More than Accuracy: Interpretability

Chenhao Tan

@MLDG 08/15/2013

Interpretability in Medical Research

- Interpretability is a fundamental desirable quality in many domains

if total cholesterol $\geq$160 **and** smoke **then** *10 year CHD risk* $\geq$ *5%*
else if smoke **and** systolic blood pressure$\geq$140 **then** *10 year CHD risk* $\geq$ *5%*
else *10 year CHD risk* $<$ *5%*

Figure 1: Example decision list created using the NHBLI Framingham Heart Study Coronary Heart Disease (CHD) inventory for a 45 year old male.

Letham et al. 2012

Interpretability is everywhere

- Introduction

Some instances are used to help people understand

MODULAIRE BUYS BOISE HOMES PROPERTY

Modulaire Industries said it acquired the design library and manufacturing rights of privately-owned Boise Homes for an undisclosed amount of cash. Boise Homes sold commercial and residential prefabricated structures, Modulaire said.

USX, CONSOLIDATED NATURAL END TALKS

USX Corp's Texas Oil and Gas Corp subsidiary and Consolidated Natural Gas Co have mutually agreed not to pursue further their talks on Consolidated's possible purchase of Apollo Gas Co from Texas Oil. No details were given.

JUSTICE ASKS U.S. DISMISSAL OF TWA FILING

The Justice Department told the Transportation Department it supported a request by USAir Group that the DOT dismiss an application by Trans World Airlines Inc for approval to take control of USAir. "Our rationale is that we reviewed the application for control filed by TWA with the DOT and ascertained that it did not contain sufficient information upon which to base a competitive review," James Weiss, an official in Justice's Antitrust Division, told Reuters.

E.D. And F. MAN TO BUY INTO HONG KONG FIRM

The U.K. Based commodity house E.D. And F. Man Ltd and Singapore's Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo's 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will rename the firm Man Pacific (Holdings) Ltd.

Figure 2: Four documents from the Reuters-21578 category "corporate acquisitions" that do not share any content words.

Joachims 2001

Interpretability is everywhere

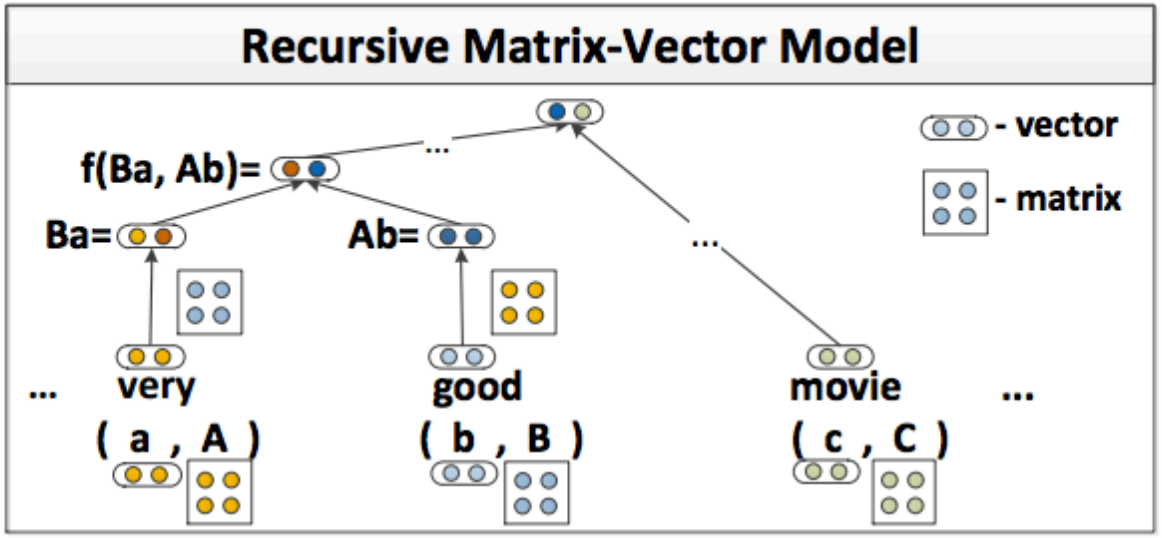


Figure 1: A recursive neural network which learns semantic vector representations of phrases in a tree structure. Each word and phrase is represented by a vector and a matrix, e.g., *very* = (a, A). The matrix is applied to neighboring vectors. The same function is repeated to combine the phrase *very good* with *movie*.

Socher et al. 2013

Interpretability is everywhere

- Introduction
- Experiment
  - Interpreting linear coefficients

LIWC+BIGRAMS$^{+}_{SVM}$		LIWC$_{SVM}$	
TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE
-	chicago	hear	i
...	my	number	family
on	hotel	allpunct	perspron
location	,_and	negemo	see
)	luxury	dash	pronoun
allpunct$_{LIWC}$	experience	exclusive	leisure
floor	hilton	we	exclampunct
(	business	sexual	sixletters
the_hotel	vacation	period	posemo
bathroom	i	otherpunct	comma
small	spa	space	cause
helpful	looking	human	auxverb
$	while	past	future
hotel_.	husband	inhibition	perceptual
other	my_husband	assent	feel

Table 5: Top 15 highest weighted truthful and deceptive features learned by LIWC+BIGRAMS$^{+}_{SVM}$ and LIWC$_{SVM}$. Ambiguous features are subscripted to indicate the source of the feature. LIWC features correspond to groups of keywords as explained in Section 4.2; more details about LIWC and the LIWC categories are available at `http://liwc.net`.

Ott et al. 2011

Interpretability is everywhere

- Introduction
- Experiment
  - Topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Blei et al. 2003

Interpretability is everywhere

- Introduction
- Experiment
- Modeling (Occam's razor)

What is interpretability?

- Sadly, I cannot find a good formulated definition
- Merriam Webster

  Interpret: to explain or tell the meaning of
  to present in understandable terms

  Understand: to grasp the meaning of
  to grasp the reasonableness of
- In the formal logic sense, an interpretation is a map- ping of a formal construct to the entities and their relations it represents. [Ruping 2006]

Interpretability in Machine Learning

- The understandability of a model
- The understandability of why the model is true or how the model is induced from data

Interpretability in Machine Learning

- *The understandability of a model*
- The understandability of why the model is true or how the model is induced from data

By definition subjective

- Interpretability is hard to formalize, as it is a very subjective concept
  - Capacity of human brain
  - Formal model vs representative examples
  - Plots vs natural languages
  - Different background in understanding box plots, vector spaces, probability distribution …

Measuring Interpretability

- A survey over human experts (probably widely used in biology, medical research, even linguistics or social scientists)

Measuring Interpretability

- A survey over human experts (probably widely used in biology, medical research, even linguistics or social scientists)

  This can be very non-trivial. A very cute test that I learned about measuring interpretability of topic models:

  Word intrusion detection, for each topic get the top 5 words, and sample a random word from the bottom half vocabulary, present the 6 words in random order to human, and test the accuracy of finding the "intruder"

  [Murphy et al. 2012]

Measuring Interpretability

- A survey over human experts (probably widely used in biology, medical research, even linguistics or social scientists)
- The difficulty in expressing with natural languages
- Formal complexity measure
- Cognitive difficulty

Measuring Interpretability

- How to prove a formal measure of complexity is valid in representing interpretability
  - No idea, psychological investigations may help, but probably measure/domain specific
- Formal complexity measures are either only applicable for a specific class of models or too coarse

Three goals of interpretability

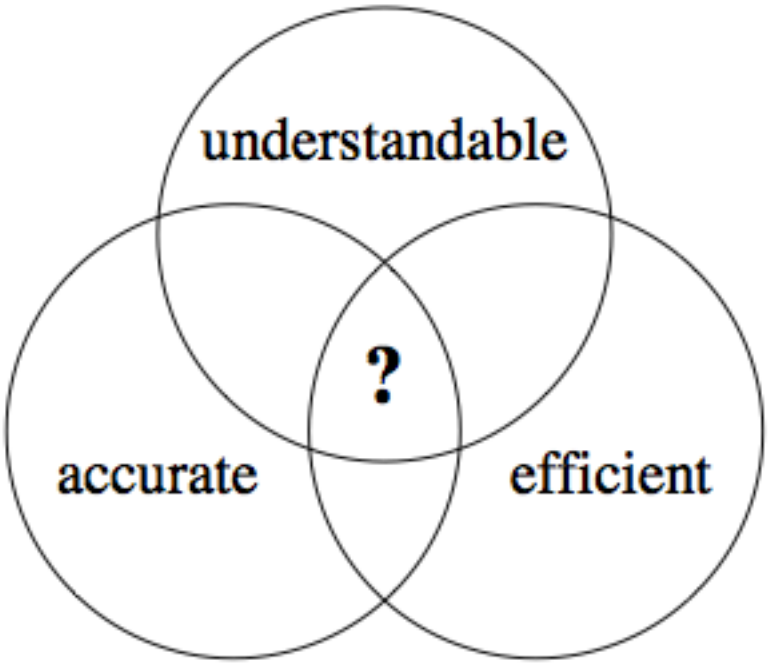


Figure 1.4: Three goals of interpretability

How can we solve the interpretability problem

An interesting high-level model

- Data = Global Model + Local Models + Noise

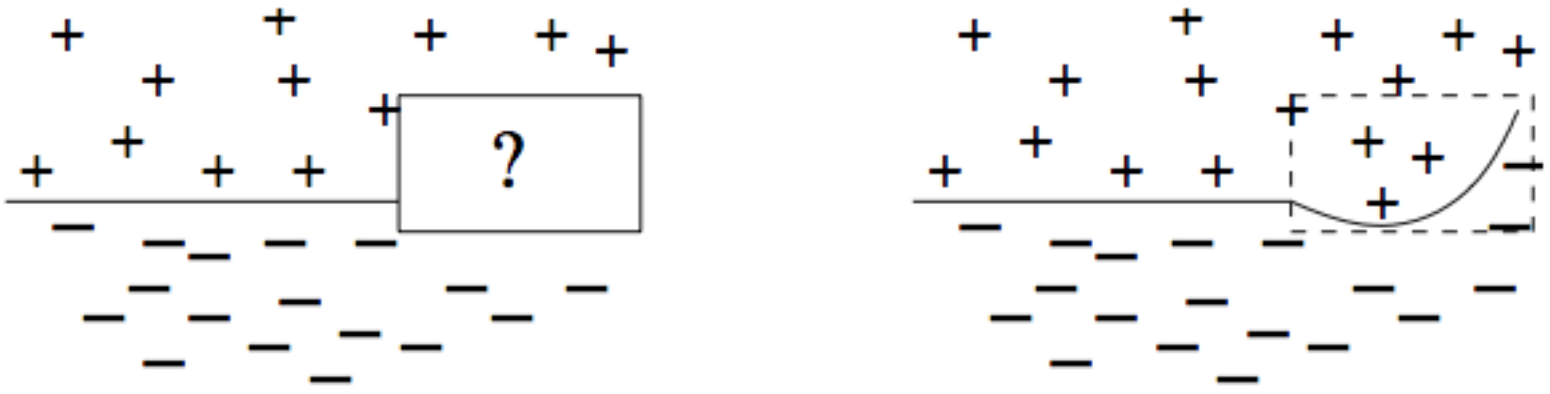


Figure 1.3: The local model idea

- Interpretability-optimal Global-plus-Local Model

Strategies to improve the interpretability

- Black Box Optimization: How can one optimize the interpretability of a classifier if one does not now how the classifier is working?
- White Box Optimization: How can knowledge about the internals of the learning algorithm help to increase understandability?
- Local Patterns: What is the best way to describe on which examples not to trust the classifier?
- Local Models: Given an understandable classifier, how can one add extra additional classification performance without hurting understandability?

Strategies to improve the interpretability

- *Black Box Optimization: How can one optimize the interpretability of a classifier if one does not now how the classifier is working?*
- *White Box Optimization: How can knowledge about the internals of the learning algorithm help to increase understandability?*
- Local Patterns: What is the best way to describe on which examples not to trust the classifier?
- Local Models: Given an understandable classifier, how can one add extra additional classification performance without hurting understandability?

Black box optimization

- Feature selection
- Instance selection
- Decomposing complex non-linear models into smaller, easier to understand linear local models
- Direct complexity reduction

Feature Selection

- Traditionally, although the impact of feature selection on interpretability is obvious, feature selection is usually motivated from the perspective of classification performance
- Adding sparsity constraints

Instance selection

- Informative instances
  - Prototype instances
  - Discriminating instances

Instance Selection

- Pro: easier for a domain expert to understand than abstract models and decision rules
- Con: risk of seeing the wood for the trees
- A missing general measure of instance importance
  - Representative, a similarity measure is needed
  - Reflect information from the view of the learner instead of from other examples

Instance Selection

- Pro: easier for a domain expert to understand than abstract models and decision rules
- Con: risk of seeing the wood for the trees
- A missing general measure of instance importance
  - Representative, a similarity measure is needed
  - Reflect information from the view of the learner instead of from other examples

  The support vectors in SVMs are discriminating instances and it can be shown that the SVM trained on the support vectors alone is identical to the SVM on the complete data set

Black box optimization

- Feature selection
- Instance selection
- *Decomposing complex non-linear models into smaller, easier to understand linear local models*
- *Direct complexity reduction*

White box optimization

- Linear models

  Weights, or
  coefficients of features

LIWC+BIGRAMS$^{+}_{SVM}$		LIWC$_{SVM}$	
TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE
-	chicago	hear	i
...	my	number	family
on	hotel	allpunct	perspron
location	,_and	negemo	see
)	luxury	dash	pronoun
allpunct$_{LIWC}$	experience	exclusive	leisure
floor	hilton	we	exclampunct
(	business	sexual	sixletters
the_hotel	vacation	period	posemo
bathroom	i	otherpunct	comma
small	spa	space	cause
helpful	looking	human	auxverb
$	while	past	future
hotel_.	husband	inhibition	perceptual
other	my_husband	assent	feel

Table 5: Top 15 highest weighted truthful and deceptive features learned by LIWC+BIGRAMS$^{+}_{SVM}$ and LIWC$_{SVM}$. Ambiguous features are subscripted to indicate the source of the feature. LIWC features correspond to groups of keywords as explained in Section 4.2; more details about LIWC and the LIWC categories are available at `http://liwc.net`.

Ott et al. 2011

A different metric

Impact:

$$\frac{w_j}{N} \sum_{i=1}^{N} f_j(x_i)$$

Bill Survival	
sponsor is in the majority party (2)	0.525
sponsor is in the majority party and on the committee (4)	0.233
sponsor is a Democrat (1)	0.135
sponsor is on the committee (3)	0.108
bill introduced in year 1 (11)	0.098
sponsor is the referred committee's chair (5)	0.073
sponsor is a Republican (1)	0.069
Bill Death	
bill's sponsor is from NY (9)	-0.036
sponsor is Ron Paul (Rep., TX) (6)	-0.023
bill introduced in December (10)	-0.018
sponsor is Bob Filner (Dem., CA) (6)	-0.013

Table 2: Baseline model: high-impact features associated with each outcome and their impact scores (eq. 4).

Yano et al. 2012

Which is better?

- Is there some principled way even if we know how the models work?

Which is better?

- Is there some principled way even if we know how the models work?

  I do not know …

Interpreting SVM

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$$

- Reduce n: a transformation of the SVM classifier in terms of different basis functions is presented
- Investigating the possibility of describing a SVM using logical formulas
- A novel visualization method for Support Vector Machines that combines the structure of the hypothesis space with the form of the decision function

Interpreting SVM

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$$

- *Reduce n: a transformation of the SVM classifier in terms of different basis functions is presented*
- *Investigating the possibility of describing a SVM using logical formulas*
- A novel visualization method for Support Vector Machines that combines the structure of the hypothesis space with the form of the decision function

Sparse models

$$
\begin{aligned}
f(x) &= \sum_{i=1}^{s} \alpha_i K(x_i, x) \\
&= \sum_{i=1}^{s} \alpha_i \Phi(x_i) * \Phi(x) \\
&= \left(\sum_{i=1}^{s} \alpha_i \Phi(x_i) \right) * \Phi(x) \\
&=: w * \Phi(x)
\end{aligned}
$$

- Pre-image w, which can be interpreted as a prototypical example, but not always possible

Definition 4.1.2 (Approximate Pre-Image Problem [Mika et al., 1998]). Given a feature map $\Phi : X \to \mathcal{X}$ and an element of the feature space $w \in \mathcal{X}$, the approximate pre-image problem is to find a $x \in X$ which approximates w in the 2-norm in feature space. I. e. $x = \arg\min_x ||w - \Phi(x)||^2_{\mathcal{H}}$.

An example

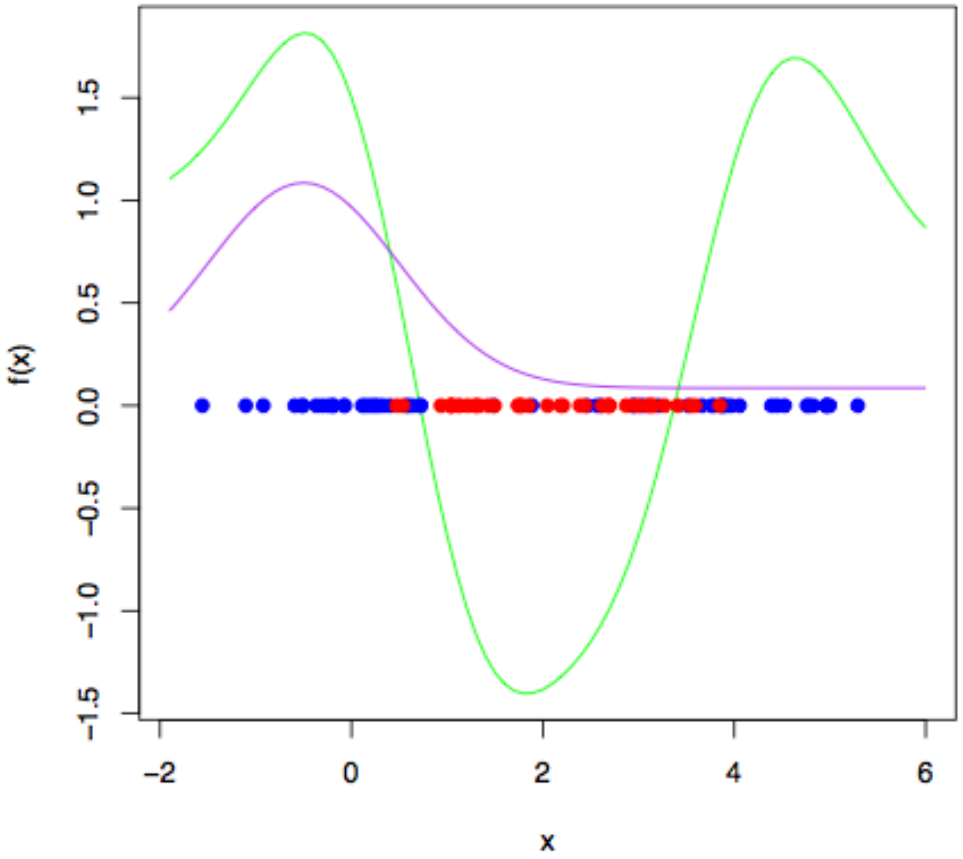


Figure 4.1: SVM decision function and pre-image

Sparse Models

$$
\begin{aligned}
x &= \arg\min_x ||w - \Phi(x)||_2^2 \\
&= \arg\min_x w * w - 2w * \Phi(x) + \Phi(x) * \Phi(x) \\
&= \arg\min_x -2w * \Phi(x) + \Phi(x) * \Phi(x) \\
&= \arg\min_x -2f(x) + K(x,x)
\end{aligned}
$$

When K(x,x) is constant, the pre-image is the point x with the highest decision function value f(x)

Other methods (not covered now):

Distance-based pre-image construction

Learning Pre-images

Reduced set methods (limit the number of support vectors or something similar to l1-regularization on how each instance is used)

Function approximation

Just a comparison

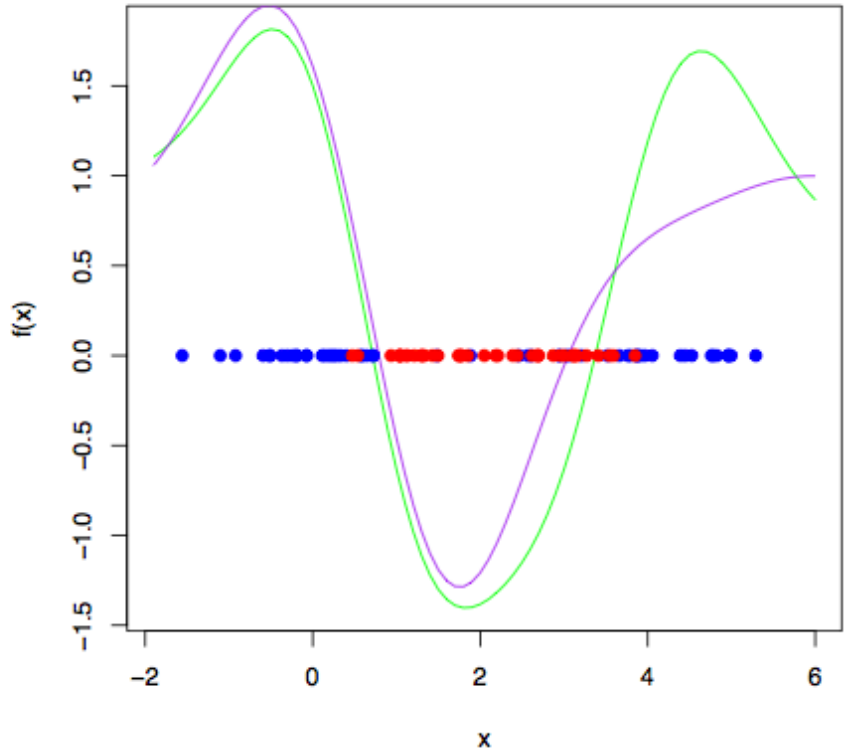


Figure 4.3: SVM decision function and reduced set approximation

Note that here what we show is actually measuring the accuracy of the interpretation instead of interpretability.

Logical Approximation

- Existence of a Logical Approximation: theoretically yes!
- Trepan algorithm [not really practical]
  - Starts with a decision tree consisting of only one leaf, which represents the whole example set
  - In each iteration it selects the leaf v where the approximation is worse according to

    Err(v)=number(v) * (1 – fidelity(v))

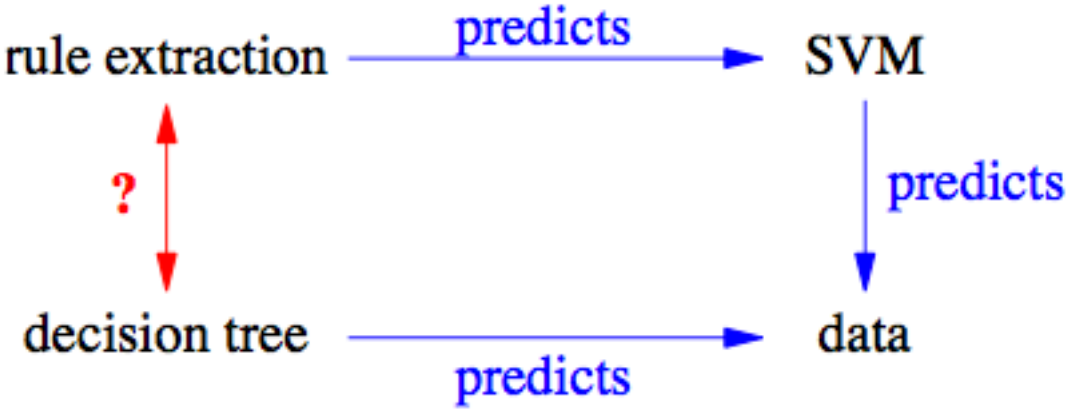
  number: the estimated examples that fall into the leaf

  fidelity: estimated accuracy

  Each leaf represents a region in the input space

A conceptual problem

- Predicting the true class vs. Predicting the classifiers output



Results

Name	linear SVM			radial basis SVM		
	class	SVM	Sig	class	SVM	Sig
Business	0.165	0.133	*o*	0.147	0.145	*o*
Covtype	0.182	0.052	++	0.159	0.093	++
Diabetes	0.132	0.066	++	0.142	0.080	++
Digits	0.010	0.009	*o*	0.008	0.015	*o*
Physics	0.301	0.095	++	0.293	0.146	++
Ionosphere	0.119	0.105	*o*	0.071	0.073	*o*
Liver	0.318	0.220	+	0.269	0.240	*o*
Medicine	0.230	0.040	++	0.191	0.063	++
Mushroom	0.000	0.000	*o*	0.001	0.001	*o*
Promoters	0.256	0.238	*o*	0.180	0.180	*o*
Insurance	0.002	0.002	*o*	0.008	0.005	++
Balance	0.165	0.149	*o*	0.142	0.156	*o*
Dermatology	0.000	0.000	*o*	0.016	0.016	*o*
Iris	0.013	0.013	*o*	0.013	0.013	*o*
Voting	0.020	0.000	++	0.038	0.025	*o*
Wine	0.061	0.066	*o*	0.220	0.135	++
Breast	0.030	0.019	*o*	0.029	0.026	*o*
Garageband	0.281	0.225	++	0.267	0.210	++

Take away: Interpretability Heuristics

- Use a small number of features and parameters
- Split up a problem into several independent sub-problems
- Use examples and basic features instead of formal models and constructed features
- Use what they already know