August 30, 2018

Measures of model interpretability for model selection

André M. Carrington
Ph.D, P.Eng, CISSP

International Cross-Domain Conference for Machine Learning and Knowledge Extraction

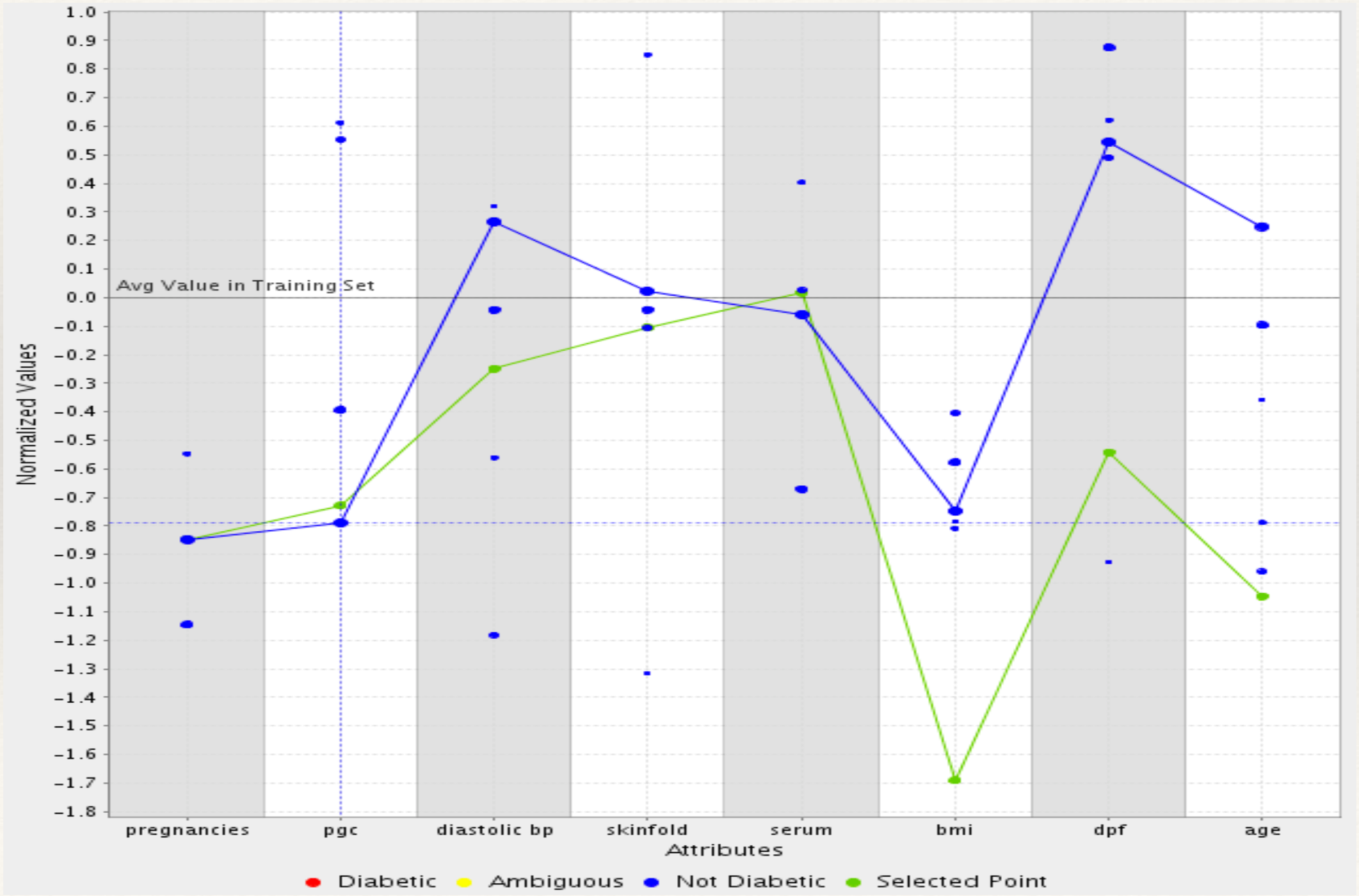
Agenda

- Problem
- Contributions
- Concepts
- Conclusions
- Other / future work

Problem

- Machine learning may be a black box.
- Some people care (per regulations), some do not.
- Interpretations are not widely known, used, understood.
- **Model selection** is based on accuracy (predominantly)
- What does the literature mean by **model interpretability**?
- How do we **quantitatively define and measure** it?

An interpretation of an SVM result from Barbella et al.



Objectives

- Clarify concepts
- Create measures
- Apply measures to model selection, assess any trade-offs
  - SVM classification for health care
  - Atomic data types only (e.g., no images, time-series)
  - Multiple kernels

Contributions

For the previous scope, focused on two kernels:

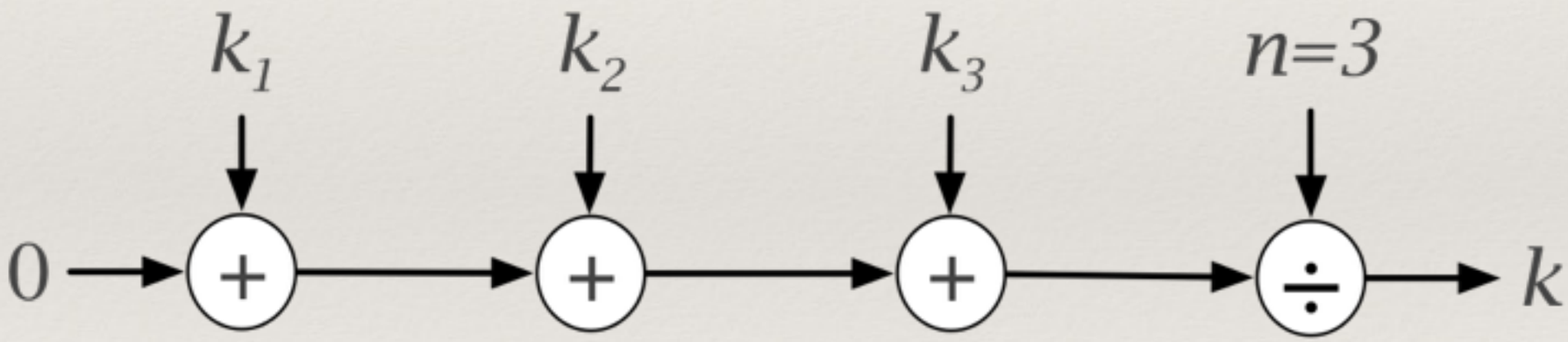
1. Defined new concepts and quantitative measures
2. No accuracy vs. **transparency** trade-off **between** kernels
3. There is no accuracy vs. **inherent model interpretability** trade-off **within** or **between** kernels
    - Possible negative linear trend at the balanced front

Concepts

	1. Doctor/HCP	2. Doctor/HCP	3. Doctor/HCP	4. Patient, Peer HCP
Actor:				
Task:	**Select models,** method, features *a priori*	Review and **select** results, **models**, **interpretations**	Construct **explanations** of results & models	Understand **explanations** of results & models
	→	learn & test →	→	→
Measure:	**transparency**	**inherent model interpretability** = understandability	explainability, clinician numeracy	patient numeracy, clinician numeracy

in focus (blue text) **not in focus (red text)**

An additive model with three parts, one for each feature



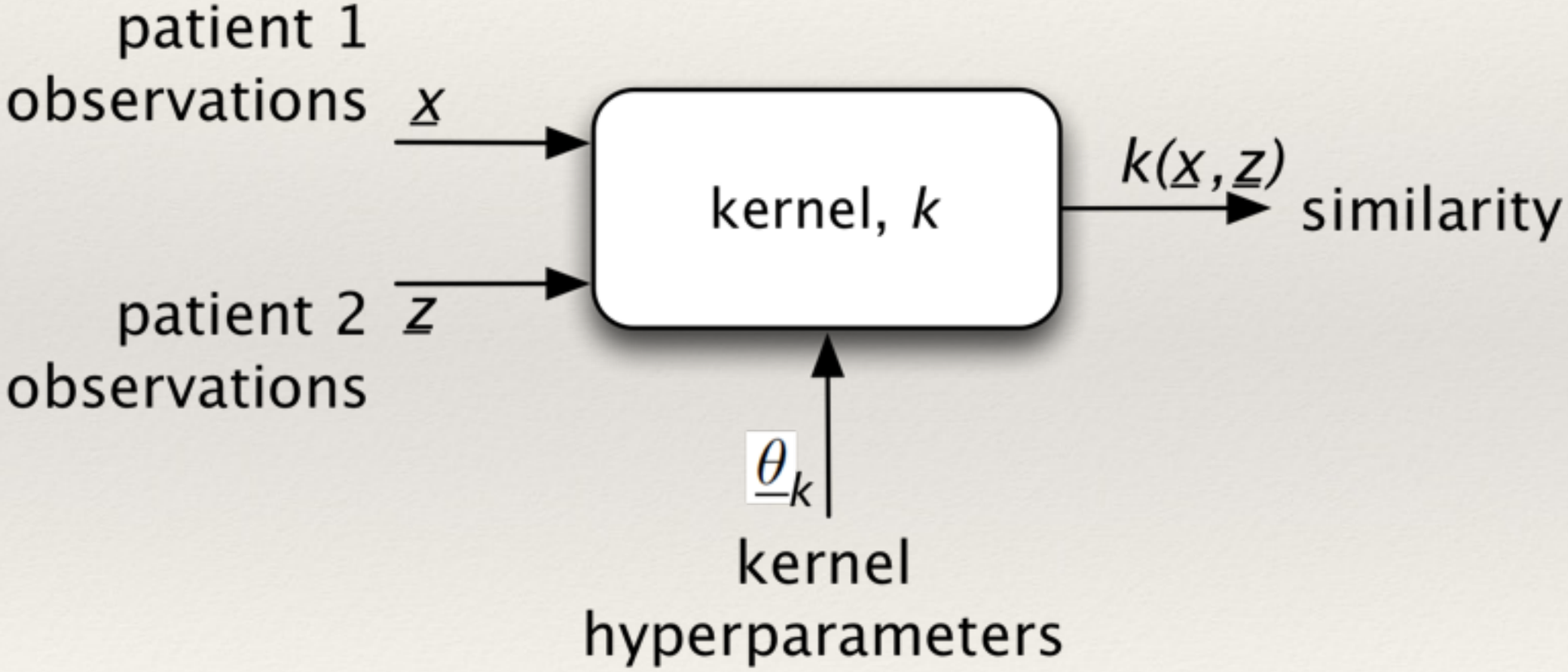
Transparency

Dirac measure	Desc.	Linear	Polynomial	Gaussian RBF	Sigmoid	Mercer Sigmoid
∂_{essep}	Exp. Sym. Sep.	✓	×	×	×	×
∂_{fin}	Finite	✓	✓	×	×	✓
∂_{eM}	Exp. Mercer	✓	×	✓[44]	×	✓
$\partial_{\times}$	Multiplicative	×	×	×	×	×
∂_{uni}	Uniform	✓	×	×	×	✓
∂_{adm}	Admissible	✓	✓	✓	×	✓
$\check{U}_{\partial}$ (%)		83	33	33	0	67

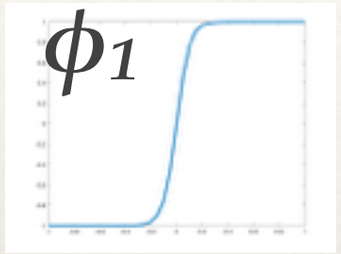
Legend: Green = top result. Light green = second best result.

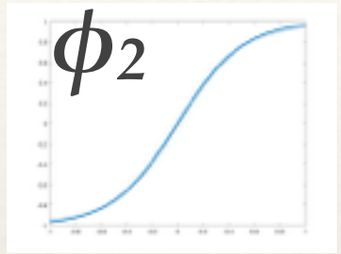
Kernels are key to SVM models

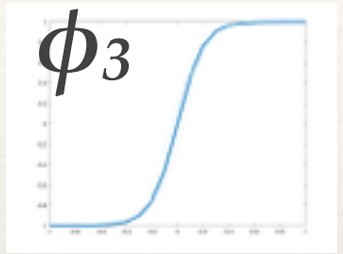
- Kernels judge the similarity between two patients



Mercer sigmoid kernel (non-uniform, 3 features)

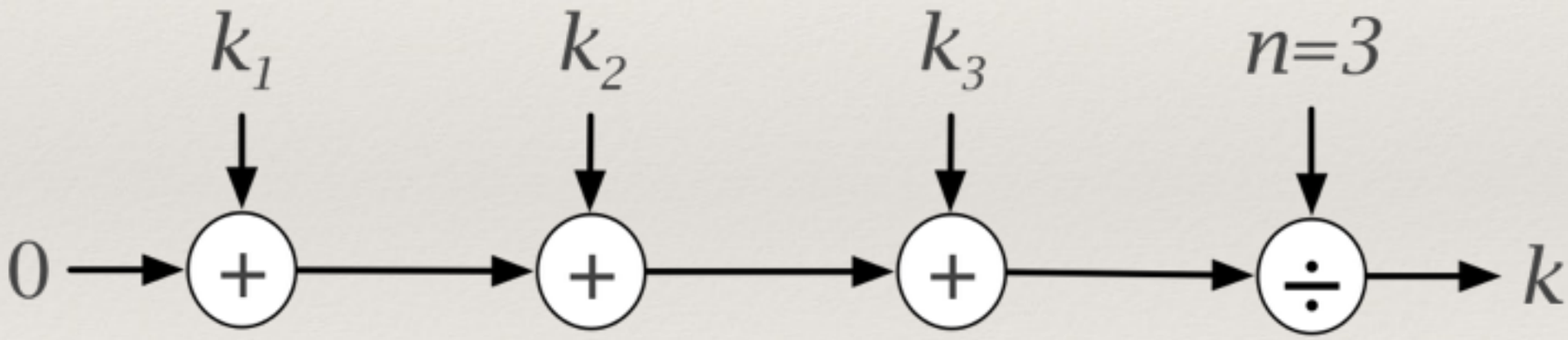






$$\phi_i(x_i) = \tanh\left(\frac{x_i - d_i}{b_i}\right)$$

$$k_i(\underline{x}, \underline{z}) = \phi_i(x_i)\,\phi_i(z_i)$$

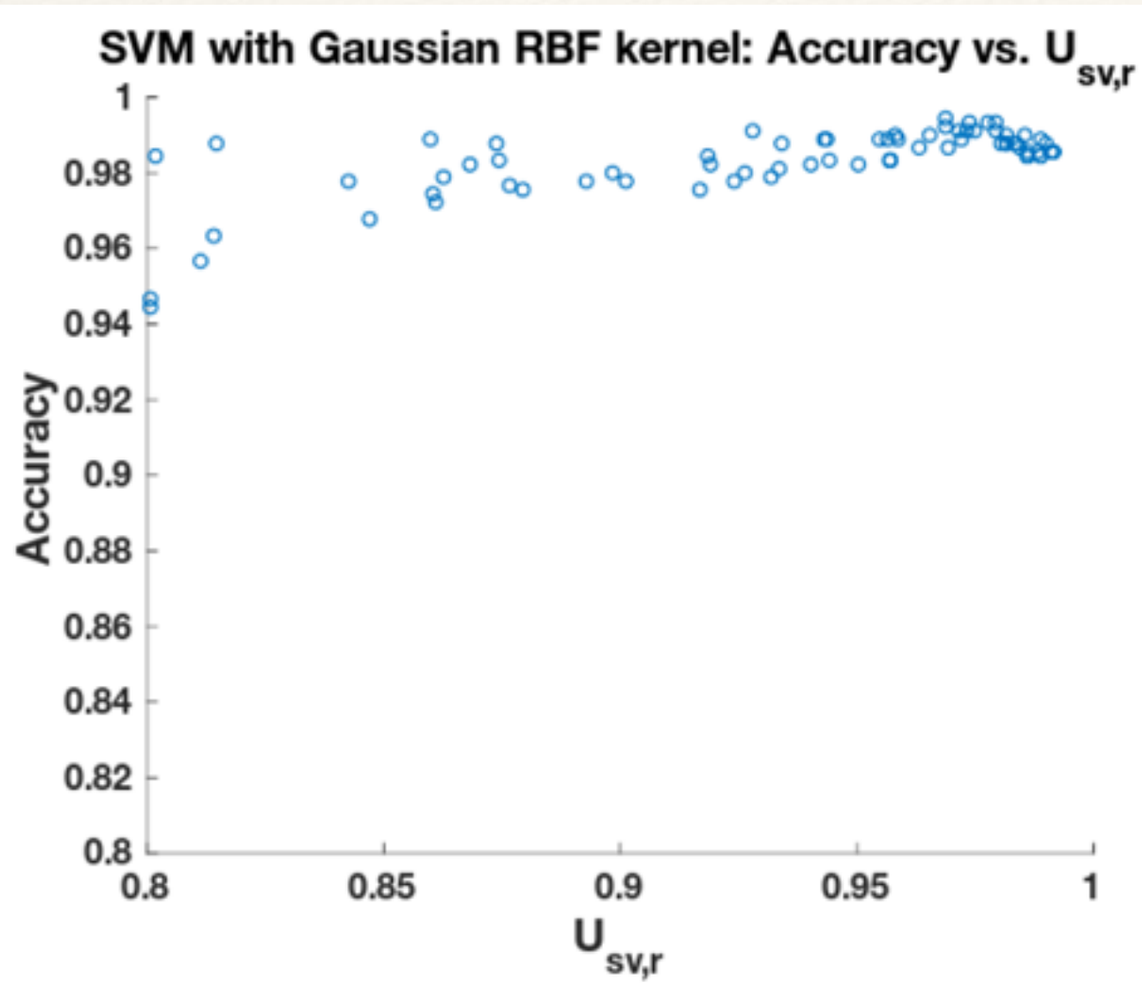


- ϕ_1 is a decision function which suggests class -1 to the left of centre and class +1 to the right, in its contribution to a sum.
- ϕ_i are sigmoids, S-shaped curves, like a square step function but where smoothness represents uncertainty from class overlap (next).

Inherent model interpretability

- The inherent simplicity (understandability) of a model, independent of any specific person
- Is linearly independent of
  - clinically meaningful features
  - model utility
  - goodness of fit
  - accuracy.

Accuracy vs. interpretability 1 (U_{sv})



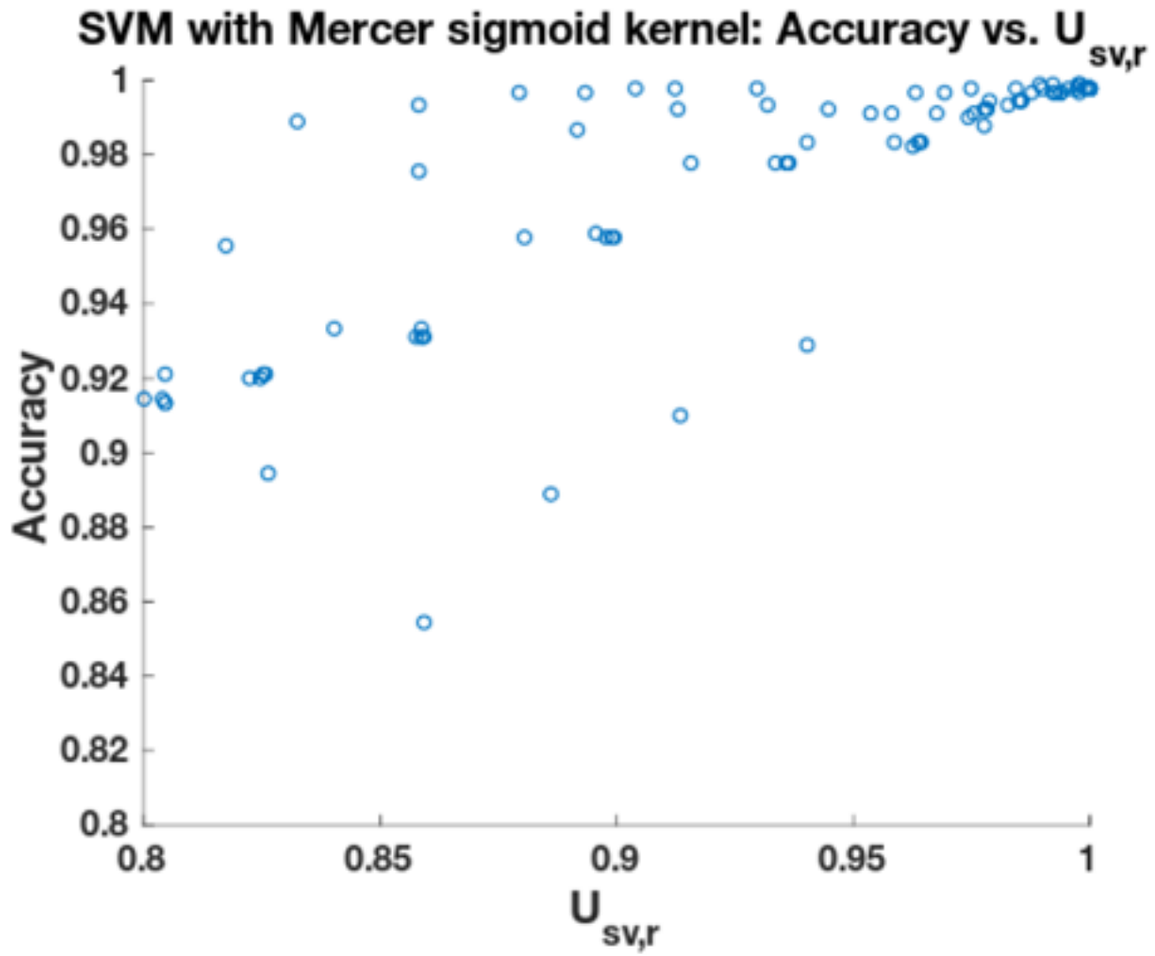


Figure 6.2: In classification for the toy problem, there are many results with high accuracy and high inherent model interpretability, with almost no sacrifice in the latter for maximum accuracy.

Accuracy vs. interpretability 2 (U_{sv})

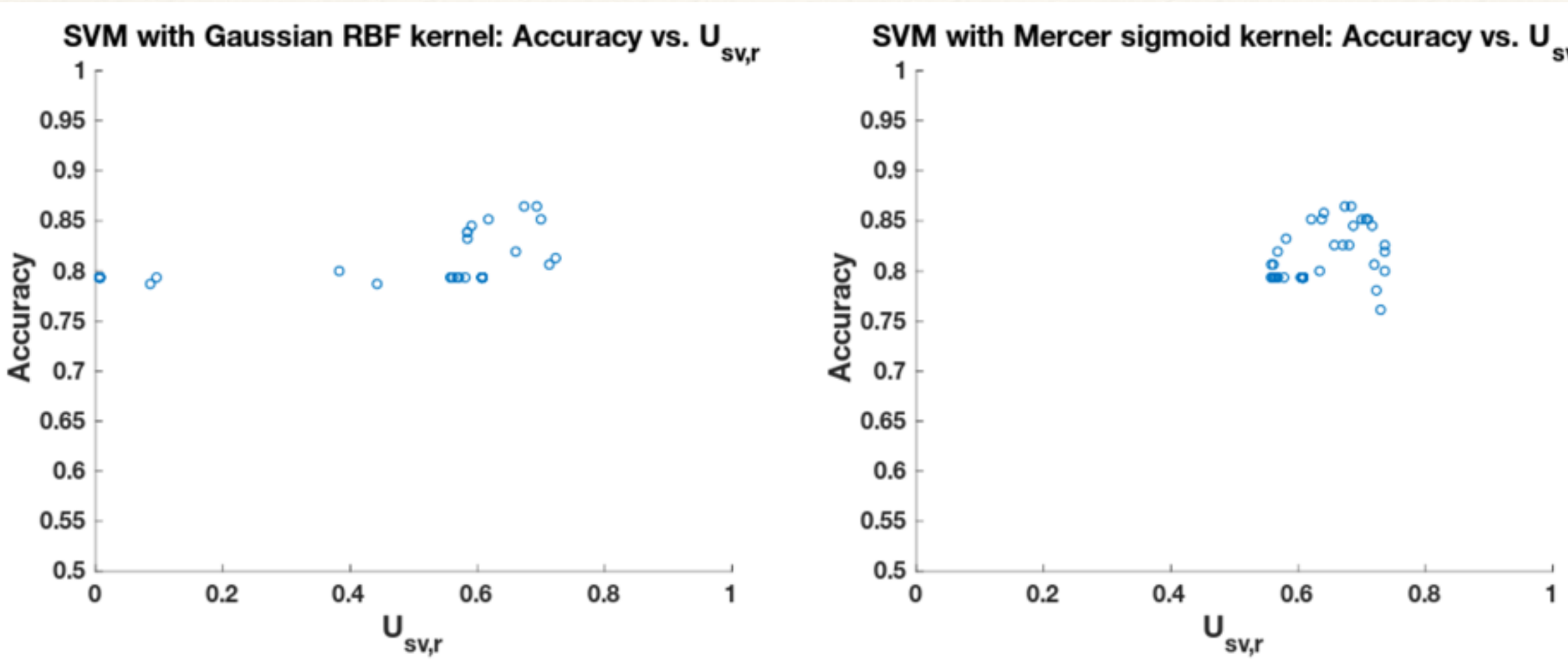
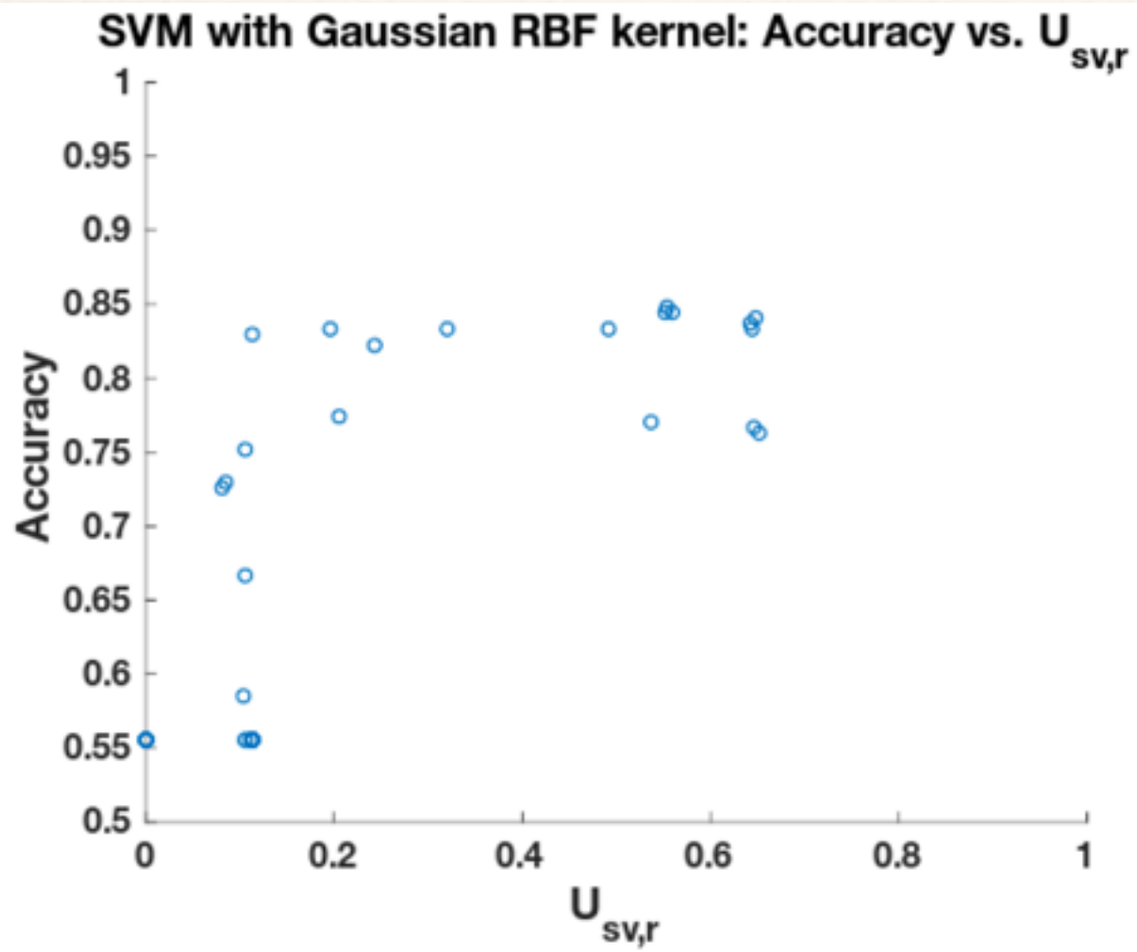


Figure 6.3: In classification with the Hepatitis data set there is a less than 5% sacrifice in inherent model interpretability for the highest accuracy.

Accuracy vs. interpretability 3 (U_{sv})



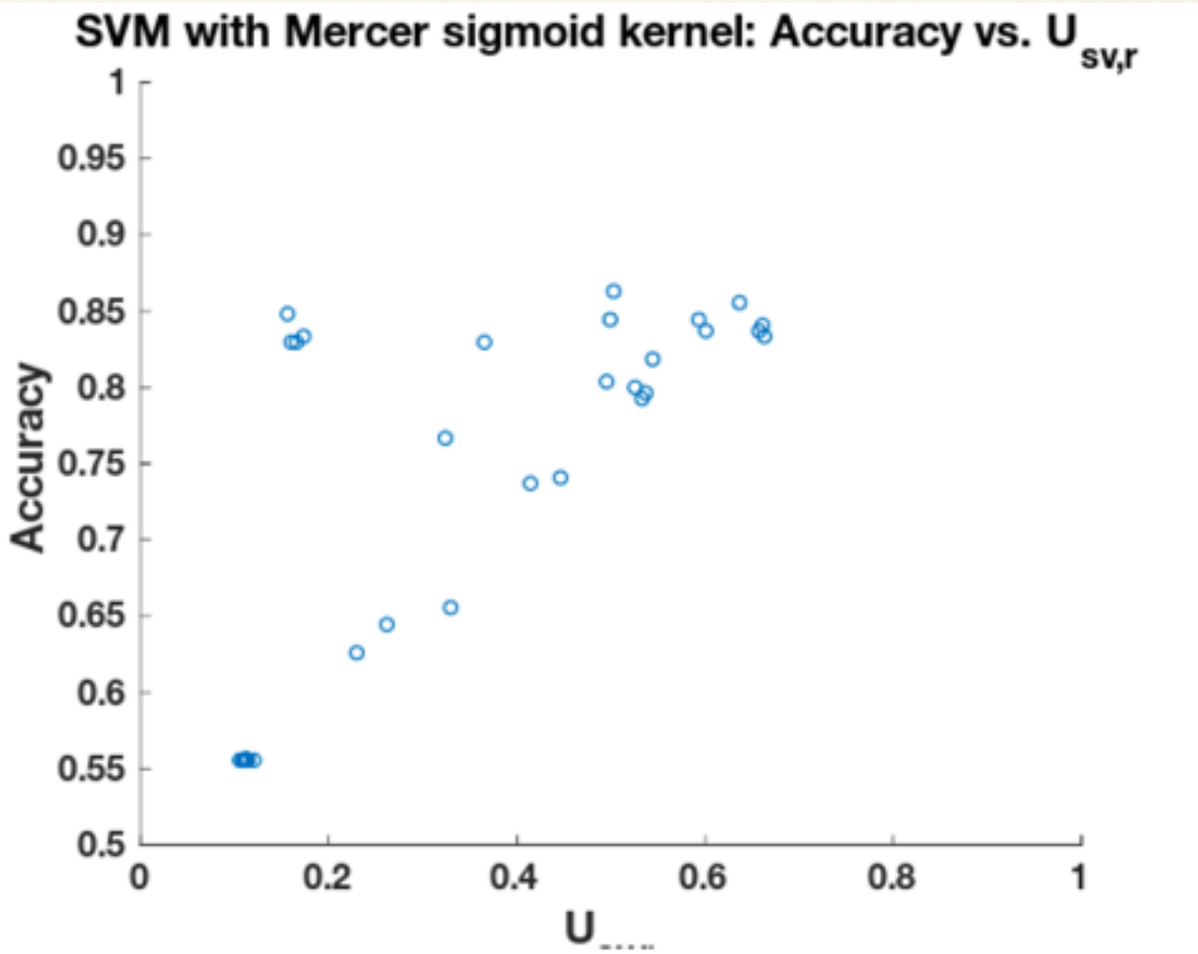
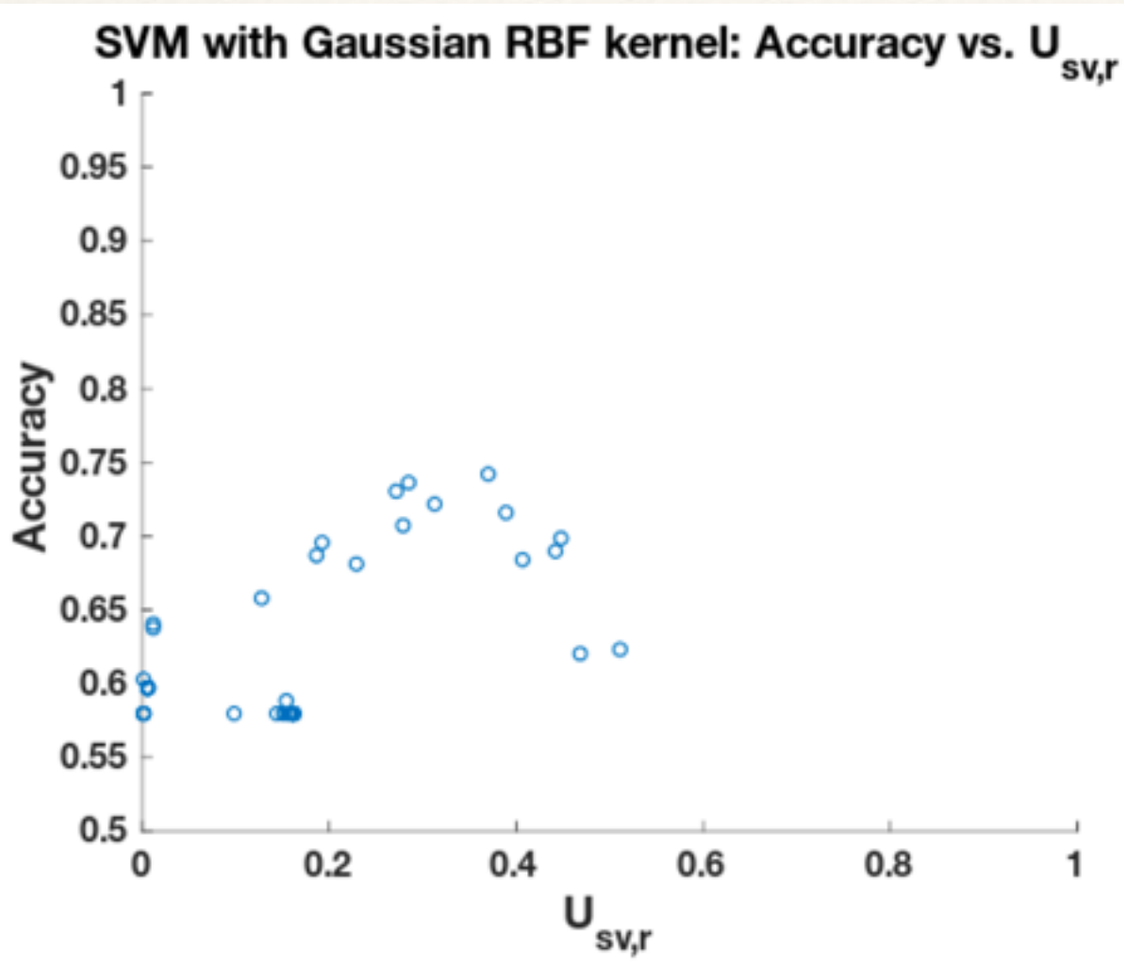


Figure 6.4: In classification with Statlog Heart data there are points with high accuracy and high inherent model interpretability, with minimal sacrifice, 1% and 2%, respectively.

Accuracy vs. interpretability 4 (U_{sv})



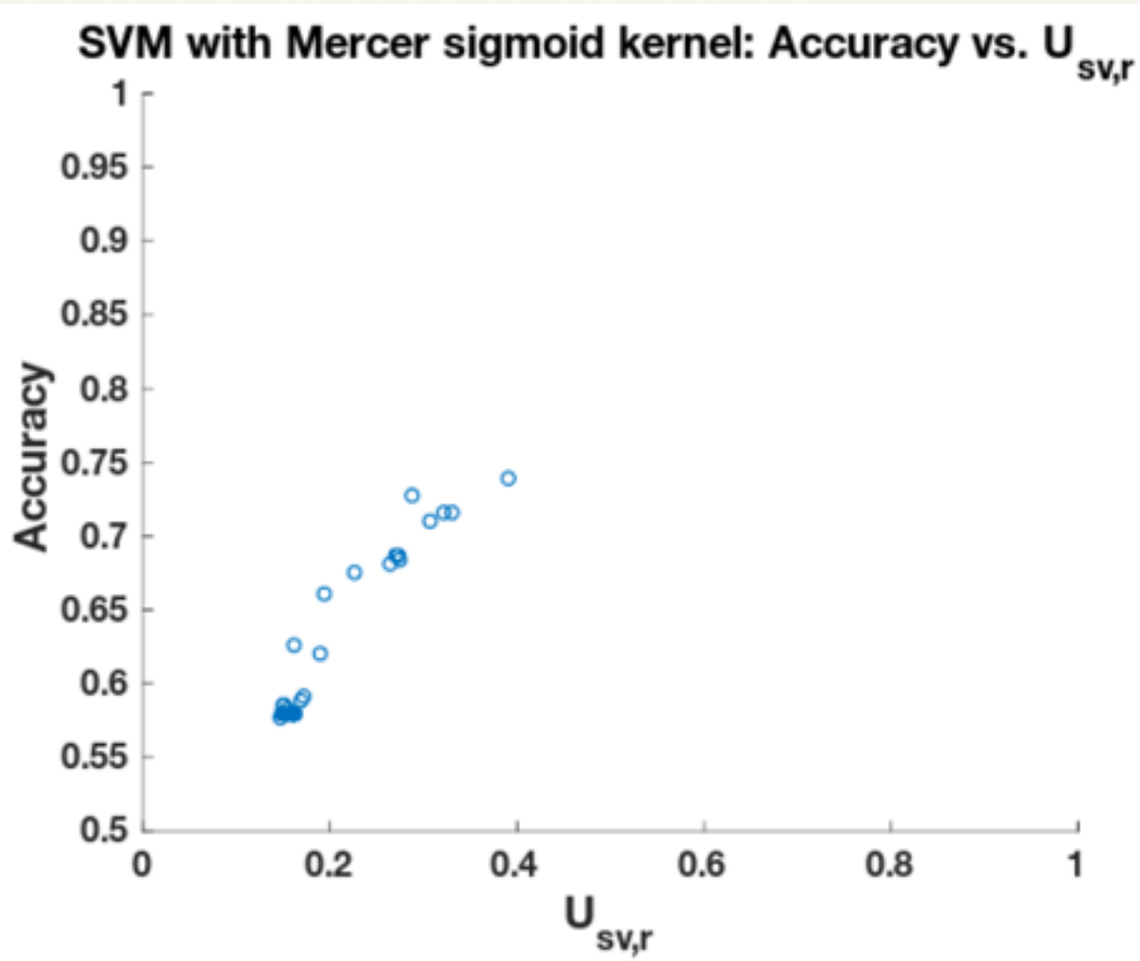


Figure 6.5: In classification with the Bupa liver data set there is a 20% and 0% sacrifice, respectively, in inherent model interpretability for the highest accuracy.

Conclusions

For binary classification in health care with atomic data types:

1. There is no accuracy / **transparency** trade-off **between** the Mercer sigmoid and the Gaussian RBF
    a. The Mercer sigmoid, as an explicit Mercer kernels is **more transparent**
    b. The Mercer sigmoid, as an explicit Mercer kernels is **at least as accurate**

Contributions

2. There is no accuracy versus inherent model interpretability trade-off within or between kernels for the Gaussian RBF and Mercer sigmoid kernels, using the support-vector measure

    a. Within kernel plots, there is no **overall** negative linear trend nor a negative exponential trend

    b. Between kernel plots, the kernels achieve similar accuracy and similar but slightly different interpretability

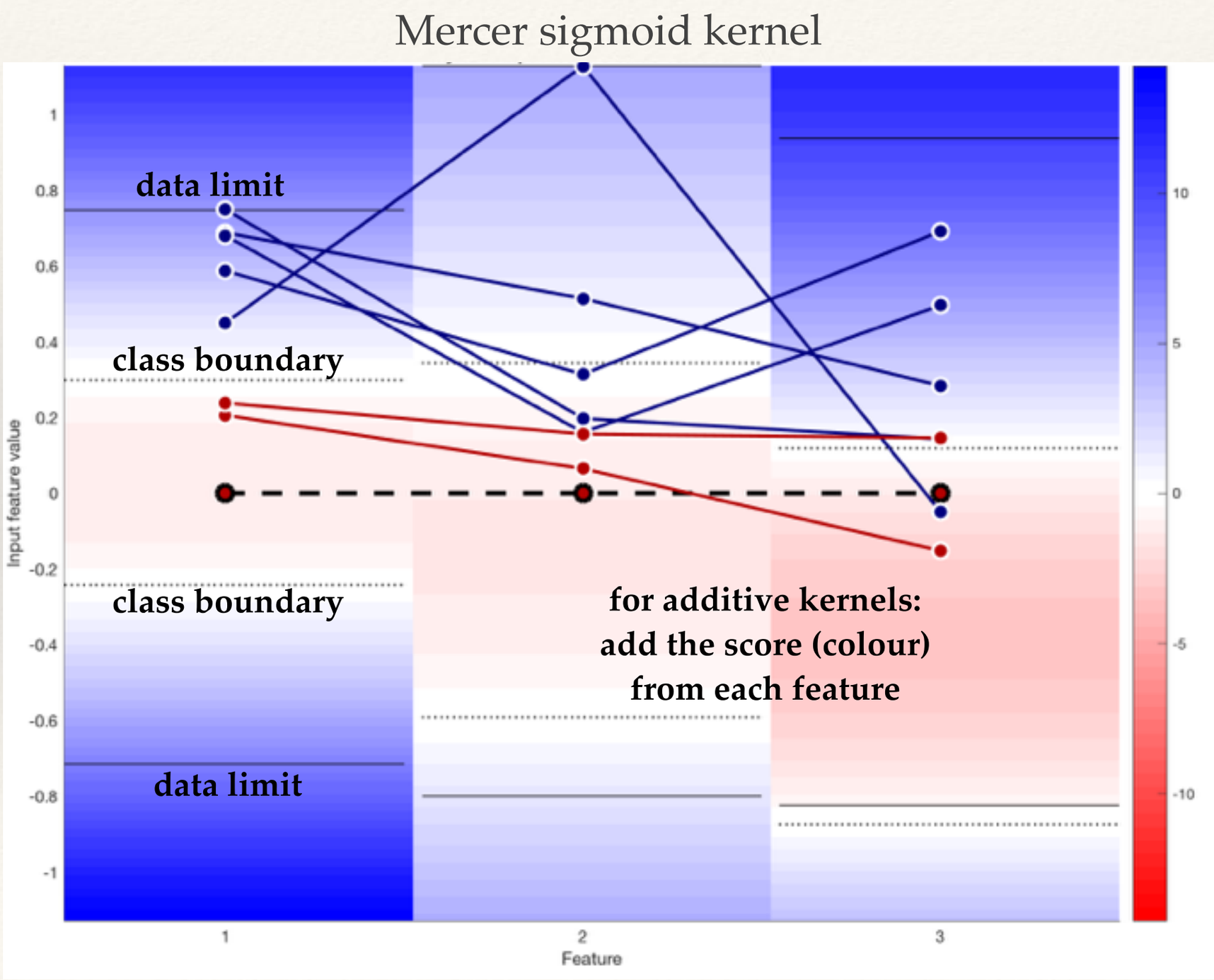
    c. Future work is needed to confirm a possible negative linear trend at the balanced portion of the optimal (pareto) front within plots

Contributions

3. I articulate new concepts and new quantitative measures of inherent model interpretability and transparency for model selection with accuracy, which can be automated.

   - Some measures are validated for use, some for limited use, and some are not yet validated.

Our proposed kernels do not confound features/columns



The End