Causal Feature Selection and its Applications in Biomedical and Health Data Science

Sisi Ma

University of Minnesota

sisima@umn.edu

November 3, 2018

My co-organizer

Constantin Aliferis, Director of IHI at UMN

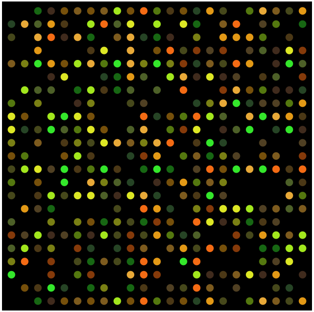
Alexander Statnikov, Global Head of Data Science and Machine Learning, SoFi

Where to find material related to this tutorial?

`https://github.com/SisiMa1729/Causal_Feature_Selection`

Overview

What is Predictive Modeling?

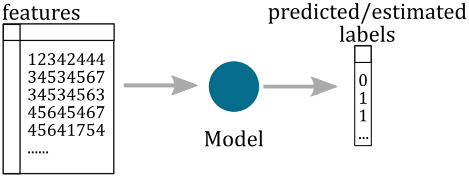


Predictive Modeling uses statistics to predict outcome [Geisser, 1993].

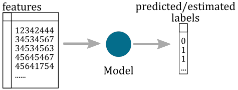
Applications in Biomedical and Health Sciences

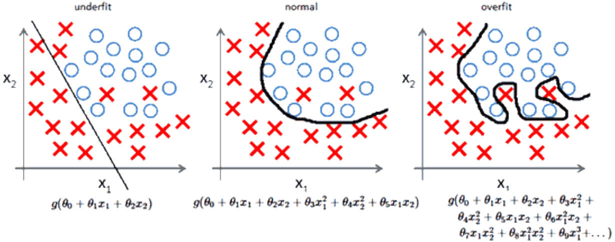
- Diagnoistics: Determine subtypes of breast cancer [Perou et al., 2000]
- Prognostics: Predicting time-to-death/time-to-relapse for cancer patients [Ray et al., 2014]
- Risk Assessment: Predicting the risk of PTSD after trauma [Saxe et al., 2017, Galatzer-Levy et al., 2017]

Basic Concepts: Model



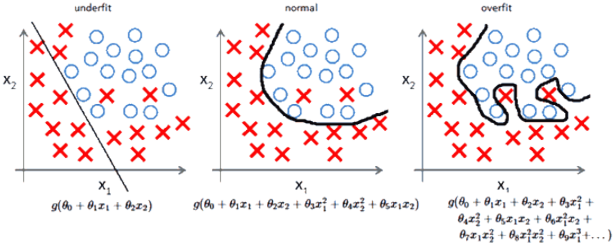
Basic Concepts: Model





* bottom figure from Andrew Ng's ML course

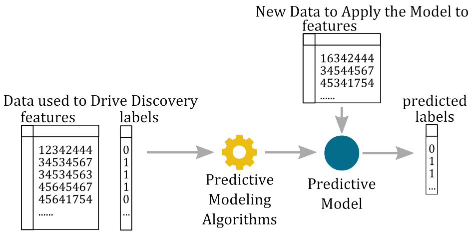
Basic Concepts: Model



"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." – von Neumann

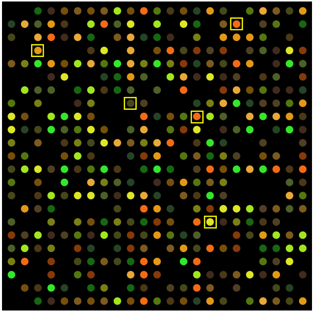
Basic Concepts: Model

How do evaluate the quality of my model?



Basic Concepts: Feature Selection

Features: The variables that the predictive model can use as inputs to make predictions.
Featuture Selection: selecting a subset of features from all the available features for constructing the predictive model.



Basic Concepts: Feature Selection

Featuture Selection: selecting a subset of features from all the available features for constructing the predictive model.

Q: How many ways one can select subset of features out of N total number of features?

Basic Concepts: Feature Selection

Goal of Feature Selection:

- Improve model predictivity
- Enhance model interpretability
- Increase cost-efficiency for model deployment

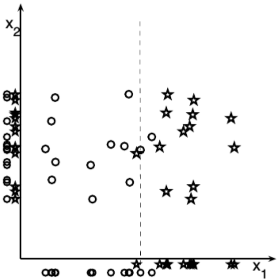
Feature Selection is Highly Relavent for Modeling Biomedical and Health Data

Goal of Feature Selection:

- Improve model predictivity: Stakes are high in health sciences
- Enhance model interpretability: Black-box models are generally not acceptable
- Increase cost-efficiency for model deployment: Data are expensive to collect in this domain

Overview

Selection based on predictivity of individual features

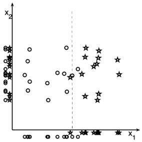


- An intuitive and popular stratergy: to select features that are univariate associated with the target of interest.

Selection based on predictivity of individual features

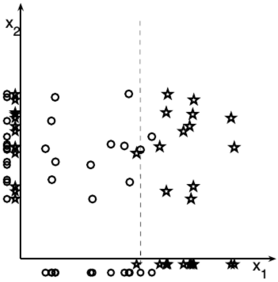
- Individual feature irrelevance: The feature X_i is individually irrelevant to the target T iff X_i is independent of T (denoted as $X_i \perp T$): $P(X_i, Y) = P(X_i)P(T)$
- Individual feature relevance: All featuares that are not individually irrelevant are individually relevant
- e.g. P(flu=True)=0.2, P(vaccine=True)=0.8, P(vaccine=True,flu=True)=0.05. Does knowing someone's vaccination status help us predict if they would get the flu?

Selection based on predictivity of individual features



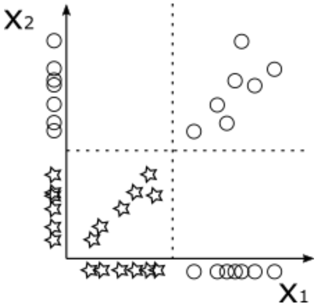
- Popular tests for continuous predictors, binary target: student t, Mann–Whitney U, Kolmogorov–Smirnov
- Popular tests for discrete predictors, binary target: χ^2, G^2, fisher exact
- Note: multiple comparisons inflate family-wise error (type I error). This may or may not be a problem for predictive modeling...

Selection based on predictivity of individual features



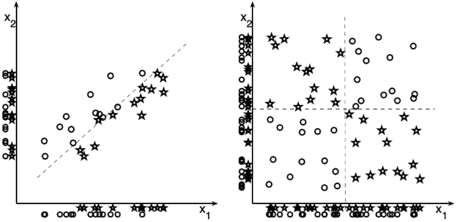
- pro: computationally efficient, good intepretation
- con: rendundancy among selected features, misses interactions among features

Selection based on predictivity of individual features



Feature X_1 and X_2 could predict the target equally well by themselves. However, feature selection based on predictivity of individual features will typically select both, resulting in redundancy.

Selection based on predictivity of individual features



Features that are useless for prediction by themselves can be useful when combined with other features (i.e. interactions among features could be predictive. Selection based on predictivity/relevance of individual features may miss useful interactions.).

Selection based on combination of features

It is critical to consider combination of features and the combined information in these features with respect to the target of interest.

Selection based on combination of features

What is a good combination of features?

Optimal feature set: Given a data set $\mathbb{D}$ (a sample from distribution $\mathbb{P}$) for variables $\mathbf{V}$, a learning algorithm $\mathbb{L}$, and a performance metric $\mathbb{M}$, a feature set $\mathbf{X} \subseteq \mathbf{V}$ is an optimal feature set of T if $\mathbf{X}$ maximizes the performance metric $\mathbb{M}$ for predicting T using learner $\mathbb{L}$ in the dataset $\mathbb{D}$.
[Statnikov et al., 2013, Kohavi and John, 1997]

Selection based on combination of features

How to measure combined information in features?

Let $\mathbf{S_i} = \mathbf{V} - V_i$, the set of all features except V_i
Strong relevance: A feature V_i is strongly relevant to T iff there exists some v_i, t, and s_i for which $p(V_i = v_i, \mathbf{S_i} = s_i) > 0$, such that $p(T = t | V_i = v_i, \mathbf{S_i} = s_i) \neq p(T = t | \mathbf{S_i} = s_i)$.
[Kohavi and John, 1997]

Selection based on combination of features

How to measure combined information in features?

Strongly relevant features are conditionally dependent of T given all other measured variables. In other words, knowing the value of features V_i gives additional information regarding the target of interest, even when we know the value of all other features.
[Kohavi and John, 1997]

Selection based on combination of features

How to measure combined information in features?

Let $\mathbf{S_i} = \mathbf{V} - V_i$, the set of all features except V_i
Weak relevance: A feature V_i is weakly relevant to T iff it is not strongly relevant, and $\exists S_i' \subset S_i$ and some v_i, t, and s_i' for which $p(V_i = v_i, \mathbf{S_i'} = s_i') > 0$, such that $p(T = t | X_i = v_i, S_i' = s_i') \neq p(T = t | S_i' = s_i')$.
[Kohavi and John, 1997]

Selection based on combination of features

How to measure combined information in features?

Weakly relevant features are not conditionally dependent of T given all other measured variables, but are conditionally dependent of T given at least one subset of all other measured variables.

In other words, knowing the value of features V_i do not give additional information regarding the target of interest, when we know the value of all other features; however if we only know the value of a subset of all other measured variables, knowing the value of V_i may provide additional information.

[Kohavi and John, 1997]

Selection based on combination of features

How to measure combined information in features?

Irrelevance:A feature V_i is relevant if it is either weakly relevant or strongly relevant; otherwise, it is irrelevant.

Irrelevant features neither provide information regarding the target by itself nor in combination with any other features.

[Kohavi and John, 1997]

Selection based on combination of features

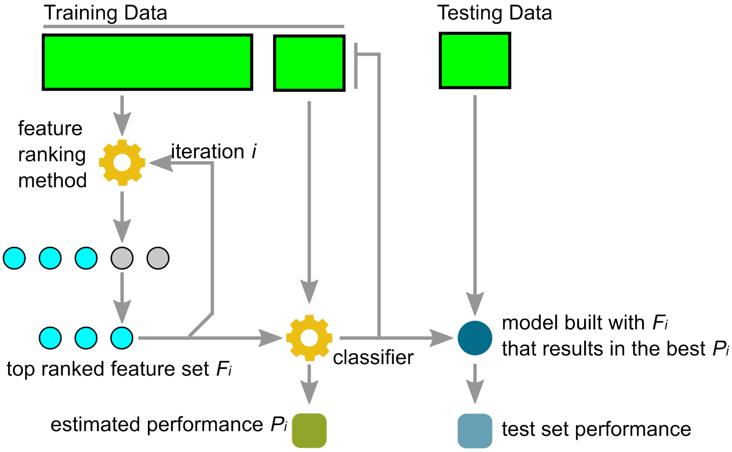
Given the definition of strong, weak and irrelevant variables, what is our stratergy?

- Strongly relevant features should always be selected. Omitting Strongly relevant features will compromise predictive performance.
- Weakly relevant features do not improve predictive performance if all strongly relevant features are selected.
- Irrelevant features do not improve predictive performance.

[Kohavi and John, 1997]

Selection based on combination of features

Recursive Feature Elimination [Guyon et al., 2002]:



Selection based on combination of features

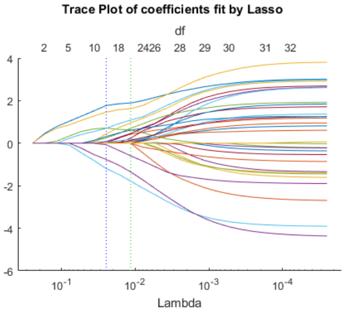
Lasso Regularization [Tibshirani, 1996]

Loss function for logistic regression:

$J(\vec{\beta}) = -\sum_{i=1}^{n} y^{(i)} log(f(\vec{x^{(i)}})) + (1 - y^{(i)}) log(1 - f(\vec{x^{(i)}}))$

Loss function for logistic regression with Lasso regularization:

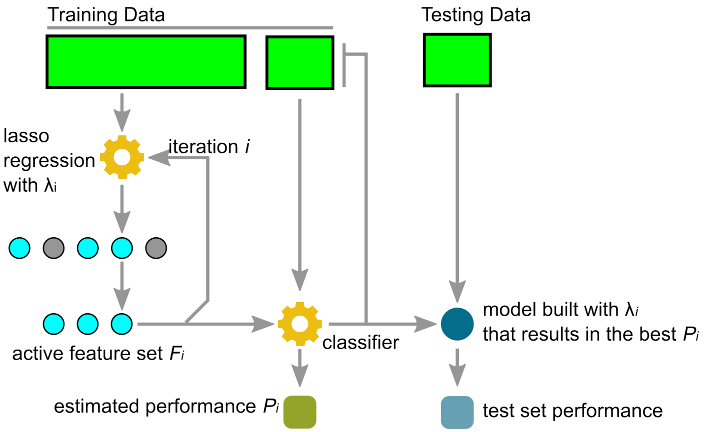
$J_{lasso}(\vec{\beta}) = J(\vec{\beta}) + \lambda \Sigma_p |\beta_p|$



* figure taken from https://www.mathworks.com/help/stats/regularize-logistic-regression.html

Selection based on combination of features

Lasso Regularization: finding λ value



* selection of λ is generally done with cross-validation on the training data. A simplified version is shown for illustrative purpose.

Optimizing Predictivity with Constraint

Generally speaking, non-causal feature selection methods, such as RFE and lasso, selects features that optimize predictive performance while minimizing number of features being selected.

Optimal feature set?

Are features selected by RFE and Lasso likely to be an optimal feature set?

Optimal feature set: Given a data set $\mathbb{D}$ (a sample from distribution $\mathbb{P}$) for variables $\mathbf{V}$, a learning algorithm $\mathbb{L}$, and a performance metric $\mathbb{M}$, a feature set $\mathbf{X} \subseteq \mathbf{V}$ is an optimal feature set of T if $\mathbf{X}$ maximizes the performance metric $\mathbb{M}$ for predicting T using learner $\mathbb{L}$ in the dataset $\mathbb{D}$.

Feature Relevance?

These methods do not directly utilize the concepts of different types of feature relevance, but in practice they produces models with excellent predictive performances in various application domains.

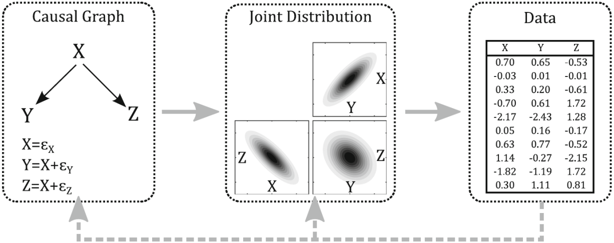
Overview

Causal Feature Selection

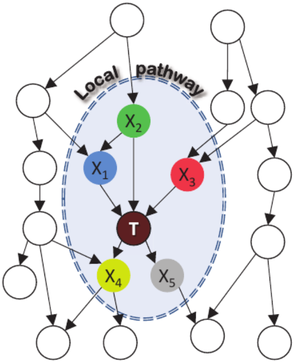
Causal feature selection methods considers the joint information in a set of features by accessing conditional independence relationships, which are directly related to the concepts of feature relevance.

Causal Structure/Causality

In this context, when we say causality, we focus on the "how the data is generated".

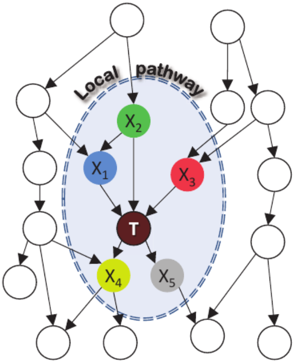


Relationship between Causality and Feature Selection



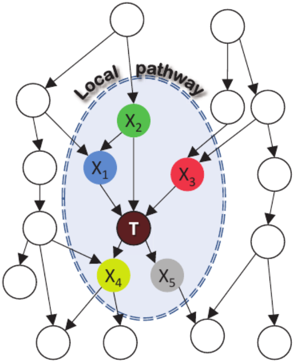
From the data generation process, it is easy to see why some variables are redundent for the predictive task.

Relationship between Causality and Feature Selection



The Markov Boundary (direct causes + direct effects + direct causes of direct effects) is the minimal feature set that contain all information regarding the target of interest.

Relationship between Causality and Feature Selection



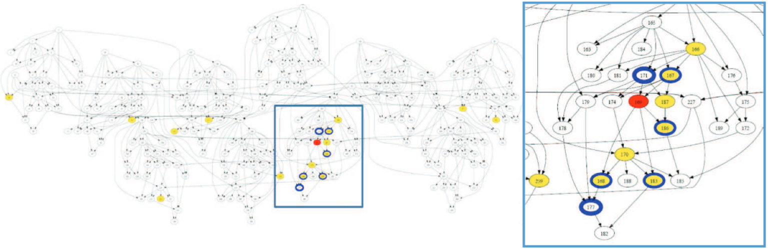
In a faithful distribution, the features that constitutes the Markov Boundary of a target T are all the strongly relevant features for T.

Benefit of Causal Feature Selection

- selected features have causal interpretations
- selected feature set is generally smaller compared to non-causal methods
- model generalize better under certain type of distribution shift

Benefit of Causal Feature Selection

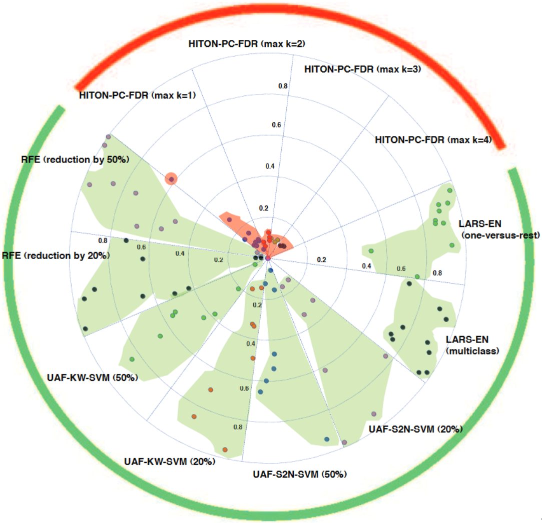
Example: selected features have causal interpretations.



[Aliferis et al., 2010a]

Benefit of Causal Feature Selection

Example: selected features have causal interpretations.



Benefit of Causal Feature Selection

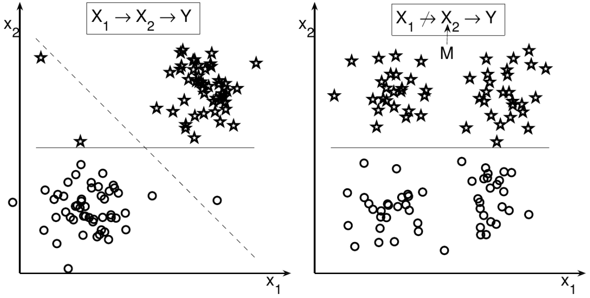
Example: selected feature set is generally smaller compared to non-causal methods, without compromising predictive performance.

	Predicitivity		*Reduction*	
Feature selection method	P-value	Nominal winner	P-value	Nominal winner
No feature selection	0.1890	Other	**<0.0001**	HITON-PC
RFE: 4 variants	0.9754	Other	**0.0046**	HITON-PC
	0.8030	Other	**0.0042**	HITON-PC
	0.1312	HITON-PC	0.3634	HITON-PC
	0.1008	HITON-PC	0.6816	Other
UAF-KruskalWallis-SVM: 4 variants	0.2248	Other	**0.0028**	HITON-PC
	0.0098	Other	**0.0004**	HITON-PC
	1.0000	HITON-PC	0.1414	HITON-PC
	0.3232	HITON-PC	0.3998	HITON-PC
UAF-Signal2Noise-SVM: 4 variants	0.0710	Other	**0.0018**	HITON-PC
	0.0752	Other	**0.0030**	HITON-PC
	0.4420	HITON-PC	0.7850	HITON-PC
	0.2820	HITON-PC	0.6604	HITON-PC
UAF-Neal-SVM: 4 variants	0.5046	Other	**<0.0001**	HITON-PC
	0.9782	HITON-PC	**<0.0001**	HITON-PC
	0.6980	HITON-PC	**0.0044**	HITON-PC
	0.3806	HITON-PC	**0.0186**	HITON-PC
Random Forest Variable Selection: 2 variants	0.6064	HITON-PC	0.3252	HITON-PC
	0.5050	HITON-PC	0.1338	Other
LARS-Elastic Net: 2 variants	1.0000	Other	0.1112	HITON-PC
	0.0832	HITON-PC	0.5216	Other

[Aliferis et al., 2010a]

Benefit of Causal Feature Selection

Example: Model built with causal feature selection methods generalize better under certain distribution shifts.



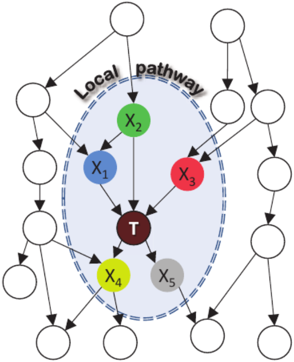
(a) No manipulation. (b) Manipulation of X_2.

[Guyon et al., 2007]

Benefit of Causal Feature Selection

- selected features have causal interpretations
- selected feature set is generally smaller compared to non-causal methods
- model generalize better under certain type of distribution shift

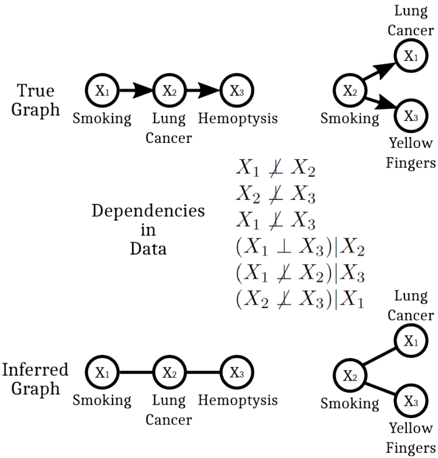
How to find MB or Strongly relevant features?



In a faithful distribution, the features that constitutes the Markov Boundary of a target T are all the strongly relevant features for T.

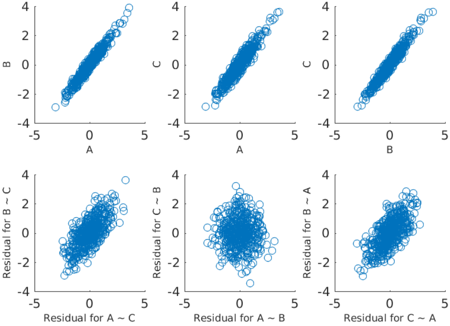
How to find MB or Strongly relevant features?

The concept of conditional independence:



How to find MB or Strongly relevant features?

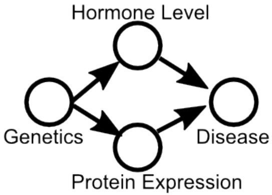
Conditional independence, What does it look like:
$A->B->C$; $A<-B<-C$; or $A<-B->C$



$A \not\perp B, A \not\perp C, B \not\perp C$
$A \not\perp B|C, A \perp C|B, B \not\perp C|A$

How to find MB or Strongly relevant features?

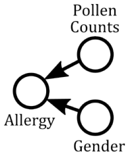
The concept of conditional independence:



How to find MB or Strongly relevant features?

The concept of conditional independence:

True Graph



Denpendencies in Data

$PollenCounts \perp Gender$

$PollenCounts \not\perp Allergy$

$Gender \not\perp Allergy$

$(PollenCounts \not\perp Gender)|Allergy$

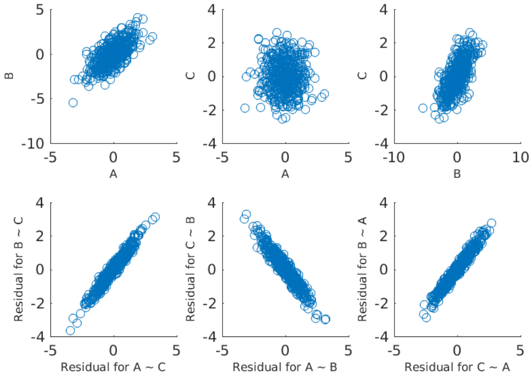
$(PollenCounts \not\perp Allergy)|Gender$

$(Gender \not\perp Allergy)|PollenCounts$

How to find MB or Strongly relevant features?

Conditional independence, What does it look like:

$A- > B < -C$



$A \not\perp B, A \perp C, B \not\perp C$

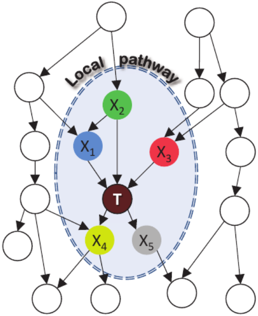
$A \not\perp B|C, A \not\perp C|B, B \not\perp C|A$

How to find MB or Strongly relevant features?

Conditional Independence Tests:

- fisher's test
- χ^2, g^2
- conditional mutual information, distance correlation
- comparison of nested models

Markov Boundary



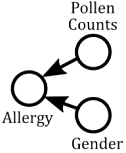
The Markov Boundary (direct causes + direct effects + direct causes of direct effects) is the minimal feature set that contain all information regarding the target of interest.

The target is conditionally independent with any other variables given MB(T).

Markov Boundary

Why is the direct causes of direct effects part of the Markov Boundary?

True Graph



Denpendencies in Data

$PollenCounts \perp Gender$

$PollenCounts \not\perp Allergy$

$Gender \not\perp Allergy$

$(PollenCounts \not\perp Gender)|Allergy$

$(PollenCounts \not\perp Allergy)|Gender$

$(Gender \not\perp Allergy)|PollenCounts$

Markov Boundary

There are a number of algorithms that discovery the local causal neibourhood of the data.

- PC-simple[Bühlmann et al., 2010]
- HITON-PC[Aliferis et al., 2003, Aliferis et al., 2010a, Aliferis et al., 2010b]
- HITON-MB[Aliferis et al., 2010a, Aliferis et al., 2010b]
- MMPC [Tsamardinos et al., 2006]
- IAMB[Tsamardinos et al., 2003]
- MBFS[Ramsey, 2006]

The PC simple algorithm

Algorithm 1 The population version of the PC-simple algorithm.

1: Set $m = 1$. Do correlation screening, and build the step$_1$ active set $\mathcal{A}^{[1]} = \{j = 1, \ldots, p;\ \mathrm{cor}(Y, X^{(j)}) \neq 0\}$ as in (5).

2: **repeat**

3: $m = m + 1$. Construct the step$_m$ active set:

$$\mathcal{A}^{[m]} = \{j \in \mathcal{A}^{[m-1]};\ \rho(Y, X^{(j)} \mid X^{(\mathcal{S})}) \neq 0 \\ \text{for all } \mathcal{S} \subseteq \mathcal{A}^{[m-1]} \setminus \{j\} \text{ with } |\mathcal{S}| = m - 1\}.$$

4: **until** $|\mathcal{A}^{[m]}| \leq m$.

[Bühlmann et al., 2010]

The PC simple algorithm: trace

Algorithm 1 The population version of the PC-simple algorithm.

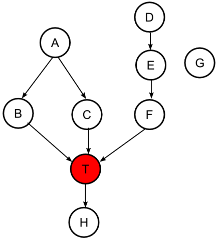
1: Set $m = 1$. Do correlation screening, and build the step$_1$ active set $\mathcal{A}^{[1]} = \{j = 1, \ldots, p;\ \mathrm{cor}(Y, X^{(j)}) \neq 0\}$ as in (5).

2: **repeat**

3: $m = m + 1$. Construct the step$_m$ active set:

$$\mathcal{A}^{[m]} = \{j \in \mathcal{A}^{[m-1]};\ \rho(Y, X^{(j)} \mid X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \subseteq \mathcal{A}^{[m-1]} \setminus \{j\} \text{ with } |\mathcal{S}| = m - 1\}.$$

4: **until** $|\mathcal{A}^{[m]}| \leq m$.

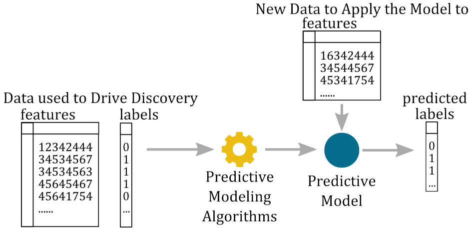


$m = 1; A^{[1]} = \{A, B, C, D, E, F, H\}$

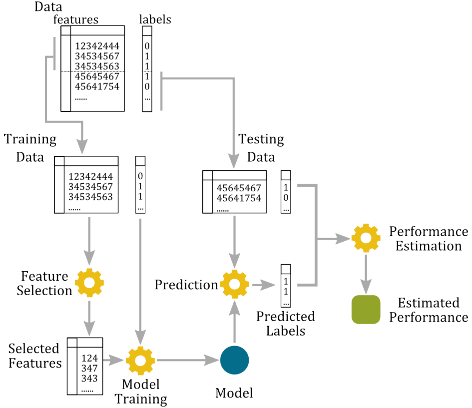
$m = 2; A^{[2]} = \{A, B, C, F, H\}$

$m = 3; A^{[3]} = \{B, C, F, H\}$

Demo: Design


New Data to Apply the Model to
features
16342444
34544567
45341754
......
Data used to Drive Discovery
features
labels
12342444
34534567
34534563
45645467
45641754
......
0
1
1
1
0
...
Predictive Modeling Algorithms
Predictive Model
predicted labels
0
1
1
...

Demo: Design


Data
features
labels
12342444
34534567
34534563
45645467
45641754
......
0
1
1
1
0
...
Training
Data
12342444
34534567
34534563
......
0
1
1
...
Testing
Data
45645467
45641754
......
1
0
...
Feature
Selection
Prediction
1
1
...
Predicted
Labels
Performance
Estimation
Estimated
Performance
Selected
Features
124
347
343
......
Model
Training
Model

Demo1: Classifying Cancer Tissue vs. Normal

Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses

Arindam Bhattacharjee[*,†], William G. Richards[‡,§], Jane Staunton[†,¶], Cheng Li[‖], Stefano Monti[¶], Priya Vasa[*], Christine Ladd[¶], Javad Beheshti[*], Raphael Bueno[‡], Michael Gillette[¶], Massimo Loda[*,**], Griffin Weber[*], Eugene J. Mark[††], Eric S. Lander[¶], Wing Wong[‖], Bruce E. Johnson[*], Todd R. Golub[¶,‡‡,§§,¶¶], David J. Sugarbaker[‡,§,¶¶], and Matthew Meyerson[*,§§,¶¶]

156 samples: Adenocarcinoma(139) vs Normal(17)
12601 features: represent different probe on microarray chips
[Bhattacharjee et al., 2001]

Demo1: SVM-RFE

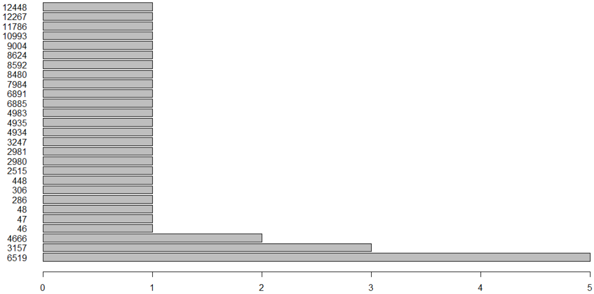

```
library(e1071)
library(caret)
library(pROC)
source('svm_rfe.R')
```

Demo1: SVM-RFE


```
folds<-createFolds(data[,1], k = 5)
for (f in 1:5){
  ...
  features[[f]] <-svm_rfe(x,y,
                          train_train_idx,
                          train_test_idx,1.2)
  mod <- svm(x[train_idx,features[[f]]],y[train_idx],
             kernel = "linear",cost=1,
             scale=T, probability=T)
  pred<-predict(mod,
            newdata=x[test_idx,features[[f]]],
           probability=T,decision.values=T)
  perf[f]<-roc( ...)
  ...
}
```

Demo1: SVM-RFE

AUC: 0.977
number of features: 6.8



Demo1: lasso


```
library(pROC)
library(caret)
library(glmnet)
```

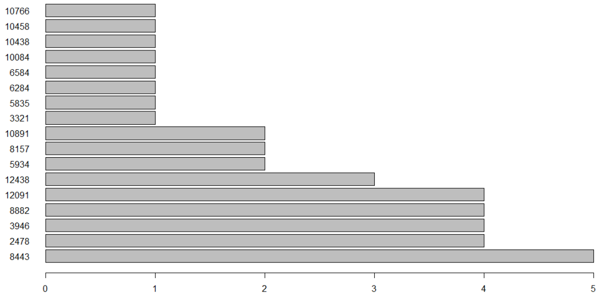
Demo1: lasso

```
folds<-createFolds(data[,1], k = 5)
for (f in 1:5){
  ...
  cv_mod <-cv.glmnet(as.matrix(x[train_idx,]),
                y[train_idx],
                family="binomial",nfolds=4)
  features[[f]]<-setdiff(which(coef(cv_mod)!=0),1)-1
      # the first coef correspond to intercept
  pred<-predict(cv_mod,
           newx=as.matrix(x[test_idx,]),
           type="response")
           # alternative setting s="lambda.min"
  perf[f]<-roc(....)
  ...
}
```

Demo1: lasso

AUC: 0.991

number of features: 7.6



Demo1: pc-simple


```
library(e1071)
library(caret)
source('fs_pcsimple.R')
    library(pcalg)
```

Demo1: pc-simple

```
folds<-createFolds(data[,1], k = 5)
for (f in 1:5){
  ...
  features[[f]] <-fs_pcsimple(x,y, train_idx)
  mod <- svm(x[train_idx,features[[f]]],y[train_idx],
        kernel = "linear",
        cost=1,
        scale=T,
        probability=T)
  pred<-predict(mod,
           newdata=x[test_idx,features[[f]]],
           probability=T,
           decision.values=T)
  perf[f]<-roc(...)
  ...
}
```

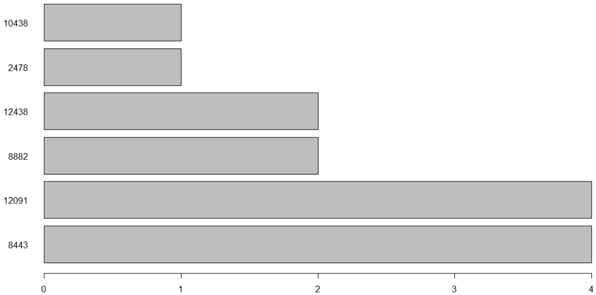
Demo1: pc-simple


```
fs_pcsimple <- function(x,y, train_idx){
  library(pcalg)
  tr_x<-x[train_idx,]
  tr_y<-y[train_idx]
  slct<-pcSelect(as.numeric(tr_y),
           as.matrix(tr_x), 0.001,
           corMethod = "standard",
           verbose = FALSE, directed = FALSE)
  features<-which(slct$G)
  return(features)
}
```

Demo1: pc-simple

AUC: 0.991

number of features: 2.8



Demo1.5: PTSD prognosis

Saxe *et al. BMC Psychiatry* (2017) 17:223
DOI 10.1186/s12888-017-1384-1

BMC Psychiatry

RESEARCH ARTICLE Open Access

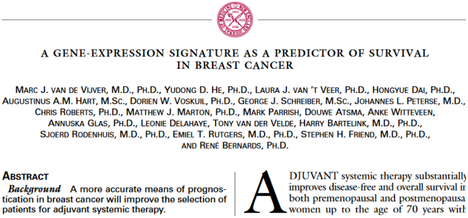
CrossMark

Machine learning methods to predict child posttraumatic stress: a proof of concept study

Glenn N. Saxe[1*], Sisi Ma[2], Jiwen Ren[3] and Constantin Aliferis[4,5]

[Saxe et al., 2017, Aliferis et al., 2010a, Aliferis et al., 2010b]

Demo2: Risk Model for Metastasis



A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D., AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.SC., JOHANNES L. PETERSE, M.D., CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE ATSMA, ANKE WITTEVEEN, ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D., SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D., AND RENÉ BERNARDS, PH.D.

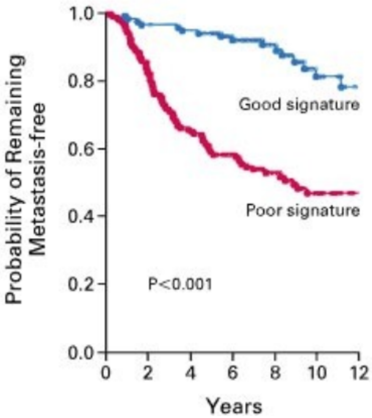
ABSTRACT

Background A more accurate means of prognostication in breast cancer will improve the selection of patients for adjuvant systemic therapy.

ADJUVANT systemic therapy substantially improves disease-free and overall survival in both premenopausal and postmenopausal women up to the age of 70 years with

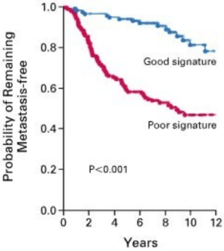
295 samples: with time to metastasis outcome
metastasis (101) vs no metastasis (194)
70 features: gene expression levels of 70 genes determined to be important for prognostication for breast cancer.
[Van De Vijver et al., 2002]
http://ccb.nki.nl/data/

Demo2: Time to Event Outcome



Survival models relate the time that passes, before some event occurs, to one or more covariates that may be associated with that quantity of time.
[Cox, 1992, Harrell, 2015, Therneau and Grambsch, 2013]

Demo2: Time to Event Outcome



Cencoring is common in time to event data and requires special treatment.
[Cox, 1992, Harrell, 2015, Therneau and Grambsch, 2013]

Demo2: lasso

```
library(survival)
library(caret)
library(glmnet)
library(cancerdata)
```

Demo2: lasso


```
data('VIJVER')
idx<-is.element(rownames(VIJVER@assayData$exprs)
          ,gene)
x<-as.data.frame(t(VIJVER@assayData$exprs[idx,]))
y<-Surv(VIJVER$Follow_up_time_or_metastasis,
      VIJVER$event_metastasis)
```

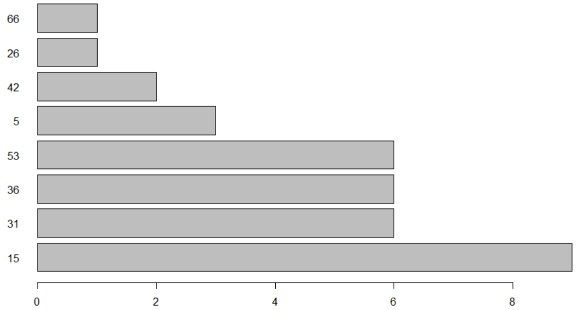
Demo2: lasso

```
folds<-createFolds(VIJVER$event_metastasis, k = 10)
for (f in 1:10){
  ...
  cv_mod <-cv.glmnet(as.matrix(x[train_idx,]),
                    y[train_idx],
                    family="cox",nfolds=9)
  features[[f]]<-which(coef(cv_mod)!=0)
  pred<-predict(cv_mod,
          newx=as.matrix(x[test_idx,,drop=FALSE]))
          # alternative setting s="lambda.min"
  perf[f]<-survConcordance(y[test_idx]~pred)
           $concordance
}
```

Demo2: lasso

c-idx: 0.654

number of features: 3.4



Demo2: pc-simple


```
library(survival)
library(caret)
library(pcalg)
library(cancerdata)
source('pcsimple_survival.R')
```

Demo2: pc-simple

```
folds<-createFolds(data[,1], k = 10)
for (f in 1:10){
  ...
  features[[f]]<-which(
                  pcsimple_survival(y[train_idx,],
                  x[train_idx,], 0.05)$G)
  cdata<-cbind(y,x[,features[[f]],drop=FALSE])
                           [train_idx,]
  cfit<-coxph(y~.,data=cdata)
  pred<-predict(cfit,newdata=
              x[test_idx,features[[f]],drop=FALSE])
  perf[f]<-survConcordance(y[test_idx]~pred)
                                $concordance
}
```

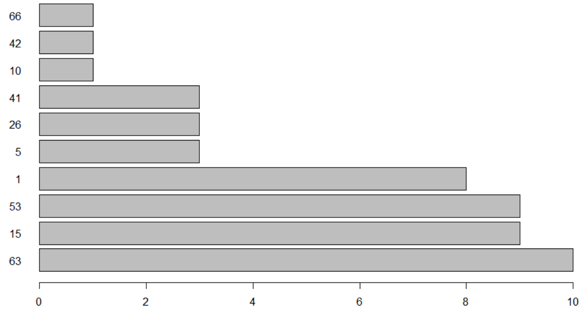
Demo2: pc-simple modified for survival


```
pcsimple_survival<-function (y, dm, alpha,
    verbose = FALSE, directed = FALSE) {
      ...
     # swaping fisher's z test for cox model
     # to accomodate survival outcome
     # z <- zStat(x, y, nbrs[S], C, n)
     cdata<-dm[,c(y,x,nbrs[S])]
     cfit<-coxph(y~.,data=cdata)
     pval<-summary(cfit)$coef[1,5]
    # alternative: Chi-square test to compare
    # nested models
...
}
```

Demo2: pc-simple

AUC: 0.682

number of features: 4.8



Demo2: Time to Event Outcome

For more information regarding feature selection for time to event outcome. [Lagani and Tsamardinos, 2010]

References I

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010a).
Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation.
Journal of Machine Learning Research, 11(Jan):171–234.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010b).
Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions.
Journal of Machine Learning Research, 11(Jan):235–284.

References II

- Aliferis, C. F., Tsamardinos, I., and Statnikov, A. (2003). Hiton: a novel markov blanket algorithm for optimal variable selection. In *AMIA Annual Symposium Proceedings*, volume 2003, page 21. American Medical Informatics Association.
- Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795.

References III

Bühlmann, P., Kalisch, M., and Maathuis, M. H. (2010).
Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm.
Biometrika, 97(2):261–278.

Cox, D. R. (1992).
Regression models and life-tables.
In *Breakthroughs in statistics*, pages 527–541. Springer.

Galatzer-Levy, I., Ma, S., Statnikov, A., Yehuda, R., and Shalev, A. (2017).
Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting ptsd.
Translational psychiatry, 7(3):e1070.

References IV

Geisser, S. (1993).
Predictive inference, volume 55.
CRC press.

Guyon, I., Aliferis, C., and Elisseeff, A. (2007).
Causal feature selection.
Computational methods of feature selection, pages 63–82.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002).
Gene selection for cancer classification using support vector machines.
Machine learning, 46(1-3):389–422.

References V

Harrell, F. E. (2015).
Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.
Springer.
Google-Books-ID: 94RgCgAAQBAJ.

Kohavi, R. and John, G. H. (1997).
Wrappers for feature subset selection.
Artificial intelligence, 97(1-2):273–324.

Lagani, V. and Tsamardinos, I. (2010).
Structure-based variable selection for survival data.
Bioinformatics, 26(15):1887–1894.

References VI

Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M., et al. (2000).
Molecular portraits of human breast tumours.
Nature, 406(6797):747.

Ramsey, J. (2006).
A pc-style markov blanket search for high dimensional datasets.
Technical Report.

References VII

Ray, B., Henaff, M., Ma, S., Efstathiadis, E., Peskin, E. R., Picone, M., Poli, T., Aliferis, C. F., and Statnikov, A. (2014).
Information content and analysis methods for multi-modal high-throughput biomedical data.
Scientific reports, 4.

Saxe, G. N., Ma, S., Ren, J., and Aliferis, C. (2017).
Machine learning methods to predict child posttraumatic stress: a proof of concept study.
BMC psychiatry, 17(1):223.

References VIII

Statnikov, A., Lytkin, N. I., Lemeire, J., and Aliferis, C. F. (2013).
Algorithms for discovery of multiple markov boundaries.
Journal of Machine Learning Research, 14(Feb):499–566.

Therneau, T. M. and Grambsch, P. M. (2013).
Modeling survival data: extending the Cox model.
Springer Science & Business Media.

Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

References IX

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., and Statnikov, E. (2003).
Algorithms for large scale markov blanket discovery.
In *FLAIRS conference*, volume 2, pages 376–380.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006).
The max-min hill-climbing bayesian network structure learning algorithm.
Machine learning, 65(1):31–78.

References X

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. (2002).
A gene-expression signature as a predictor of survival in breast cancer.
New England Journal of Medicine, 347(25):1999–2009.