Feature Selection via Mutual Information: New Theoretical Insights

Mario Beraha, Alberto Maria Metelli, Matteo Papini, Andrea Tirinzoni and Marcello Restelli

International Joint Conference on Neural Networks

16 July 2019

POLITECNICO MILANO 1863

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Presentation Plan

1. Introduction
2. Theoretical guarantees on feature selection
3. Algorithm
4. Experiments

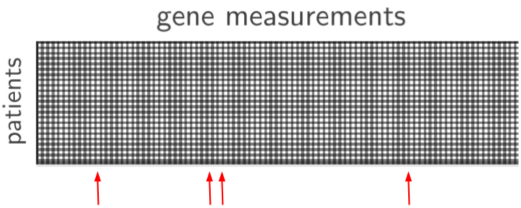
Plan

1. Introduction
2. Theoretical guarantees on feature selection
3. Algorithm
4. Experiments

Why feature selection

Datasets with thousands (or millions) of features have become a standard in Machine Learning tasks.

- Interpretability issues

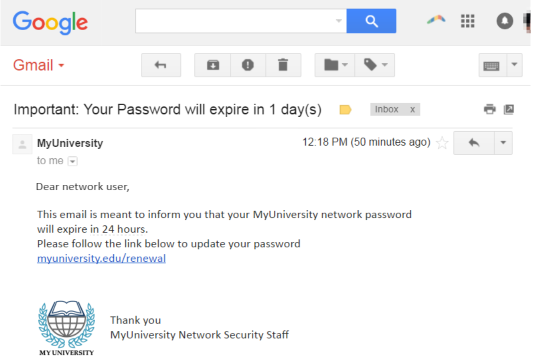


Very few samples
Need to give doctors meaningful results

Why feature selection

Datasets with thousands (or millions) of features have become a standard in Machine Learning tasks.

- Interpretability issues
- Generalization issues



Vocabulary size $\sim 100K$

Very easy to overfit.

Why feature selection

Datasets with thousands (or millions) of features have become a standard in Machine Learning tasks.

- Interpretability issues
- Generalization issues
- Computational issues

“ More data beats clever algorithms, but better data beats more data. ”
Peter Norvig

Feature selection: Main Idea

Feature selection: distinguish between

- Relevant features
- Irrelevant features

$$y = x_1^2 + x_2 + 0 \times x_3 + 3x_4$$

x_1, x_2 and x_4 are relevant
x_3 is irrelevant

Feature selection: Main Idea

Feature selection: distinguish between

- Relevant features
- Irrelevant features
- Redundant features

$$y = x_1^2 + x_2 + 0 \times x_3 + 3x_4$$

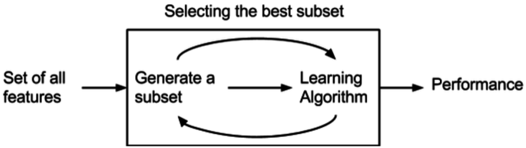
x_1, x_2 and x_4 are relevant
x_3 is irrelevant

If $x_4 = -x1 + 10x_2$
$\rightarrow x_4$ is redundant!

Feature selection at a glance

Wrappers

- Learning as a sub-routine of feature selection algorithm

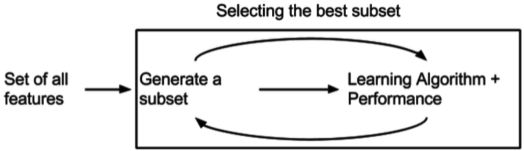


Wrappers

- Learning as a sub-routine of feature selection algorithm

Embedded methods

- Feature selection and learning carried out together
- Example: LASSO regression, Feature Selection for SVMs (Weston et al. 2001)



Feature selection at a glance

Wrappers

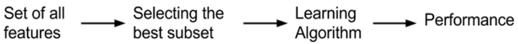
- Learning as a sub-routine of feature selection algorithm

Embedded methods

- Feature selection and learning carried out together
- Example: LASSO regression, Feature Selection for SVMs (Weston et al. 2001)

Filter methods

- No knowledge of the learning algorithm



Mutual Information is a measure of statistical dependence between random variables.

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

Feature selection via Mutual Information

Mutual Information is a measure of statistical dependence between random variables.

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

Conditional Mutual Information $I(X;Y \mid Z)$ used by Brown et al. (2012):
"A feature can be discarded if it is useless for predicting the target or it is predictable from the other features".

$$I(X;Y \mid Z) = \int_Z D_{LK}\left(P_{(X,Y)|Z} \,||\, P_{X|Z} P_{Y|Z}\right) p(z) dz$$

Feature selection via Mutual Information

Mutual Information is a measure of statistical dependence between random variables.

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

Conditional Mutual Information $I(X;Y \mid Z)$ used by Brown et al. (2012):
"A feature can be discarded if it is useless for predicting the target or it is predictable from the other features".

$$I(X;Y \mid Z) = \int_Z D_{LK}\left(P_{(X,Y)|Z} \,||\, P_{X|Z}P_{Y|Z}\right) p(z)dz$$

So far, proposed filter methods based on MI are "empirical" as they do not investigate the relation between the mutual information of a feature set and the prediction error

Plan

1. Introduction

2. Theoretical guarantees on feature selection

3. Algorithm

4. Experiments

Problem statement

Let $\mathcal{X}$ be the space of covariates and $\mathcal{Y}$ the space of response.

$$g : \mathcal{X} \to \mathcal{Y}$$

- A is the index set of features to be removed, $\bar{A}$ its complementary.
- $\mathcal{X}_{\bar{A}} \subset \mathcal{X}$ which includes only the features with indices in $\bar{A}$
- $\mathcal{G}_{\bar{A}} = \{g : \mathcal{X}_{\bar{A}} \to \mathcal{Y}\}$, $\mathcal{G} = \{g : \mathcal{X} \to \mathcal{Y}\}$

We want to bound:

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\boldsymbol{X},Y} \left[L\left(Y, g(\boldsymbol{X}_{\bar{A}})\right)\right]$$

Where L is a suitable loss function.

Under Mean Squared Error Loss, we know that

$$\arg\inf_{g\in\mathcal{G}} \mathbb{E}_{\boldsymbol{X},Y}\left[(Y-g(\boldsymbol{X}))^2\right] = \mathbb{E}[Y \mid \boldsymbol{X}]$$

Theorem 1

$$\inf_{g\in\mathcal{G}_{\bar{A}}} \mathbb{E}_{\boldsymbol{X},Y}\left[(Y-g(\boldsymbol{X}_{\bar{A}}))^2\right] \leq \sigma^2 + 2B^2 I(Y;\boldsymbol{X}_A|\boldsymbol{X}_{\bar{A}}) \tag{1}$$

- $\sigma^2 = \mathbb{E}_{\boldsymbol{X},Y}\left[(Y-\mathbb{E}[Y|\boldsymbol{X}])^2\right]$ is the irreducible error
- B s.t. $\mid Y \mid \leq B$ a.s.

Bound on the regression error - Sketch of the proof

$$\begin{aligned}
\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\boldsymbol{X},Y}\left[(Y - g(\boldsymbol{X}_{\bar{A}}))^2\right] &= \mathbb{E}_{\boldsymbol{X},Y}\left[(Y - \mathbb{E}[Y|\boldsymbol{X}_{\bar{A}}])^2\right] \\
&= \int p(\boldsymbol{x}) \int p(y|\boldsymbol{x}) \left(y - \mathbb{E}[Y|\boldsymbol{x}_{\bar{A}}] \pm \mathbb{E}[Y|\boldsymbol{x}]\right)^2 \mathrm{d}y \mathrm{d}\boldsymbol{x} \\
&= \sigma^2 + \int p(\boldsymbol{x}) \left(\mathbb{E}[Y|\boldsymbol{x}] - \mathbb{E}[Y|\boldsymbol{x}_{\bar{A}}]\right)^2 \mathrm{d}\boldsymbol{x} \\
&= \sigma^2 + \int p(\boldsymbol{x}) \left(\int y \left(p(y|\boldsymbol{x}) - p(y|\boldsymbol{x}_{\bar{A}})\right) \mathrm{d}y\right)^2 \mathrm{d}\boldsymbol{x} \\
&\leq \sigma^2 + B^2 \int p(\boldsymbol{x}) \left(\int |p(y|\boldsymbol{x}) - p(y|\boldsymbol{x}_{\bar{A}})| \,\mathrm{d}y\right)^2 \mathrm{d}\boldsymbol{x} \\
&\leq \sigma^2 + 2B^2 \int p(\boldsymbol{x}) D_{\mathrm{KL}}\left(p(\cdot|\boldsymbol{x}) \| p(\cdot|\boldsymbol{x}_{\bar{A}})\right) \mathrm{d}\boldsymbol{x} \\
&= \sigma^2 + 2B^2 I(Y; \boldsymbol{X}_A | \boldsymbol{X}_{\bar{A}}).
\end{aligned}$$

Bound on the classification error

Theorem 2

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\boldsymbol{X},Y}\left[\mathbb{1}_{\{Y \neq g(\boldsymbol{X}_{\bar{A}})\}}\right] \leq \epsilon + \sqrt{2I(Y;\boldsymbol{X}_A|\boldsymbol{X}_{\bar{A}})} \quad (2)$$

- $\epsilon = \mathbb{E}_{\boldsymbol{X},Y}\left[\mathbb{1}_{\{Y \neq \arg\max_{y \in \mathcal{Y}} p(y|\boldsymbol{X})\}}\right]$ is the Bayes error

Plan

1. Introduction

2. Theoretical guarantees on feature selection

3. Algorithm

4. Experiments

Backward elimination

- Select a threshold $\delta \geq 0$, the maximum error that the filter is allowed to introduce.
- Start with the full feature set
- At each step remove the feature that minimizes the

$$I(Y; X_i | \boldsymbol{X}_{\bar{A}_t} \setminus X_i)$$

- Stop as soon as $\sum_{h=1}^{t} I_h \geq \frac{\delta}{2B^2}$ for regression and $\sum_{h=1}^{t} I_h \geq \frac{\delta^2}{2}$ for classification.

Forward selection

In a similar fashion, we can define a forward search algorithm

- Start with no features
- At each step, look for the feature that maximizes the $I(Y; X_i | \boldsymbol{X}_{A_t})$
- Stop as soon as a threshold is met

Theorem 3

Backward elimination achieves an error of $\sigma^2 + \delta$ for regression, where σ^2 is the irreducible error and $\epsilon + \delta$ for classification, where ϵ is the Bayes error.

Theorem 4

Forward selection achieves an error of $\sigma^2 - \delta + 2B^2 I(Y; \boldsymbol{X})$ for regression, where σ^2 is the irreducible error and $\epsilon - \delta + \sqrt{2I(Y; \boldsymbol{X})}$ for classification, where ϵ is the Bayes error.

The proofs are based on recursively applying the equality

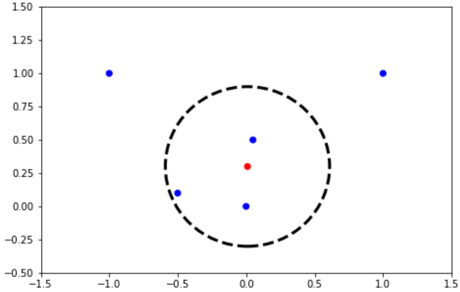
$$I(Y; X \mid Z) = I(Y; X, Z) + I(Y; Z)$$

Mutual Information can be written in the form

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Problem when X is continuous and Y is discrete (classification).
We resort to the KSG estimator (Kraskov et al. 2004)

$$I(X,Y) = \psi(k) + \psi(N) - \mathbb{E}\left[\psi(n_x+1) + \psi(n_y+1)\right]$$



Plan

1. Introduction

2. Theoretical guarantees on feature selection

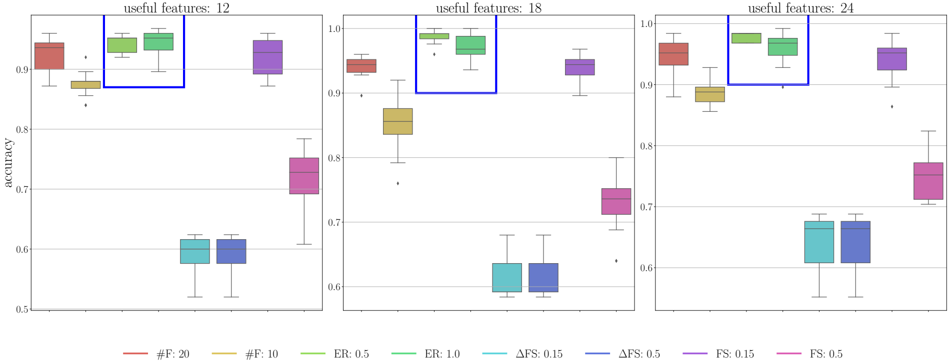
3. Algorithm

4. Experiments

Synthetic Experiments

$X_1, \dots X_{500} \in \mathbb{R}^{30}$

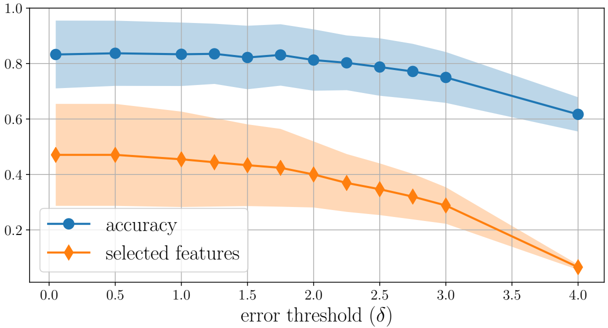
- 30 features, only k useful (fixed)
- $X_1, \dots X_k \mid Y = 1 \sim N_k(0, 1)$
- $X_1, \dots X_k \mid Y = 0 \sim N_k(0, 1) \mid \sum_{l=1}^{k} X_i > 3(k-2)$



Synthetic Experiments

$X_1, \dots X_{500} \in \mathbb{R}^{15}$

- 30 features, only k useful, $k \sim \mathcal{U}(3, 15)$.
- $X_1, \dots X_k \mid Y = 1 \sim N_k(0, 1)$
- $X_1, \dots X_k \mid Y = 0 \sim N_k(0, 1) \mid \sum_{l=1}^{k} X_i > 3(k-2)$



Dataset	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 1.0$
ORL	0.8	0.75	0.7	0.7375	0.7125
warpAR10P	0.97	0.98	0.98	0.98	0.98
glass*	0.99	0.99	0.99	0.99	0.99
wine	0.96	0.96	0.96	0.95	0.83
ALLAML	1.0	1.0	1.0	0.92	0.78

*: no feature removed

- New stopping condition on Mutual Information based filter feature selection
- Theoretical guarantees on the introduced error
- Less sensitive to hyperparameters

Feature work

- How to parallelize the backward elimination?
- Faster (approximate) CMI estimation in high dimension?
- How to leverage information about the learning algorithm ?