0.8774032	0.7337275

Model selection for binary classification: a motivating example

Automated disease diagnosis a binary classification problem

- Features/predictors: $\sim$ 20K human gene expression levels
- Response: binary disease status: 0 (diseased) and 1 (healthy)

sample	A1BG	A1CF	A2BP1	A2LD1	A2M
breast cancer tissue	0.2313966	0.6515808	0.4277823	0.8855574	0.6718676
breast cancer tissue	0.2479010	0.6686116	0.4653367	0.8820843	0.7234769
breast cancer tissue	0.2253524	0.7167720	0.3317310	0.7430520	0.8042019
breast cancer tissue	0.1732131	0.7957590	0.5316374	0.8725860	0.7594307
breast cancer tissue	0.2202075	0.6620401	0.3481031	0.7282685	0.7953009
normal tissue	0.2245508	0.7429916	0.5128035	0.8923495	0.7125915
breast cancer tissue	0.2641653	0.7466348	0.5220898	0.9173302	0.8038506
normal tissue	0.2418038	0.7126189	0.5523689	0.8840687	0.7019090
normal tissue	0.2168959	0.6747223	0.4917510	0.8137360	0.6560015
breast cancer tissue	0.2202682	0.7067618	0.5223232	0.8774032	0.7337275

Given a classification method (e.g., logistic regression),

- What subset of genes exhibits the highest diagnostic power?
  A model selection problem

Two binary classification paradigms: classical vs. Neyman-Pearson

Paradigm	Oracle classifier	Practical classifier
Classical	$\phi^* = \arg\min_{\phi} R(\phi)$	$\widehat{\phi} = \arg\min_{\phi} \widehat{R}(\phi)$
Neyman-Pearson	$\phi^*_\alpha = \arg\min_{R_0(\phi)\le\alpha} R_1(\phi)$	$\widehat{\phi}_\alpha$ by the NP umbrella algorithm
	[Rigollet and Tong (2011)]	[Tong, Feng and Li (2018)]

where α is a user-specified upper bound on the type I error

Two binary classification paradigms: classical vs. Neyman-Pearson

Paradigm	Oracle classifier	Practical classifier
Classical	$\phi^* = \arg\min_{\phi} R(\phi)$	$\widehat{\phi} = \arg\min_{\phi} \widehat{R}(\phi)$
Neyman-Pearson	$\phi^*_\alpha = \arg\min_{R_0(\phi)\leq\alpha} R_1(\phi)$	$\widehat{\phi}_\alpha$ by the NP umbrella algorithm
	[Rigollet and Tong (2011)]	[Tong, Feng and Li (2018)]

where α is a user-specified upper bound on the type I error

The Neyman-Pearson paradigm is suitable for disease diagnosis

Two binary classification paradigms: classical vs. Neyman-Pearson

Paradigm	Oracle classifier	Practical classifier
Classical	$\phi^* = \arg\min_{\phi} R(\phi)$	$\widehat{\phi} = \arg\min_{\phi} \widehat{R}(\phi)$
Neyman-Pearson	$\phi^*_\alpha = \arg\min_{R_0(\phi)\le\alpha} R_1(\phi)$	$\widehat{\phi}_\alpha$ by the NP umbrella algorithm
	[Rigollet and Tong (2011)]	[Tong, Feng and Li (2018)]

where α is a user-specified upper bound on the type I error

The Neyman-Pearson paradigm is suitable for disease diagnosis

- The two classes have asymmetric importance
  - Mispredicting a normal tissue sample as malignant
    $\Rightarrow$ patient anxiety & additional medical costs—type II error $R_1(\phi)$
  - Mispredicting a tumor sample as normal
    $\Rightarrow$ life loss 😭—type I error $R_0(\phi)$

  Policy makers often like to enforce a pre-specified threshold α on $R_0(\phi)$

Two binary classification paradigms: classical vs. Neyman-Pearson

Paradigm	Oracle classifier	Practical classifier
Classical	$\phi^* = \arg\min_{\phi} R(\phi)$	$\widehat{\phi} = \arg\min_{\phi} \widehat{R}(\phi)$
Neyman-Pearson	$\phi^*_\alpha = \arg\min_{R_0(\phi)\leq\alpha} R_1(\phi)$	$\widehat{\phi}_\alpha$ by the NP umbrella algorithm
	[Rigollet and Tong (2011)]	[Tong, Feng and Li (2018)]

where α is a user-specified upper bound on the type I error

The Neyman-Pearson paradigm is suitable for disease diagnosis

Two binary classification paradigms: classical vs. Neyman-Pearson

Paradigm	Oracle classifier	Practical classifier
Classical	$\phi^* = \arg\min_{\phi} R(\phi)$	$\widehat{\phi} = \arg\min_{\phi} \widehat{R}(\phi)$
Neyman-Pearson	$\phi^*_\alpha = \arg\min_{R_0(\phi)\le\alpha} R_1(\phi)$	$\widehat{\phi}_\alpha$ by the NP umbrella algorithm
	[Rigollet and Tong (2011)]	[Tong, Feng and Li (2018)]

where α is a user-specified upper bound on the type I error

The Neyman-Pearson paradigm is suitable for disease diagnosis

- The two classes have imbalanced sample sizes

$$R(\phi) = \mathbb{P}(Y = 0)R_0(\phi) + \mathbb{P}(Y = 1)R_1(\phi)$$

When $\mathbb{P}(Y = 0) = \mathbb{P}(\text{diseased}) \ll \mathbb{P}(Y = 1) = \mathbb{P}(\text{healthy})$,

  - The classical oracle classifier may have an excessively large $R_0(\phi)$
  - The NP oracle classifier will have $R_0(\phi) \le \alpha$

Paper and software of the Neyman-Pearson umbrella algorithm

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

Neyman-Pearson classification algorithms and NP receiver operating characteristics

Xin Tong,[1*†] Yang Feng,[2†] Jingyi Jessica Li[3*]

In many binary classification applications, such as disease diagnosis and spam detection, practitioners commonly face the need to limit type I error (that is, the conditional probability of misclassifying a class 0 observation as class 1) so that it remains below a desired threshold. To address this need, the Neyman-Pearson (NP) classification paradigm is a natural choice; it minimizes type II error (that is, the conditional probability of misclassifying a class 1 observation as class 0) while enforcing an upper bound, α, on the type I error. Despite its century-long history in hypothesis testing, the NP paradigm has not been well recognized and implemented in classification schemes. Common practices that directly limit the empirical type I error to no more than α do not satisfy the type I error control objective because the resulting classifiers are likely to have type I errors much larger than α, and the NP paradigm has not been properly implemented in practice. We develop the first umbrella algorithm that implements the NP paradigm for all scoring-type classification methods, such as logistic regression, support vector machines, and random forests. Powered by this algorithm, we propose a novel graphical tool for NP classification methods: NP receiver operating characteristic (NP-ROC) bands motivated by the popular ROC curves. NP-ROC bands will help choose α in a data-adaptive way and compare different NP classifiers. We demonstrate the use and properties of the NP umbrella algorithm and NP-ROC bands, available in the R package nproc, through simulation and real data studies.

R package `nproc`

https://CRAN.R-project.org/package=nproc

Email: jli@stat.ucla.edu

Model selection under the Neyman-Pearson paradigm (population)

- **Goal**: Develop a model selection criterion to compare models (i.e., feature subsets) under the Neyman-Pearson (NP) paradigm

Model selection under the Neyman-Pearson paradigm (population)

- **Goal**: Develop a model selection criterion to compare models (i.e., feature subsets) under the Neyman-Pearson (NP) paradigm

- **Idea**: prediction-based model selection
  - **Compare two feature subsets based on the type II errors of their corresponding NP oracle classifiers**
    in contrast to
  - Compare two feature subsets based on the risks of their corresponding classical oracle classifiers

Model selection under the Neyman-Pearson paradigm (population)

- **Goal**: Develop a model selection criterion to compare models (i.e., feature subsets) under the Neyman-Pearson (NP) paradigm

- **Idea**: prediction-based model selection
  - **Compare two feature subsets based on the type II errors of their corresponding NP oracle classifiers**
    in contrast to
  - Compare two feature subsets based on the risks of their corresponding classical oracle classifiers

- **Question**: Will the model selection results be different under the two paradigms?

Prediction-based model selection under the two paradigms (population)

A linear discriminant analysis (LDA) example

Two features $X_1, X_2 \in \mathbb{R}$ with the following class conditional distributions:

$$X_{\{1\}} \mid (Y = 0) \sim \mathcal{N}(-5, 2^2), \qquad X_{\{1\}} \mid (Y = 1) \sim \mathcal{N}(0, 2^2),$$
$$X_{\{2\}} \mid (Y = 0) \sim \mathcal{N}(-5, 2^2), \qquad X_{\{1\}} \mid (Y = 1) \sim \mathcal{N}(1.5, 3.5^2).$$

We would like to select the better feature between the two

- Classical oracle classifiers:

$$R\left(\phi^*_{\{1\}}\right) = 0.106 < R\left(\phi^*_{\{2\}}\right) = 0.113$$

So X_1 is the better feature

Prediction-based model selection under the two paradigms (population)

A linear discriminant analysis (LDA) example

Two features $X_1, X_2 \in \mathbb{R}$ with the following class conditional distributions:

$$X_{\{1\}} \mid (Y = 0) \sim \mathcal{N}(-5, 2^2), \qquad X_{\{1\}} \mid (Y = 1) \sim \mathcal{N}(0, 2^2),$$
$$X_{\{2\}} \mid (Y = 0) \sim \mathcal{N}(-5, 2^2), \qquad X_{\{1\}} \mid (Y = 1) \sim \mathcal{N}(1.5, 3.5^2).$$

We would like to select the better feature between the two

- Classical oracle classifiers:

$$R\left(\phi^*_{\{1\}}\right) = 0.106 < R\left(\phi^*_{\{2\}}\right) = 0.113$$

  So X_1 is the better feature

- Neyman-Pearson (NP) oracle classifiers:

$$R_1\left(\phi^*_{\alpha\{1\}}\right) \text{ vs. } R_1\left(\phi^*_{\alpha\{2\}}\right)$$

  - $\alpha = 0.01$, $R_1\left(\phi^*_{\alpha\{1\}}\right) = 0.431 > R_1\left(\phi^*_{\alpha\{2\}}\right) = 0.299 \Longrightarrow X_2$ better
  - $\alpha = 0.20$, $R_1\left(\phi^*_{\alpha\{1\}}\right) = 0.049 < R_1\left(\phi^*_{\alpha\{2\}}\right) = 0.084 \Longrightarrow X_1$ better

Prediction-based model selection under the two paradigms (population)

Special scenarios where prediction-based model selection is the same under the two paradigms

Lemma 1

$$X_{\{1\}} \mid (Y=0) \sim \mathcal{N}\left(\mu_1^0, (\sigma_1^0)^2\right), \qquad X_{\{1\}} \mid (Y=1) \sim \mathcal{N}\left(\mu_1^1, (\sigma_1^1)^2\right),$$
$$X_{\{2\}} \mid (Y=0) \sim \mathcal{N}\left(\mu_2^0, (\sigma_2^0)^2\right), \qquad X_{\{2\}} \mid (Y=1) \sim \mathcal{N}\left(\mu_2^1, (\sigma_2^1)^2\right).$$

Let $\alpha \in (0,1)$, $\phi^*_{\alpha\{1\}}$ *and* $\phi^*_{\alpha\{2\}}$ *be the NP oracle classifiers based on* $X_{\{1\}}$ *and* $X_{\{2\}} \in \mathbb{R}$ *respectively, and* $\phi^*_{\{1\}}$ *and* $\phi^*_{\{2\}}$ *be the corresponding classical oracle classifiers. If*

$$\frac{\sigma_2^0}{\sigma_2^1} = \frac{\sigma_1^0}{\sigma_1^1},$$

then

$$\operatorname{sign}\left(R_1\left(\phi^*_{\alpha\{1\}}\right) - R_1\left(\phi^*_{\alpha\{2\}}\right)\right) = \operatorname{sign}\left(R\left(\phi^*_{\{1\}}\right) - R\left(\phi^*_{\{2\}}\right)\right).$$

for all α. *The reverse statement also holds.*

Prediction-based model selection under the two paradigms (population)

Special scenarios where prediction-based model selection is the same under the two paradigms

Lemma 2

Let A_1, $A_2 \subseteq \{1, \ldots, d\}$ be two index sets. For a random vector $\boldsymbol{X} \in \mathbb{R}^d$, let $\boldsymbol{X}_{A_1}$ and $\boldsymbol{X}_{A_2}$ be sub-vectors of $\boldsymbol{X}$ comprising of coordinates with indexes in A_1 and A_2 respectively, and assume they follow the class conditional distributions:

$$\boldsymbol{X}_{A_1} \,|\, (Y = 0) \sim \mathcal{N}(\boldsymbol{\mu}_1^0, \boldsymbol{\Sigma}_1)\,, \qquad \boldsymbol{X}_{A_1} \,|\, (Y = 1) \sim \mathcal{N}(\boldsymbol{\mu}_1^1, \boldsymbol{\Sigma}_1)\,,$$
$$\boldsymbol{X}_{A_2} \,|\, (Y = 0) \sim \mathcal{N}(\boldsymbol{\mu}_2^0, \boldsymbol{\Sigma}_2)\,, \qquad \boldsymbol{X}_{A_2} \,|\, (Y = 1) \sim \mathcal{N}(\boldsymbol{\mu}_2^1, \boldsymbol{\Sigma}_2)\,,$$

where $\boldsymbol{\mu}_j^i \in \mathbb{R}^\ell$ denotes the mean vector of feature set A_j in class i, and $\boldsymbol{\Sigma}_j \in \mathbb{R}^{\ell \times \ell}$ denotes the variance-covariance matrix of feature set A_j, $j = 1, 2$, $i = 0, 1$.

The selected feature set $A_\alpha^* = A_1$ or A_2 under the NP paradigm is invariant to α and is consistent with the selected feature set A^* under the classical paradigm.

Model selection under the Neyman-Pearson paradigm (sample)

- **Goal**: Develop a practical prediction-based model selection criterion under the NP paradigm

Model selection under the Neyman-Pearson paradigm (sample)

- **Goal**: Develop a practical prediction-based model selection criterion under the NP paradigm

- **Idea**: Compare two feature subsets based on the estimated type II errors of their corresponding NP classifiers (constructed by our NP umbrella algorithm)

Model selection under the Neyman-Pearson paradigm (sample)

- **Goal**: Develop a practical prediction-based model selection criterion under the NP paradigm

- **Idea**: Compare two feature subsets based on the estimated type II errors of their corresponding NP classifiers (constructed by our NP umbrella algorithm)

- **Question**: How to construct a "good" estimator of the type II error of an NP classifier?

Model selection under the Neyman-Pearson paradigm (sample)

- **Goal**: Develop a practical prediction-based model selection criterion under the NP paradigm

- **Idea**: Compare two feature subsets based on the estimated type II errors of their corresponding NP classifiers (constructed by our NP umbrella algorithm)

- **Question**: How to construct a "good" estimator of the type II error of an NP classifier?
  Leave out some class 1 data!

The model selection criterion: NPC (Neyman-Pearson Criterion)

Statistical formulation

Given $\alpha, \delta \in [0,1]$, a classification method, and a feature subset $A \subseteq \{1, \ldots, p\}$, a practical NP classifier $\widehat{\phi}_{\alpha A}$ is constructed by the NP umbrella algorithm. Then the NPC for model A at level α is defined as

$$\text{NPC}_{\alpha A} := \widehat{R}_1(\widehat{\phi}_{\alpha A})$$

where $\widehat{R}_1(\phi)$ is the estimated type II error of any classifier ϕ on leave-out class 1 data

- Sample splitting: split training data into three parts
  - mixed classes 0 and 1 sample
  - left-out class 0 sample

  $\Big\}$ NP umbrella algorithm $\Longrightarrow$ $\widehat{\phi}_{\alpha A}$

  - left-out class 1 sample

  $\overset{\widehat{\phi}_{\alpha A}}{\Longrightarrow}$ $\text{NPC}_{\alpha A}$
- Multiple random splits can be used to construct an ensemble estimator with a smaller variance

Theoretical properties of NPC

Concentration of $\widehat{R}_1(\widehat{\phi}_{\alpha A})$ at $R_1(\phi^*_{\alpha A})$

Given reasonable conditions on

(1) the two class conditional distributions of $X \mid (Y = 0)$ and $X \mid (Y = 1)$

(2) the scoring function of the given classification method

(3) the two class samples sizes

we can show that with probability at least $1 - \delta'$

$$|\widehat{R}_1(\widehat{\phi}_{\alpha A}) - R_1(\phi^*_{\alpha A})| \leq C(\delta')$$

where

- $R_1(\phi^*_{\alpha A})$ is the population type II error of the method-specific oracle classifier $\phi^*_{\alpha A}$ (the classifier that shares the same scoring function as the 'best' classical classifier constrained by the classification method)
- The deviation upper bound $C(\delta')$ increases as δ' decreases; also, $C(\delta')$ converges to zero as the sample sizes go to infinity

Real data application: DNA methylation features for cancer diagnosis

A glance at data (Fleischer *et al.*, 2014 *Genome Biology*):

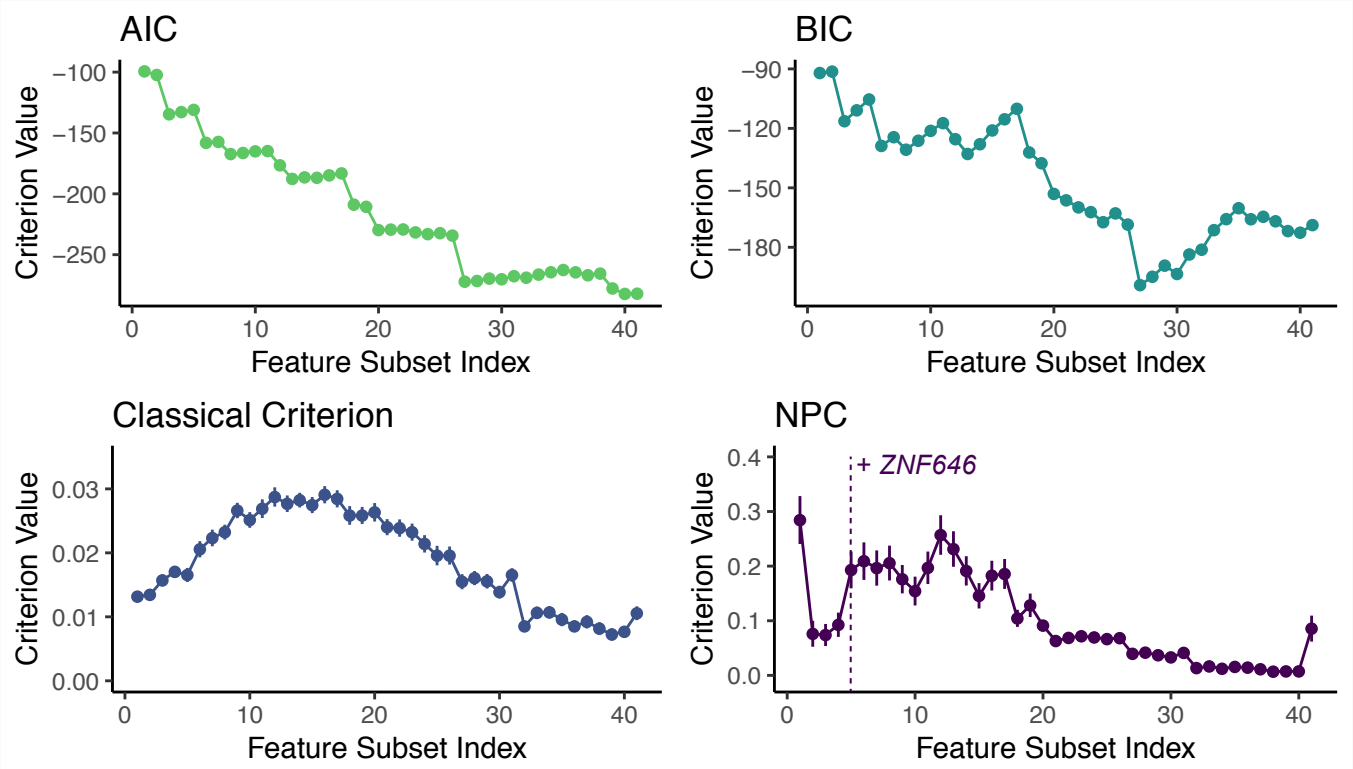
- 46 (class 1) normal tissues V.S. 239 breast cancer (class 0) tissue
- Methylation levels are measured at $468,424$ genome locations in every tissue
- After preprocessing and normalization, $d = 19,363 \gg n = 285$

Real data application: DNA methylation features for cancer diagnosis

Use L_1 penalized logistic regression to generate a solution path

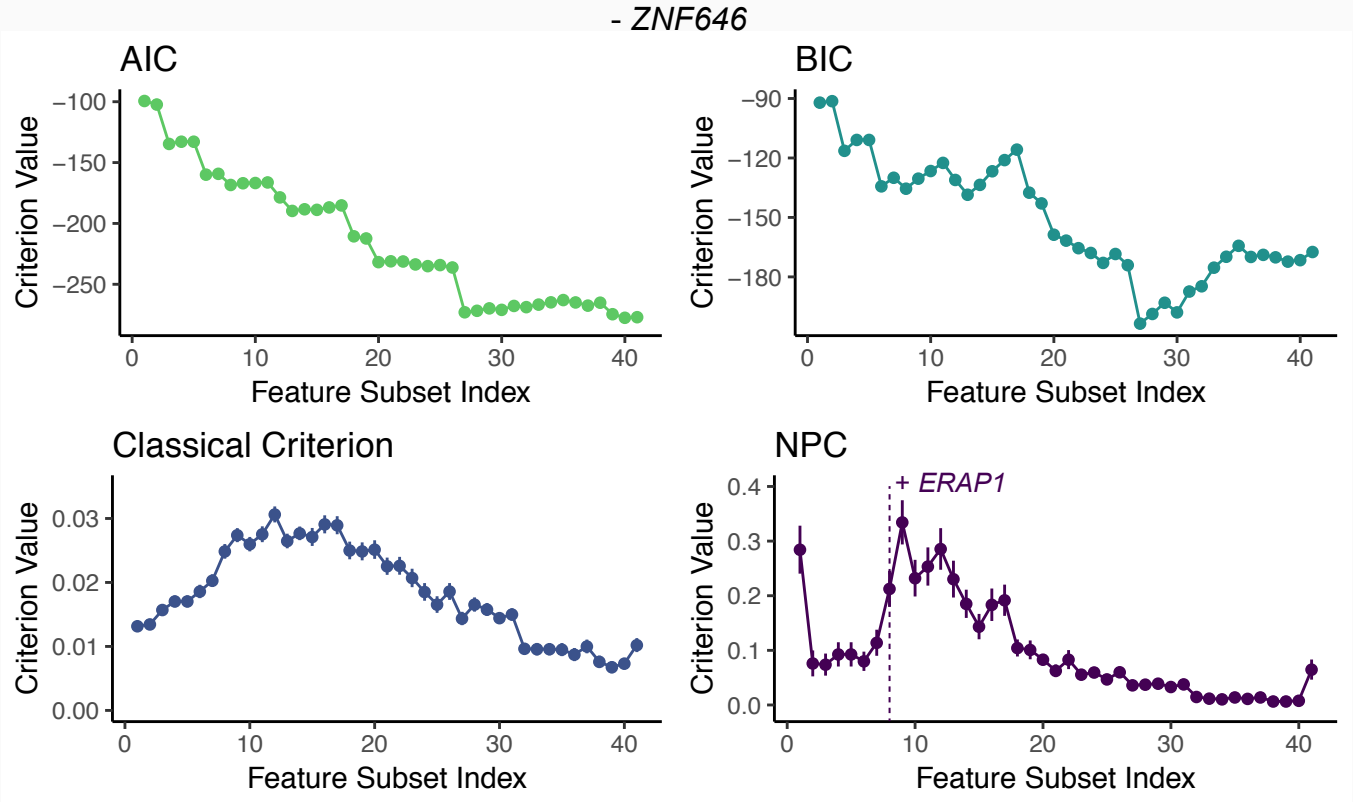
$\Longrightarrow$ 41 nested feature sets

Apply model selection criteria:



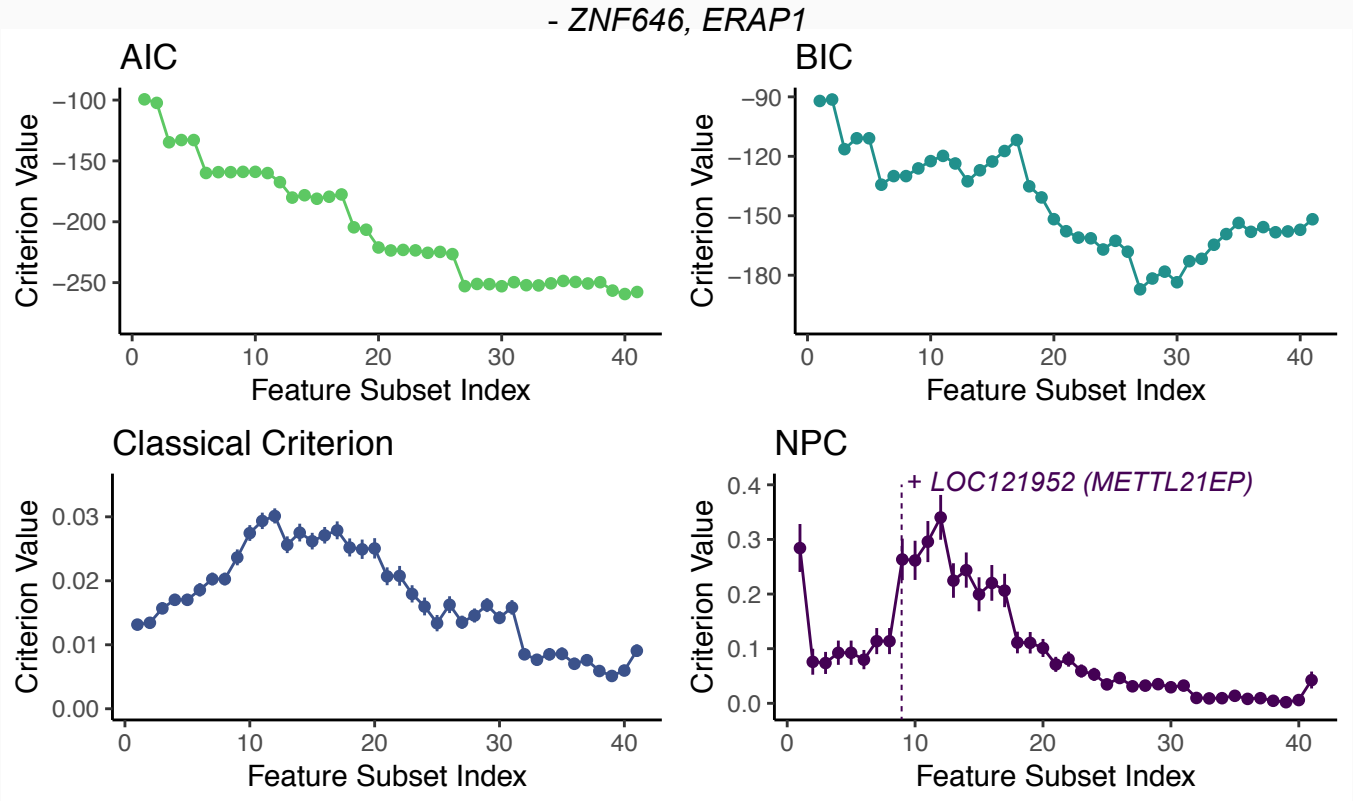
Remove: *ZNF646* . . .

Real data application: DNA methylation features for cancer diagnosis



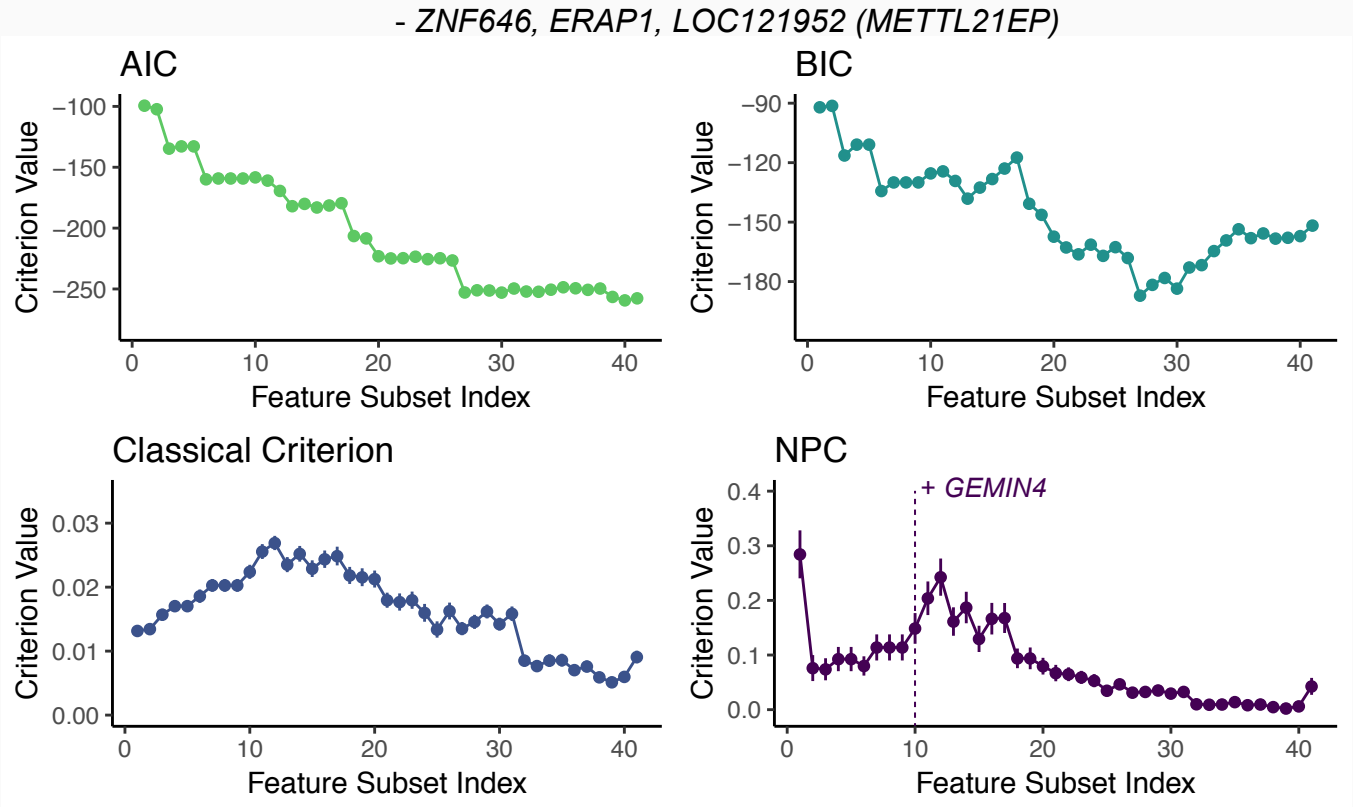
Remove: *ZNF646*, *ERAP1* . . .

Real data application: DNA methylation features for cancer diagnosis



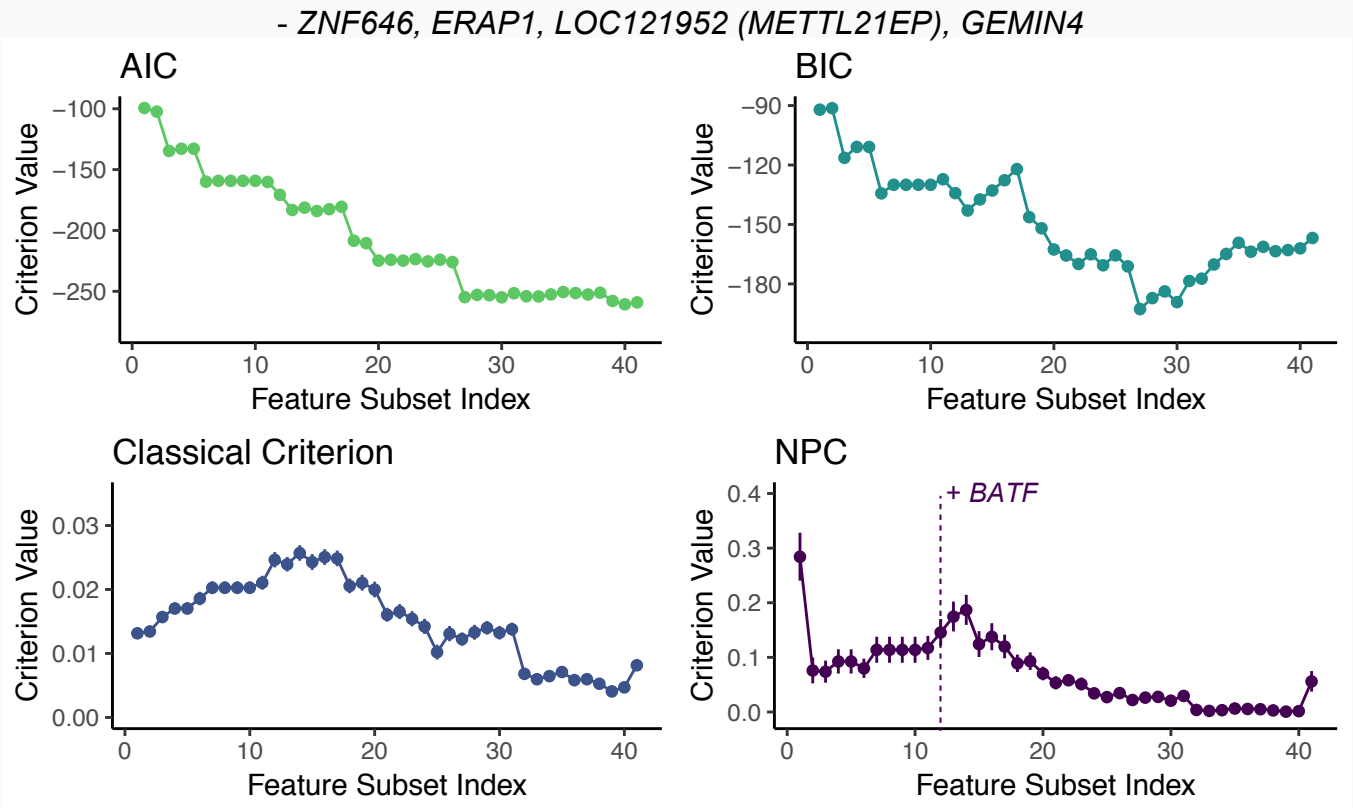
Remove: *ZNF646*, *ERAP1*, *LOC121952* . . .

Real data application: DNA methylation features for cancer diagnosis



Remove: *ZNF646*, *ERAP1*, *LOC121952*, *GEMIN4* . . .

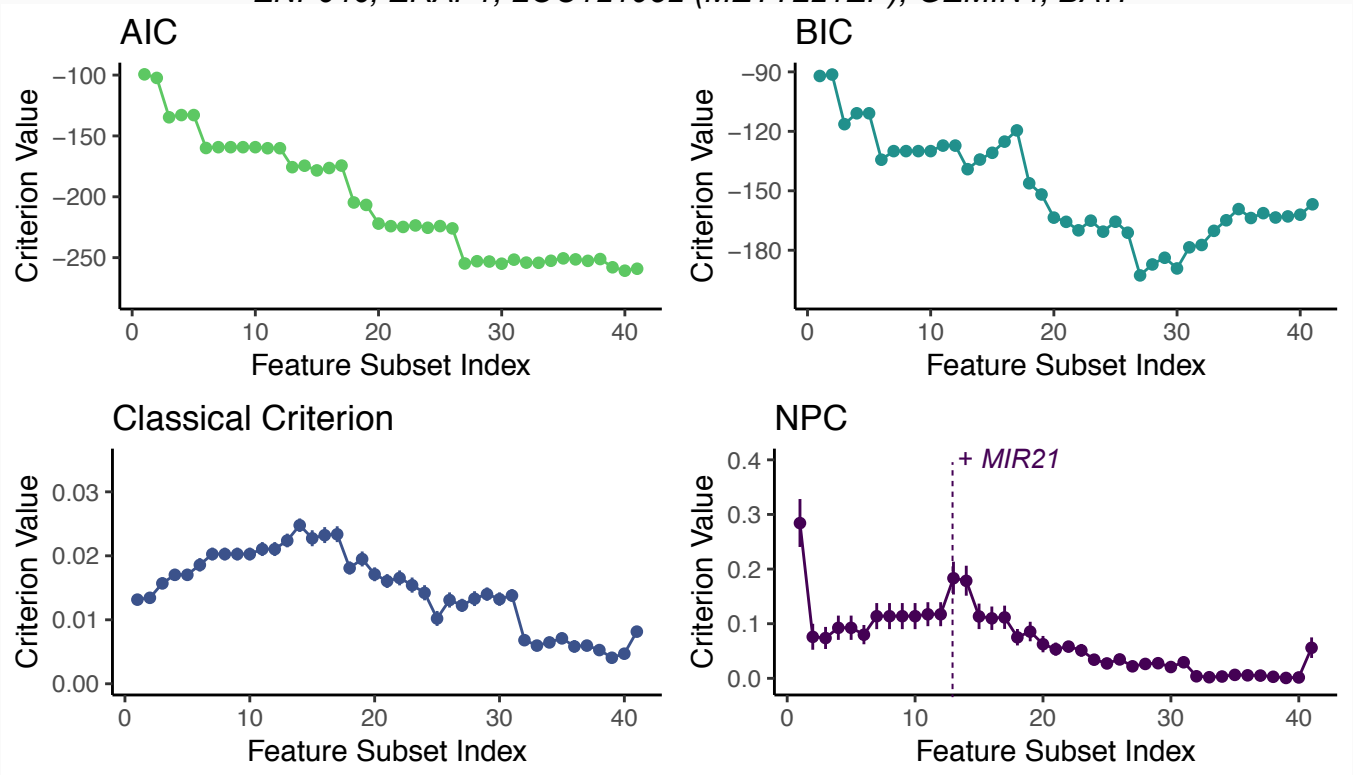
Real data application: DNA methylation features for cancer diagnosis



Remove: *ZNF646*, *ERAP1*, *LOC121952*, *GEMIN4*, *BATF* ...

Real data application: DNA methylation features for cancer diagnosis

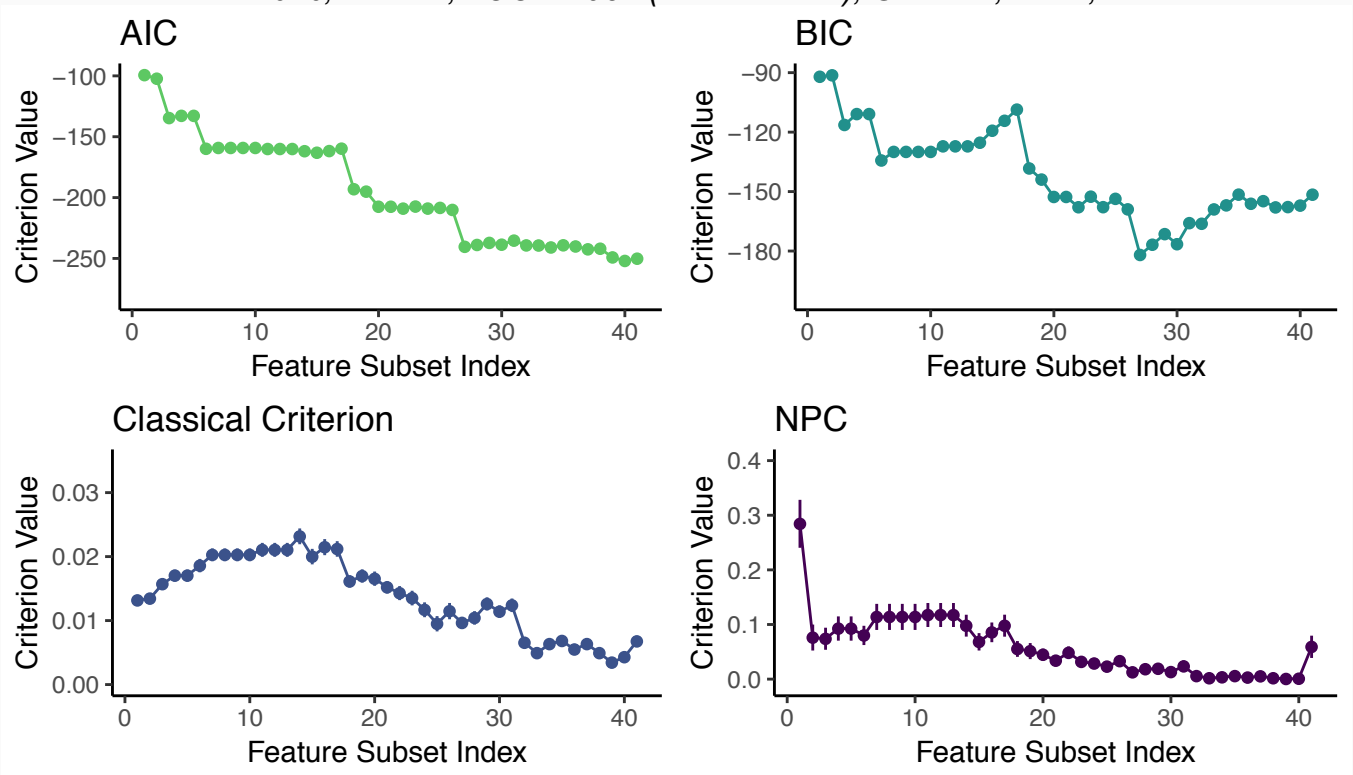
- ZNF646, ERAP1, LOC121952 (METTL21EP), GEMIN4, BATF



Remove: *ZNF646*, *ERAP1*, *LOC121952*, *GEMIN4*, *BATF*, *MIR21* . . .

Real data application: DNA methylation features for cancer diagnosis

- ZNF646, ERAP1, LOC121952 (METTL21EP), GEMIN4, BATF, MIR21



No obvious rise in NPC

Real data application: DNA methylation features for cancer diagnosis

Among the 41 genes,

	protein-coding	**microRNA**	**pseudogene**
remained	32 (20)	3	0
removed	4 (2)	1	1

() means the number of genes that express proteins in breast cancer tissues according to the Human Protein Atlas database

Observations:

- 9 out of 32 genes do not yet have available protein expression data in breast cancer tissues in the Human Protein Atlas database
- A specificity higher than 90% is achievable with only three gene markers: *HMGB2*, *MIR195* and *SPARCL1*. Inclusion of *SPARCL1* increases specificity from $\sim$ 70% to over 90%

Conclusions

- A new model selection criterion: NPC, tailored for asymmetric binary classification under the NP paradigm
- NPC allows users to select a model that achieve the best specificity among candidate models while maintaining a high sensitivity
- Useful in disease diagnosis, and other applications (network security control, loan screening and prediction of regional conflicts)
- Flexible to the choice of classification methods and the way of constructing NP classifiers

Acknowledgements

arXiv.org > stat > arXiv:1903.05262

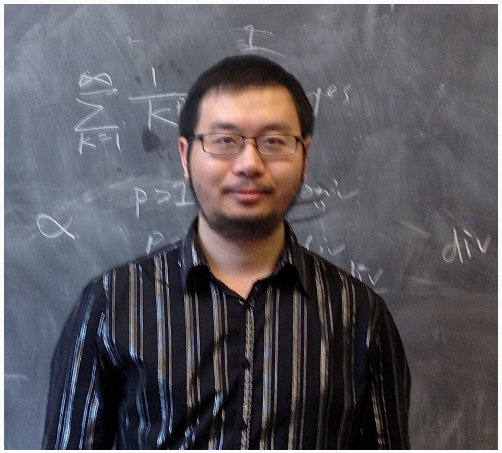
Statistics > Methodology

Neyman-Pearson Criterion (NPC): A Model Selection Criterion for Asymmetric Binary Classification

Yiling Chen, Jingyi Jessica Li, Xin Tong

(Submitted on 12 Mar 2019)

Yiling Chen
(UCLA)



Dr. Xin Tong
(USC)