

# Automatically mining relevant variable interactions via sparse Bayesian learning

Ryoichiro Yafune<sup>1</sup>, Daisuke Sakuma<sup>1</sup>, Mirai Takayanagi<sup>1</sup>

Yasuo Tabei<sup>2</sup>, Noritaka Saito<sup>1</sup>, and Hiroto Saigo<sup>1</sup>

<sup>1</sup> Kyushu University, <sup>2</sup> RIKEN Center for Advanced Intelligence Project

# Index

- Motivation
- Preparation
- Proposed Method
- Experiment HIV-1 dataset
- Experiment \*\*\* dataset
- Conclusion

# Motivation

	Interaction	Interpre- tation	Probabilistic model
Least Squares Method	X	✓	X
Bayesian Linear Regression	X	✓	✓
S V M	✓	X	X
Gaussian Process	✓	X	✓
Neural Network	✓	X	X
Decision Tree	✓	✓	X
LASSO	✓	✓	X
iSB (Proposed)	✓	✓	✓

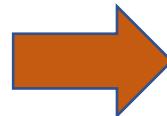
# Proposed method: itemset Sparse Bayes(iSB)

- Interaction considerations
- High accuracy
- Bayesian estimation
  - There is no need to tune parameters.
  - Applications such as Bayesian optimization can be used.
- Interpretable
  - High sparsity reveals effective features.
  - Functions that use kernel functions such as SVM, Gaussian processes and neural networks cannot interpret functions.

# What is itemset?

- It means combination of items.
- It can consider interaction but the number of itemset is too large.
- We can get all itemset from dataset by using itemset mining algorithm. [T.Uno 2003]

TID	Data
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}



{A}	{B,D}
{B}	{C,D}
{C}	{A,B,C}
{D}	{A,B,D}
{A,B}	{A,C,D}
{A,C}	{B,C,D}
{A,D}	{A,B,C,D}
{B,C}	

# Sparse Bayes Learning [Tipping 2001]

$$\mathbf{y} = \mathbf{X} \mathbf{w} + \epsilon$$

Design matrix                            error  
Target vector                              weight

If  $\epsilon \sim N(0, \sigma^2)$ , likelihood of  $\mathbf{w}$  is expressed in the following formula:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{x}_i^T \mathbf{w} - y_i)^2}{2\sigma^2}\right)$$

, which  $\mathbf{x}_i$  means vector of  $\mathbf{X}$ 's  $i$ th rows.

# Sparse Bayes Learning [Tipping 2001]

To make  $\mathbf{w}$  sparse, we introduce hyperparameters  $\alpha$  into the prior distribution. [Wipf 2004]

We calculate posterior from prior and likelihood using Bayes theorem.

$$p(\mathbf{w} | \mathbf{y}, \alpha) \sim p(\mathbf{y} | \mathbf{w}, \alpha)p(\mathbf{w} | \alpha)$$

This follows normal distribution  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  :

$$\begin{aligned}\boldsymbol{\Sigma} &= (\mathbf{A} + \sigma^{-2} \mathbf{X}^T \boldsymbol{\Phi})^{-1}, \\ \boldsymbol{\mu} &= \sigma^{-2} \boldsymbol{\Sigma} \mathbf{X}^T \mathbf{y}\end{aligned}$$

, where  $\mathbf{A} = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_p)$ .

# Algorithm

Input :  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ,  $\mathbf{y}$

1:  $\sigma^2 = \text{var}(\mathbf{y})$

2: Select the best itemset  $\mathbf{x}_i$

3: Calculate  $\alpha_i$  and Initialize  $\Sigma$  and  $\mu$

while do

4: Select the best itemset  $\mathbf{x}_k$ .

5: Calculate  $\alpha_k$  and update  $\sigma^2$ ,  $\Sigma$  and  $\mu$

end while

Output :  $\Sigma$ ,  $\mu$ ,  $\sigma^2$

# Experiment of HIV-1 dataset

- We experiment using anti-HIV-1 drug mutation and drug resistance data. [Rhee 2003]
- 6 drugs were used. (3TC, ABC, AZT, DDI, D4T, TDF)
- This data includes non-linearity in which drug resistance increases suddenly when multiple mutations appear.

	3tc	ABC	AZT	DDI	D4T	TDF
# of samples	633	628	630	632	630	353
# of mutation	348	348	348	348	348	348

# Experiment of HIV-1 dataset

## Accuracy

$$r^2 = \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- We compared the accuracy of test data .
- This table reveals 5 Fold Cross Validation.
- iSB is better than SB because of non-linear effect.

No-parameter tuning		Anti-HIV drug dataset						mean
		3TC(633)	ABC(628)	AZT(630)	D4T(630)	DDI(632)	TDF(353)	
linear	OLS ✓	0.67(0.23)	-100≤	0.77(0.28)	-0.31(0.41)	-0.54(0.56)	-53.18(61.76)	0.24(0.37)
	SBL/ARD ✓	0.71(0.05)	0.50(0.06)	0.64(0.05)	0.62(0.06)	0.57(0.16)	0.47(0.08)	0.59(0.09)
	LARS	0.41(0.03)	0.59(0.09)	0.61(0.10)	0.63(0.07)	0.61(0.09)	0.47(0.15)	0.55(0.09)
	KNN	0.59(0.03)	0.47(0.09)	0.69(0.09)	0.45(0.05)	0.34(0.06)	0.20(0.18)	0.46(0.17)
Non-linear	CART	<b>0.96(0.02)</b>	0.61(0.09)	<b>0.92(0.16)</b>	0.57(0.10)	0.55(0.09)	-0.08(0.28)	0.60(0.35)
	GP ✓	0.89(0.02)	0.41(0.19)	0.85(0.11)	0.34(0.15)	0.21(0.14)	-0.10(0.33)	0.62(0.32)
	SVR	<b>0.92(0.02)</b>	<b>0.70(0.06)</b>	<b>0.87(0.08)</b>	<b>0.67(0.10)</b>	<b>0.60(0.04)</b>	<b>0.51(0.06)</b>	<b>0.71(0.16)</b>
	MLP	<b>0.92(0.02)</b>	<b>0.69(0.04)</b>	<b>0.85(0.04)</b>	<b>0.65(0.06)</b>	0.52(0.02)	<b>0.49(0.08)</b>	<b>0.69(0.17)</b>
	iSB(Proposed) ✓	0.79(0.06)	<b>0.64(0.04)</b>	0.69(0.04)	<b>0.67(0.07)</b>	<b>0.63(0.08)</b>	<b>0.50(0.06)</b>	<b>0.65(0.09)</b>

# Conclusion

- We introduced the itemset Sparse Bayes (iSB), which incorporates interaction into sparse Bayes learning.
- We introduced the theorem for effective search from a huge number of combinations.
- Through experiments, we demonstrated the benefits of interaction, high predictive power, interpretability, and Bayesian estimation.