RELEVANCE LEARNING FOR REDUNDANT FEATURES

Lukas Pfannschmidt

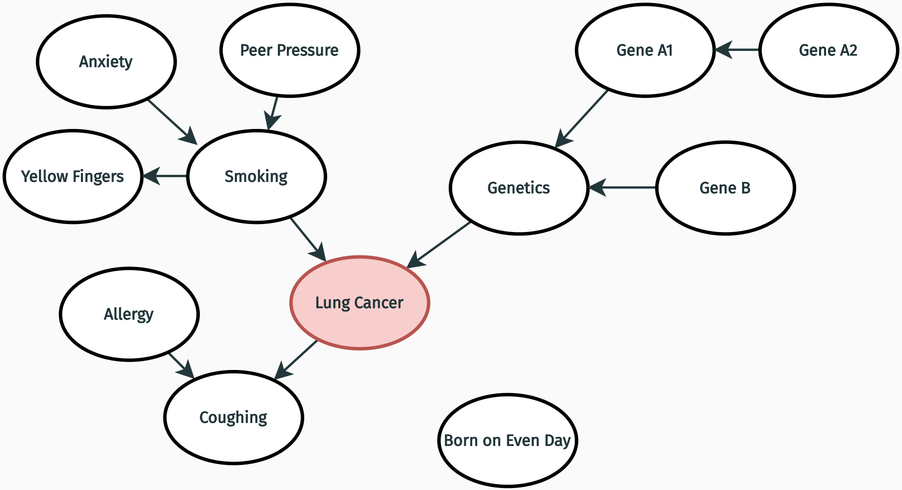
August 23, 2021

Bielefeld University, Faculty of Technology

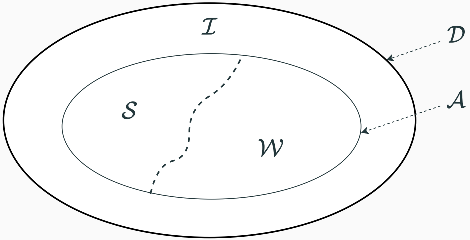
TABLE OF CONTENTS

Motivation

RELEVANT & REDUNDANT FEATURES



ALL-RELEVANT FEATURE SELECTION



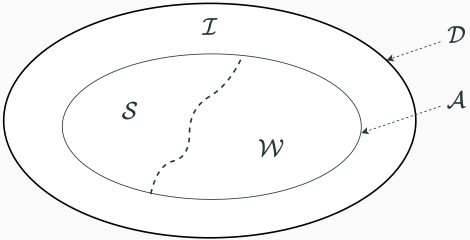
Definition (Relevancy classes)
(Kohavi et al., 1997):

- strongly relevant $\mathcal{S}$
- **weakly relevant** $\mathcal{W}$
- irrelevant $\mathcal{I}$

which result in

- all relevant features ($\mathcal{A}$)
- all input features ($\mathcal{D}$)

THESIS - RESEARCH QUESTIONS

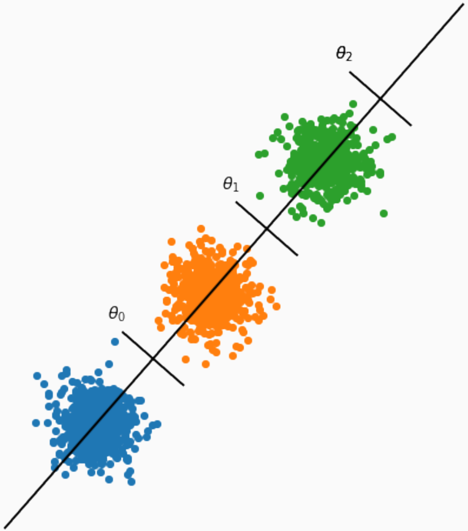


Questions

1. Can we extract the relevant feature set including redundant features?
2. Can we explore alternative solutions given by weakly relevant features?
3. Can we group related features?

Feature Relevance Bounds for Ordinal Regression

ORDINAL REGRESSION



Classification:

- model maps to two classes

Ordinal Regression:

- model maps to $1 \ldots n$ classes
- can exploit ordinal information
- often encountered in sociodemographic data (e.g. discrete rankings)

Very relevant in many medical applications and augmented decision making

ORDINAL REGRESSION SVM

Definition by Chu and Keerthi [CK07]:

$$\min_{\boldsymbol{w},\boldsymbol{b},\chi,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_1 + C\sum_{i,j}\left(\chi_i^j + \xi_i^j\right)$$

s.t. for all i,j

$$\begin{aligned}
&\boldsymbol{w}^\top \boldsymbol{x}_i^j - b_j \leq -1 + \chi_i^j \\
&\boldsymbol{w}^\top \boldsymbol{x}_i^{j+1} - b_j \geq +1 - \xi_i^{j+1} \\
&b_j \leq b_{j+1} \\
&\chi_i^j \geq 0, \xi_i^j \geq 0
\end{aligned}$$

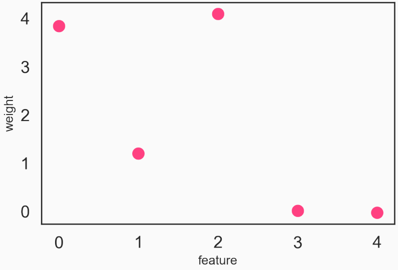
$$\min_{\boldsymbol{w},\boldsymbol{b},\chi,\xi} \quad \|\boldsymbol{w}\|_1$$

Example:

$\tilde{\boldsymbol{w}} = [3.844, 1.201, 4.096, 0.013, 0.005]$

Unique solution?

- Other viable solutions $\tilde{\boldsymbol{w}}$ could be possible, especially for high dimensional data.



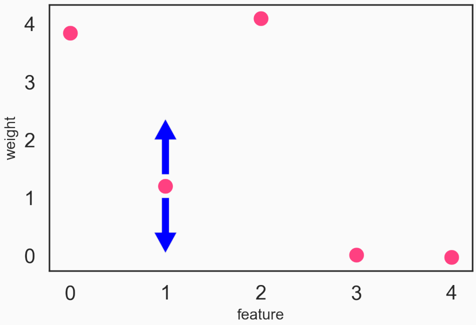
Selection:

- Features 0,1,2: > 0 (**relevant**)
- Features 3,4: ≈ 0 (**irrelevant?**)

FEATURE RELEVANCE BOUNDS

Idea
Explore all alternative solutions efficiently and classify based on solution set.

- Assure alternative solutions have identical quality
- Find maximum and minimum allowed weights



OPTIMAL SOLUTION

$$(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{b}}, \tilde{\xi}, \tilde{\chi}) = \min_{\boldsymbol{w}, \boldsymbol{b}, \chi, \boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_1 + C \sum_{i,j} \left(\chi_i^j + \xi_i^j\right)$$

$$\begin{aligned}
\text{s.t. for all i,j} \quad & \\
& \boldsymbol{w}^\top \boldsymbol{x}_i^j - b_j \leq -1 + \chi_i^j \\
& \boldsymbol{w}^\top \boldsymbol{x}_i^{j+1} - b_j \geq +1 - \xi_i^{j+1} \\
& b_j \leq b_{j+1} \\
& \chi_i^j \geq 0, \xi_i^j \geq 0
\end{aligned} \tag{1}$$

ALTERNATIVE SOLUTION SET

Optimal solution $(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{\xi}}, \tilde{\chi})$ results in:

$$\mu_X := \frac{1}{2}\|\tilde{\boldsymbol{w}}\|_1 + C \cdot \sum_{i,j}\left(\tilde{\chi}_i^j + \tilde{\xi}_i^j\right)$$

which defines the set of all similar solutions:

$$\mathcal{F}_\delta(X) := \Big\{\boldsymbol{w} \in \mathbb{R}^n \mid \exists \boldsymbol{b}, \boldsymbol{\xi}, \chi \text{ such that constraints (1) hold,}$$
$$\frac{1}{2}\|\boldsymbol{w}\|_1 + C \cdot \sum_{i,j}\left(\xi_i^j + \chi_i^j\right) \leq (1+\delta) \cdot \mu_X\Big\}$$

MINIMUM RELEVANCE BOUND

What is the minimal possible weight for feature ℓ?

$$\begin{aligned} \operatorname{minRel}(\ell) = \min_{\boldsymbol{w},\boldsymbol{b},\chi,\boldsymbol{\xi}} & \; |\boldsymbol{w}_\ell| \\ \text{s.t. for all } i,j & \text{ Ordinal SVM constraints hold and} \\ & \frac{1}{2}\|\boldsymbol{w}\|_1 + C \cdot \sum_{k,l} \left(\chi_k^l + \xi_k^l \right) \leq (1+\delta) \cdot \mu_X \end{aligned} \tag{2}$$

MAXIMUM RELEVANCE BOUND

What is the maximal possible weight for feature ℓ?

$$\text{maxRel}(\ell) = \max_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\chi},\boldsymbol{\xi}} |\boldsymbol{w}_\ell| \tag{3}$$

$$\text{s.t. for all } i, j \text{ Ordinal SVM constraints hold and}$$

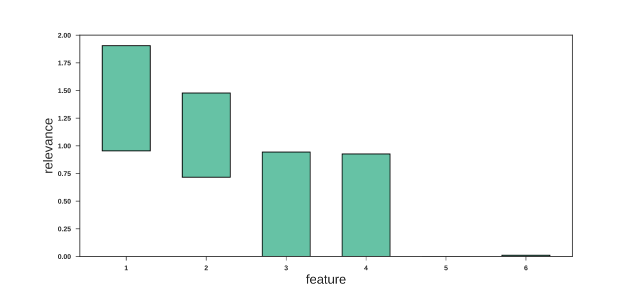
$$\frac{1}{2}\|\boldsymbol{w}\|_1 + C \cdot \sum_{k,l} \left(\chi_k^l + \xi_k^l\right) \leq (1+\delta) \cdot \mu_X$$

- can be converted into linear program and solved

RELEVANCE INTERVALS

Interpret bounds as intervals of possible feature weights.

$$\mathrm{RI}(X) := \begin{pmatrix} \mathrm{minRel}(0) & \mathrm{maxRel}(0) \\ \mathrm{minRel}(1) & \mathrm{maxRel}(1) \\ \vdots & \vdots \\ \mathrm{minRel}(d) & \mathrm{maxRel}(d) \end{pmatrix}$$



FEATURE CLASSIFICATION

Irrelevant

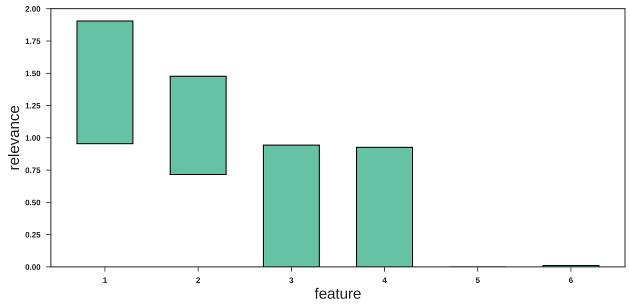
- Lower bound $= 0$
- Upper bound ≈ 0

Strongly relevant

- Lower bound > 0

Weakly relevant

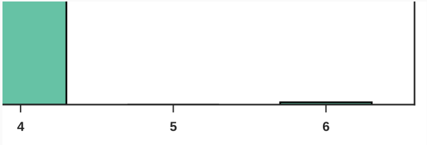
- Lower bound $= 0$
- Upper bound > 0



IMPROVE THRESHOLD

Issues

- slack in model allows false positive features
- linear programming solvers are inaccurate



How can we replace fixed threshold at 0?

Solution: Estimate rejection interval

1. Generate irrelevance samples by permuting real features
2. Determine rejection interval according to their distribution
3. Check if real relevance bound is inside the interval $\Rightarrow$ *irrelevant*

STATISTICAL FEATURE CLASSIFICATION - PREDICTION INTERVAL

Create α samples for randomly permuted input features:

$$\pi(\mathsf{maxRel}) := \mathsf{maxRel}(X \diamond \ell, \ell) \quad \ell \sim \mathcal{U}(1, d)$$

$$\widehat{\pi}(\mathsf{maxRel})_{(\alpha)} := \Big(\pi(\mathsf{maxRel})_i\Big)_{i \in \{1,\ldots,\alpha\}}$$

Define interval Π:

$$\Pi(\mathsf{maxRel}, \alpha) := \overline{\widehat{\pi}(\mathsf{maxRel})}_{(\alpha)} \pm \mathcal{T}_{\alpha-1}(p) \cdot \sigma\left(\widehat{\pi}(\mathsf{maxRel})_{(\alpha)}\right)\sqrt{1 + (1/\alpha)}$$

STATISTICAL FEATURE CLASSIFICATION - RULES

Irrelevant

- $\text{maxRel}(\ell) \in \Pi(\text{maxRel})$
- $\text{minRel}(\ell) \in \Pi(\text{minRel})$

Strongly relevant

- $\text{maxRel}(\ell) \notin \Pi(\text{maxRel})$
- $\text{minRel}(\ell) \notin \Pi(\text{minRel})$

Weakly relevant

- $\text{maxRel}(\ell) \notin \Pi(\text{maxRel})$
- $\text{minRel}(\ell) \in \Pi(\text{minRel})$

Evaluation

HOW TO EVALUATE FEATURE SELECTION PERFORMANCE?

Question

1. Can you extract the relevant feature set including redundant features?

Method:

- Create toy data sets with known feature composition
- Measure similarity of ground truth set with selected set
- Compare against alternative methods

METHODS

Problem
No comparable methods exist for the ordinal regression problem. How to compare?

Replicated Elastic Net (Zou and Hastie, 2004)

- Based on ordinal regression SVM
- Add combined L_1 and L_2 regularization
- Perform naive feature selection

FRI (Feature Relevance Intervals)

- Selection based on relevance bounds and statistical feature classification

TOY SETS (ORDINAL REGRESSION)

Datasets with 5 ordinal classes.

Dataset	#Instances	#Strong	#Weak	#Irrelevant
Set 1	150	6	0	6
Set 2	150	0	6	6
Set 3	150	3	4	3
Set 4	256	6	6	6
Set 5	512	1	2	11
Set 6	200	1	20	0
Set 7	200	1	20	20
Set 8	1000	10	20	10
Set 9	1000	10	20	200

Each set was generated in a variant with and without gaussian noise added.

FS PERFORMANCE ON TOY DATA

		Clean data		Noisy data	
score	data	EN	FRI	EN	FRI
F_1	Set 1	0.94	1.0	0.92	0.95
	Set 2	0.79	1.0	0.89	0.97
	Set 3	0.81	1.0	0.85	0.97
	Set 4	0.83	1.0	0.80	0.96
	Set 5	0.83	1.0	0.86	1.0
	Set 6	0.25	1.0	0.56	0.94
	Set 7	0.49	1.0	0.46	0.90
	Set 8	0.95	1.0	0.80	0.98
	Set 9	0.53	0.98	0.60	1.0

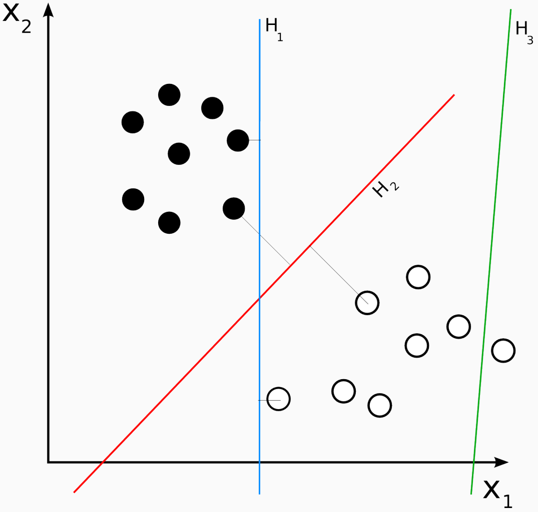
APPLIED ON REAL DATA

Dataset	#Instances	d	K	Ordered Class Distribution
Automobile	205	71	6	(3,22,67,54,32,27)
Bondrate	57	37	5	(6,33,12,5,1)
Contact-lenses	24	6	3	(15,5,4)
Eucalyptus	736	91	5	(180,107,130,214,105)
Newthyroid	215	5	3	(30,150,35)
Pasture	36	25	3	(12,12,12)
Squash-stored	52	51	3	(23,21,8)
Squash-unstored	52	52	3	(24,24,4)
TAE	151	54	3	(49,50,52)
Winequality-red	1599	11	6	(10,53,681,638,199,18)

FEATURE SET SIZE

	Average Feature Set Size			
data	FRI^s		FRI^w	EN
Automobile	4.5	$\cup$	12.6	4.0
Bondrate	0.0	$\cup$	5.4	2.0
Contact-lenses	0.9	$\cup$	1.1	2.0
Eucalyptus	2.1	$\cup$	33.2	15.6
Newthyroid	0.0	$\cup$	4.7	2.0
Pasture	0.0	$\cup$	15.5	6.0
Squash-stored	2.4	$\cup$	7.9	11.1
Squash-unstored	1.8	$\cup$	3.3	8.0
TAE	1.9	$\cup$	5.4	16.8
Winequality-red	0.0	$\cup$	7.6	5.4

BINARY CLASSIFICATION



Binary classification:

- special case of ordinal regression (2 classes)
- many methods available for comparison

$\rightarrow$ Extend comparison

RELATED METHODS: BINARY CLASSIFICATION

- **Elastic Net**: weighted sum of L1 and L2 regularization scheme (Zou and Hastie, 2004)
- **Boruta**: wrapper around Random Forest, using random contrast variables and statistical tests (Kursa and Rudnicki, 2010)
- **Ensemble Feature Selection**: combination of multiple other feature relevancy scores (Neumann et al., 2016)
- **Stability Selection**: aggregation of multiple noisy bootstrap samples of original data to compute stability score (Meinshausen and Bühlmann, 2010)
- **FRI**: Our method

TOY SETS (BINARY CLASSIFICATION)

Datasets for linear separable binary classification problem ($n = 500$ samples):

	# Strongly relevant	# Weakly relevant	# Irrelevant
Sim1	4	4	22
Sim2	12	8	10
Sim3	4	0	26
Sim4	18	0	12
Sim5	0	20	10

COULD WE IDENTIFY RELEVANT FEATURES?

F_1 score for the relevant feature set:

score	data	Boruta	EFS	EN	FRI	SS
	Sim1	**0.98**	0.96	0.62	**0.98**	0.77
	Sim2	0.82	0.76	0.84	**0.98**	0.75
F_1	Sim3	0.91	0.71	0.44	0.99	**1.00**
	Sim4	0.82	0.84	0.82	**0.99**	0.91
	Sim5	0.98	0.94	0.80	**0.99**	0.27

AVERAGE SELECTED FEATURE SET SIZE

	Boruta	EFS	EN	SS	FRI	FRI^s	FRI^w
Sim1	8.1	8.7	17.8	5.0	8.1	5.1	3.0
Sim2	14.3	12.3	26.6	12.1	19.4	12.4	7.0
Sim3	4.6	7.2	14.8	4.0	4.1	4.0	0.1
Sim4	12.6	13.2	26.2	15.0	17.9	17.9	0.0
Sim5	19.1	17.9	29.7	3.2	19.9	0.0	19.9
colp.	35.1	25.4	46.5	41.5	20.3	5.9	14.4
flip	18.8	8.1	16.9	9.1	8.9	8.8	0.1
spectf	44.0	20.3	43.1	5.9	19.9	5.9	14.0
t21	15.5	7.9	14.2	9.6	9.6	6.6	3.0
wbc	29.9	12.5	26.9	4.7	15.6	4.0	11.6

RQ1: CONCLUSION

Question

1. Can you extract the relevant feature set including redundant features?

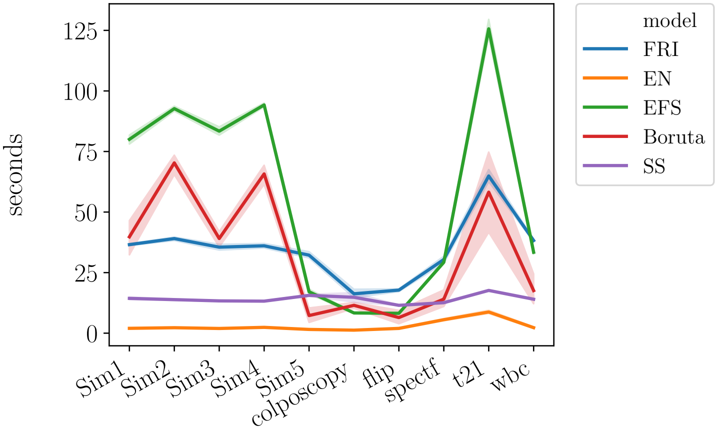
Results

- Very high selection accuracy in toy data sets (≥ 0.90 F_1)
- Overall outperforms all existing methods
- Can discriminate between weak and strong relevance
- For linear separable binary classification and ordinal regression

Efficiency of implemented tool

RUNTIME COMPARISON

Average single threaded runtime:



IMPROVE THROUGH PARALLEL PROCESSING

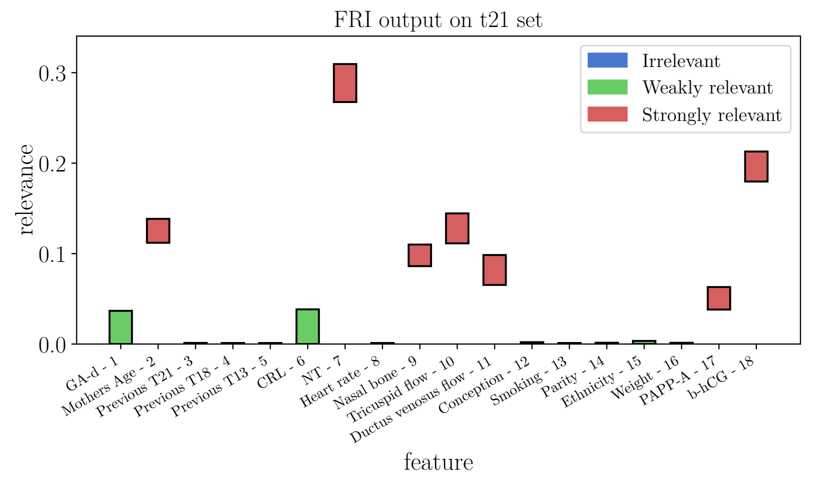
Interactive workflow requires fast results:

- relevance bounds can be computed independently per feature
- structure program for parallel processing (Joblib library)
- use frameworks for distributed computing for batch workloads

$\rightarrow$ ability to use all CPUs on machine or nodes in cluster (Dask framework)

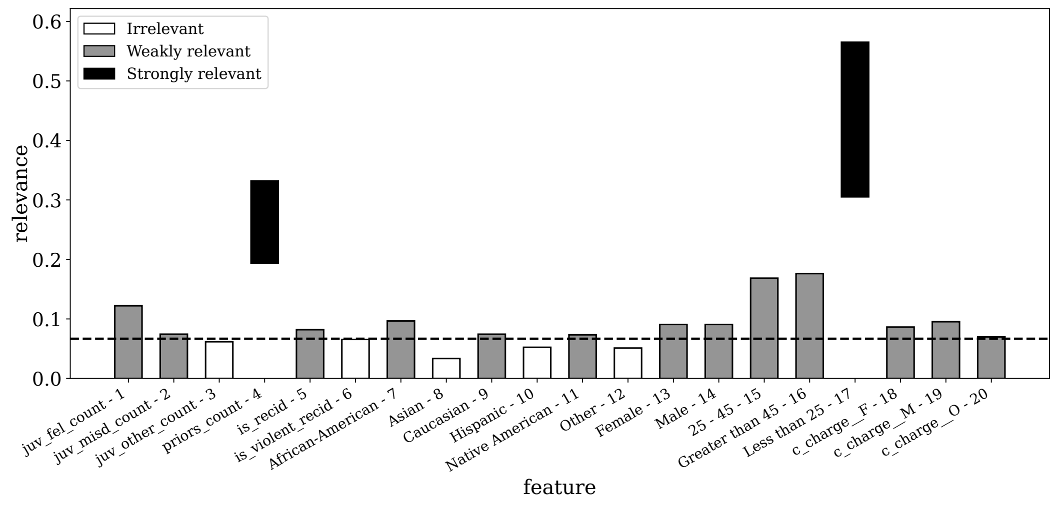
Visualization of real data examples

TRISOMY 21 - PRENATAL FEATURES



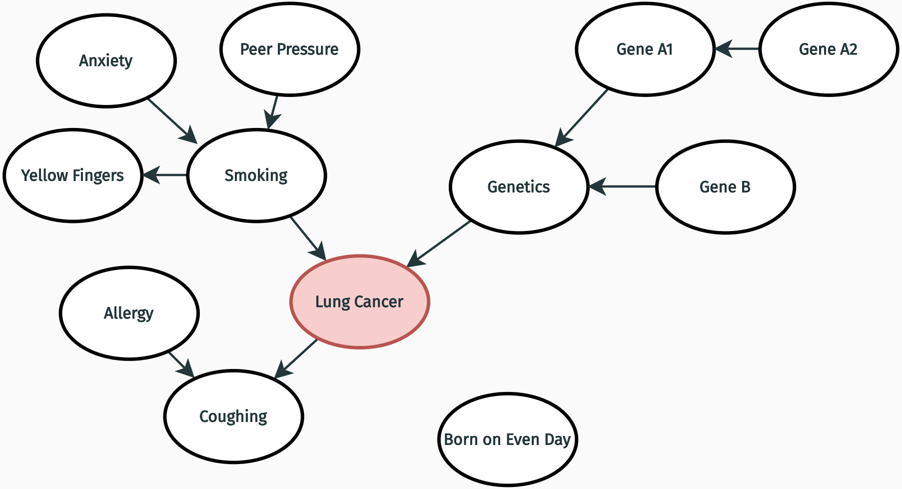
COMPASS - RECIDIVISM SCORE

Relevant features for recidivism score in US judicial system:



Alternative Solutions & Grouping

ALTERNATIVES POSSIBLE?



ALTERNATIVE SOLUTIONS

Question

- Can we explore alternative solutions given by weakly relevant features?

Idea:

- Optimization problem allows addition of feature constraints
- Constrain features based on relevance bounds

CONSTRAINED RELEVANCE BOUND

Given set of constraints K:

$$\mathrm{minRelC}(X, \ell, K) : \min_{\boldsymbol{w}, b, \boldsymbol{\chi}} |\boldsymbol{w}_\ell|$$

s.t. for all i

$$y_i \left(\boldsymbol{w}^\top x_i - b \right) \geq 1 - \chi_i, \quad \chi_i \geq 0$$

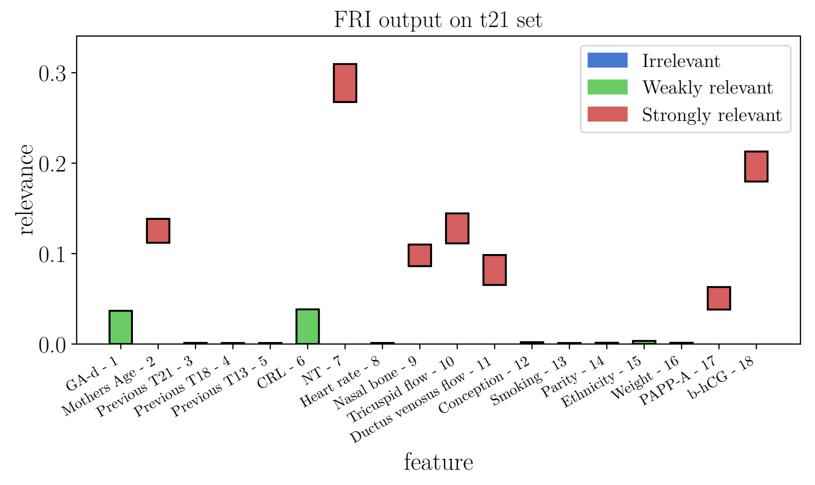
and

$$\sum_{i=1}^{n} \chi_i \leq \rho$$

$$\|\boldsymbol{w}\|_1 \leq (1 + \delta) \cdot \mu$$

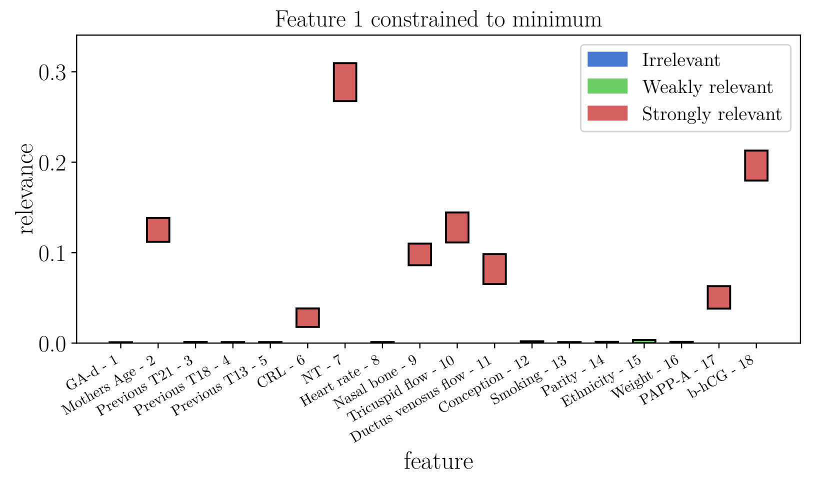
$$K_{k,\min} \geq |\boldsymbol{w}_k| \geq K_{k,\max} \quad \forall k \in K \ .$$

UNCONSTRAINED



ONE FEATURE DISABLED

$$K_{1,\text{max}} = 0$$



GROUPING RELATED FEATURES

Observation
Relevancies of related features show specific variance when they are constrained.

- Use this variance to automatically group related features

Compute the difference between normal and constrained bounds for each feature:

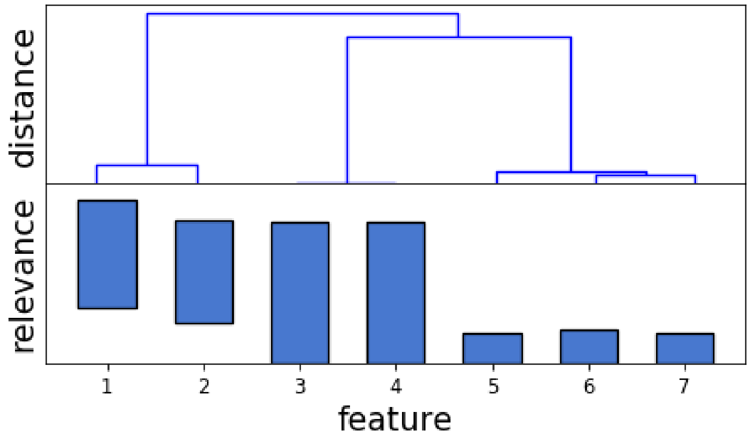
$$\mathrm{context}(k) := (\mathrm{RI} - \mathrm{RIC}(k, \min), \mathrm{RI} - \mathrm{RIC}(k, \max)) \in \mathbb{R}^{2d}$$

Similarity measure:

$$\delta(\ell, k) := \sqrt{\sum_{\substack{i=0 \\ i \neq \ell \\ i \neq k}}^{d} (\mathrm{context}(\ell)_i - \mathrm{context}(k)_i)^2}$$

$\rightarrow$ Perform aggloremative clustering with similarity measure.

Similarity measure visualized as connected tree:



EVALUATION OF GROUPING

Question

- Can we group related features?

Utilizing clustering measures and toy data:

- **V-measure**: mean of homogeneity and completeness

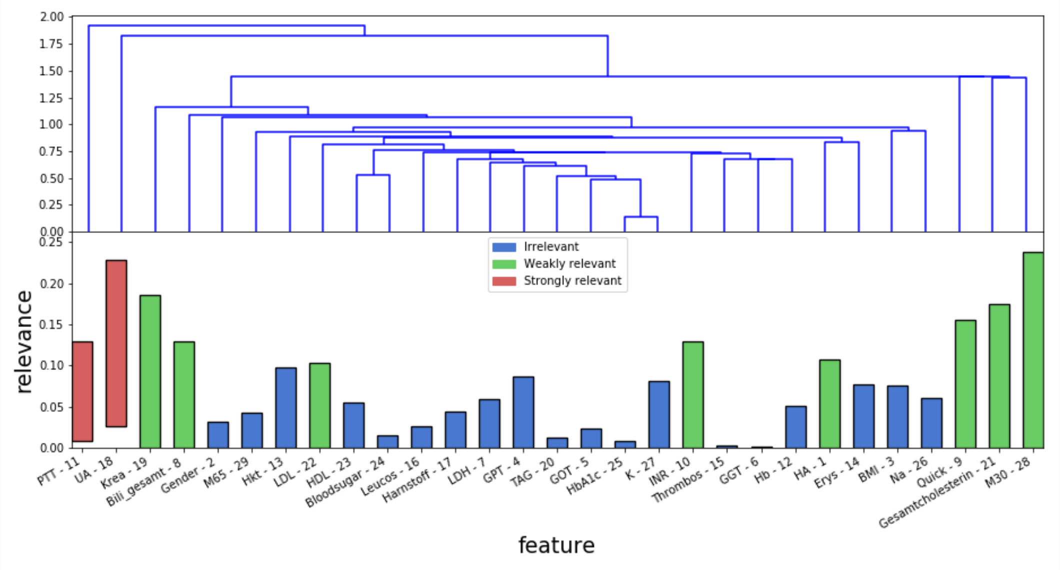
Compare against:

- **HDBSCAN**: density based clustering
- **corr-clust**: aggloremative clustering with pairwise correlation between features and maximal distance cutoff
- **FRI**: aggloremative clustering with feature relevance bound variation (context) and maximal distance cutoff

GROUPING QUALITY

model	data	V-measure	runtime (ms)
corr-clust	Set 1	0.72	1.11
	Set 2	0.89	0.93
	Set 3	**1.00**	0.90
FRI	Set 1	**1.00**	38464
	Set 2	**0.90**	30186
	Set 3	**1.00**	14520
HDBSCAN	Set 1	0.79	1.17
	Set 2	0.81	1.02
	Set 3	**1.00**	0.98

- Set 1: 20 overall features, 5 pairs of identical relevant features, **10 random noise features**
- Set 2: 18 overall features, 3 unique relevant features, 5 pairs, **5 noise features**
- Set 3: 10 overall features, only 5 identical pairs, **no noise features**


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.25
0.20
0.15
0.10
0.05
0.00
relevance
Irrelevant
Weakly relevant
Strongly relevant
PTT - 11
UA - 18
Krea - 19
Bili_gesamt - 8
Gender - 2
M65 - 29
Hkt - 13
LDL - 22
HDL - 23
Bloodsugar - 24
Leucos - 16
Harnstoff - 17
LDH - 7
GPT - 4
TAG - 20
GOT - 5
HbA1c - 25
K - 27
INR - 10
Thrombos - 15
GGT - 6
Hb - 12
HA - 1
Erys - 14
BMI - 3
Na - 26
Quick - 9
Gesamtcholesterin - 21
M30 - 28
feature

Conclusion

CONCLUSION

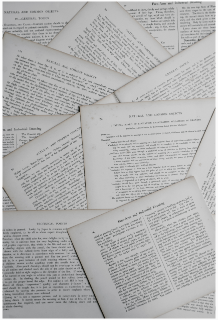
Questions answered today

1. Can we extract the relevant feature set including redundant features?
2. Can we explore alternative solutions given by weakly relevant features?
3. Can we group related features?

In this talk

- Relevance bounds for ordinal regression
- Statistical selection threshold
- Feature constraints for alternative solutions
- Clustering of weakly relevant features

PUBLICATIONS



- **"Feature Relevance Bounds for Linear Classification"**, Göpfert, Pfannschmidt, and Hammer, 2017
- **"Interpretation of Linear Classifiers by Means of Feature Relevance Bounds"**, Göpfert, Pfannschmidt, Göpfert, and Hammer, 2018
- **"FRI – Feature Relevance Intervals for Interpretable and Interactive Data Exploration"**, Pfannschmidt, Göpfert, Neumann, Heider, and Hammer, 2019
- **"Feature Relevance Bounds for Ordinal Regression"**, Pfannschmidt, Jakob, Biehl, Tino, and Hammer, 2019
- **"Feature Relevance Determination for Ordinal Regression in the Context of Feature Redundancies and Privileged Information"**, Pfannschmidt, Jakob, Hinder, Biehl, Tino, and Hammer, 2020

Thank you!

FRI: ALL-RELEVANT FEATURE SELECTION SENSITIVITY - PRECISION/RECALL

score	precis					recall				
data	Sim1	Sim2	Sim3	Sim4	Sim5	Sim1	Sim2	Sim3	Sim4	Sim5
Boruta	0.99	1.00	0.87	1.00	1.00	1.00	0.72	0.98	0.70	0.95
EFS	0.93	1.00	0.57	1.00	1.00	1.00	0.62	0.98	0.73	0.90
ElasticNet	0.46	0.74	0.28	0.69	0.67	1.00	0.98	1.00	1.00	1.00
FRI	0.98	1.00	0.98	0.99	1.00	0.99	0.97	1.00	0.98	0.99
StabilitySelection	1.00	1.00	1.00	1.00	1.00	0.62	0.60	1.00	0.83	0.16

DID WE RETAIN ALL INFORMATION?

Table 1: ROC-AUC values of a logistic regression model using features selected by listed models. The values are averaged over 50 bootstraps.

	ROC-AUC				
data	Boruta	EFS	EN	FRI	SS
colposcopy	0.568	0.586	0.640	**0.661**	0.625
flip	0.804	0.652	**0.815**	0.743	0.705
spectf	0.871	0.874	0.867	0.880	**0.888**
t21	0.971	0.977	0.971	0.975	**0.978**
wbc	0.997	0.998	0.998	0.998	0.999

MMAE ON REAL DATA

	Macro-averaged Mean Absolute Error (MMAE)			
data	p-OGMLVQ	a-OGMLVQ	$\mathrm{M}_e^{L_1}$	$\mathrm{M}_i^{L_1}$
Automobile	0.482	**0.446**	0.532	0.516
Bondrate	0.768	**0.737**	0.939	0.949
Contact-lenses	0.243	0.221	**0.190**	0.265
Eucalyptus	0.450	0.477	**0.390**	**0.390**
Newthyroid	0.124	0.097	**0.043**	0.045
Pasture	**0.307**	0.318	0.374	0.430
Squash-stored	0.415	0.411	**0.371**	**0.371**
Squash-unstored	0.488	**0.228**	0.280	0.300
TAE	0.553	**0.537**	0.552	0.664
Winequality-red	1.078	1.069	0.868	**0.790**