36th Annual International Conference of the Engineering, Medicine and Biology Society (EMBC'14)

A new Mercer sigmoid kernel
for clinical data classification

André Carrington, M.Math, P.Eng.
Paul Fieguth, Ph.D.
Helen Chen, Ph.D.

WATERLOO | ENGINEERING
SYSTEMS DESIGN ENGINEERING

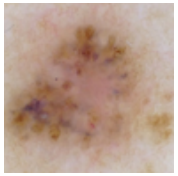
Agenda

- Context
- Problem (with existing kernel)
- Contribution (a new kernel)
- Comparison
- Test results
- Conclusions

Clinical data classification

Patient

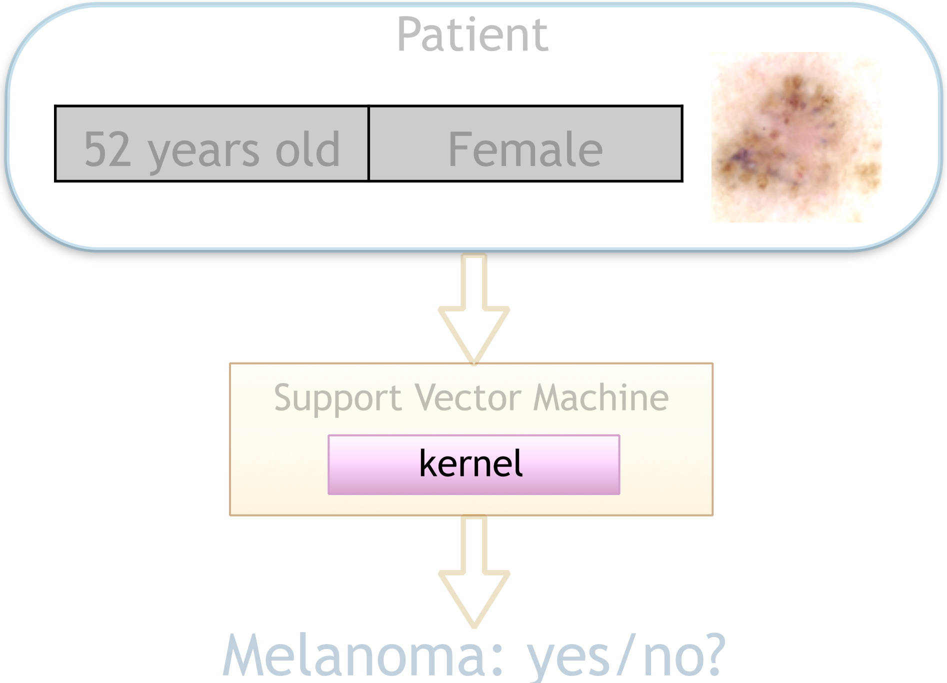
52 years old	Female



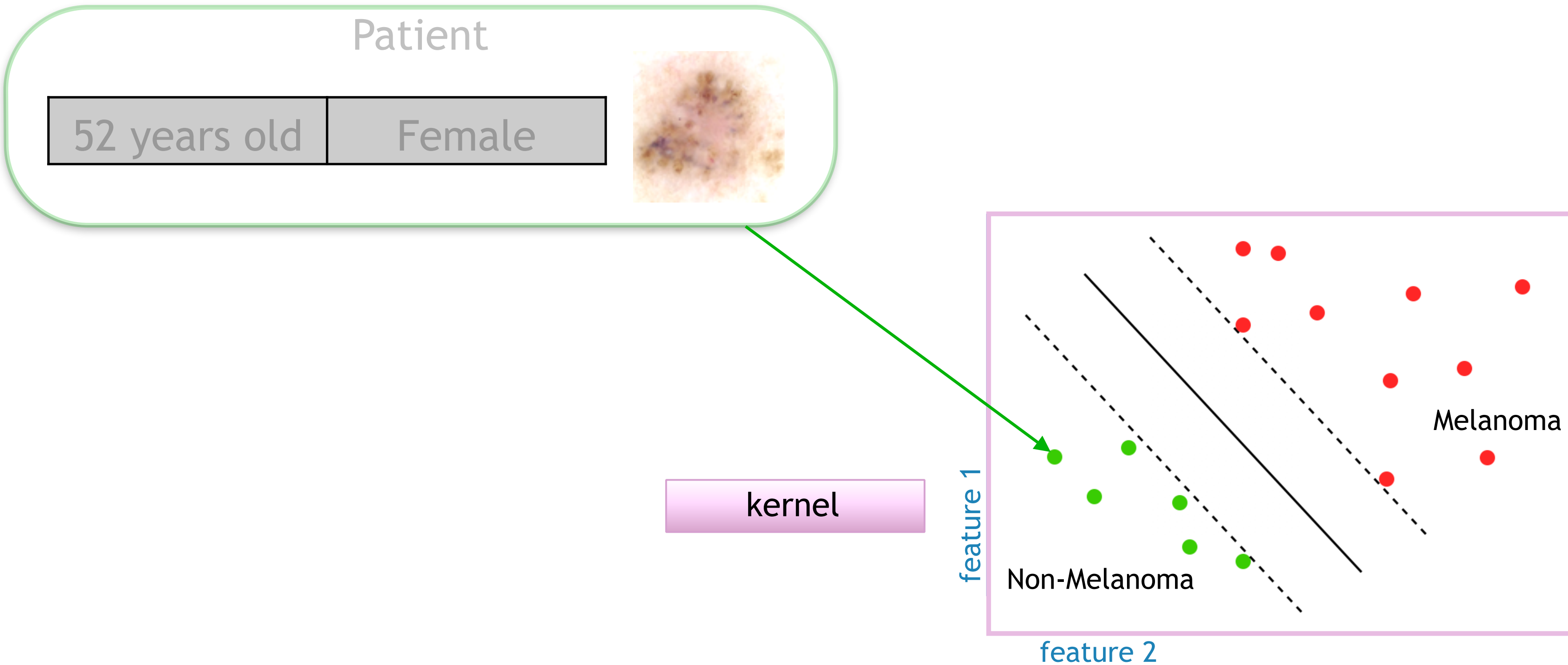
Support Vector Machine (SVM)

Melanoma: yes/no?

Kernels


Patient
52 years old
Female
Support Vector Machine
kernel
Melanoma: yes/no?

Feature-space geometry



Rightmost figure derived from Alexander Smola, Machine Learning Summer School, 2008.

Kernels

Formally, a kernel is

- a function of two inputs $\mathbf{x}, \mathbf{z} \in X$, mapped to a feature space *F*, $\phi : X \to F$
- an inner product for all inputs, $k\left(\mathbf{x}, \mathbf{z}\right) = \left\langle \phi\left(\mathbf{x}\right), \phi\left(\mathbf{z}\right)\right\rangle_F$
- real-valued and symmetric in its arguments, $k : X \times X \to \mathbb{R}$

Mercer or p.d. kernels

A kernel function is Mercer or positive definite (p.d.) if

$$\iint_{\mathcal{X}\times\mathcal{X}} \kappa(\mathbf{x},\mathbf{z})\, f(\mathbf{x})\; f(\mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z} \geq 0$$

$$\text{for all } \mathbf{x},\mathbf{z} \in \mathcal{X},\;\; f \in L_2(\mathcal{X}),\;\; \kappa \,\text{symmetric}$$

Mercer or p.d. kernels

A kernel function is Mercer or positive definite (p.d.) if

$$\iint\limits_{\mathcal{X}\times\mathcal{X}} \kappa(\mathbf{x},\mathbf{z})\, f(\mathbf{x})\; f(\mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z} \geq 0$$

$$\text{for all } \mathbf{x},\mathbf{z} \in \mathcal{X},\; f \in L_2(\mathcal{X}),\; \kappa \,\text{symmetric}$$

k is an inner product

Mercer's condition in matrix (finite) form, where kernel/Gram matrix G is p.s.d. if…

positive singular values and eigenvalues

convex optimization; a cone in the vector space of pxp matrices

Mercer or p.d. kernels

A kernel function is Mercer or positive definite (p.d.) if

$$\iint_{\mathcal{X}\times\mathcal{X}} \kappa(\mathbf{x},\mathbf{z})\, f(\mathbf{x})\, f(\mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z} \geq 0$$

$$\text{for all } \mathbf{x},\mathbf{z} \in \mathcal{X},\; f \in L_2(\mathcal{X}),\; \kappa \text{ symmetric}$$

SVMs are derived with positive definite (p.d.) kernels

The problem

The sigmoid kernel:

- is not: Mercer-compliant $\Leftrightarrow$ positive definite (p.d.)
- thus **not prima facie valid for SVM**

But it is used in health care!

- For 2011-2014, a search on “sigmoid kernel” AND “clinical” yields 33 and 451 hits on ScienceDirect and Google scholar
- 1 of 4 widely implemented kernels e.g. Matlab, R, SAS, SPSS, libsvm, Shogun, Orange, etc.
- Two fuzzy-logic (non-Mercer) sigmoid kernels were created in 2004 and 2006 to improve/replace it, but neither perform as well as the sigmoid kernel.

The sigmoid kernel

is conditionally positive definite (c.p.d.), but is that valid for SVM?

- Against - most lit': only Mercer kernels are valid
  - Smola: proves the sigmoid is not Mercer
- For - Boughorbel: $\underset{\alpha}{\operatorname{argmin}}\ 2*W(K_{cpd},\alpha/2) = \underset{\alpha}{\operatorname{argmin}}\ W(\tilde{K}_{pd},\alpha)$
  - Scholkopf: argues the sigmoid is valid

The sigmoid kernel

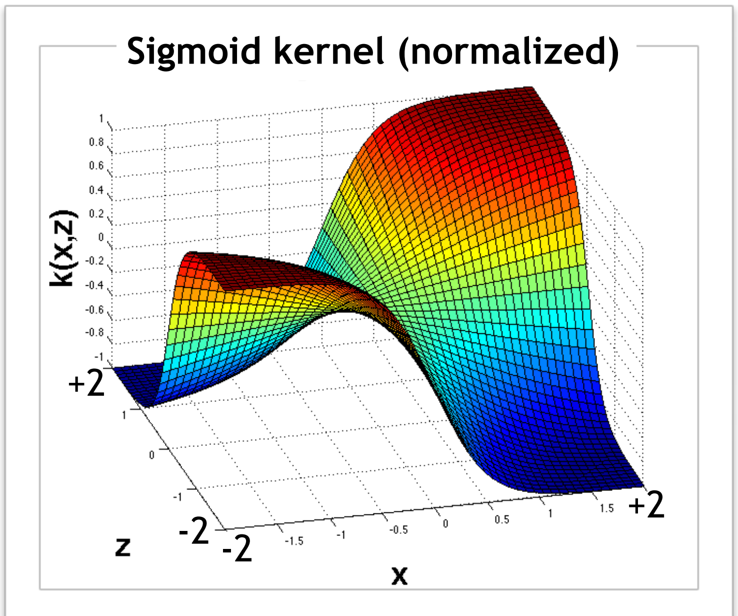
If invalid
- the optimality & stability of results are not guaranteed
- but health care applications need trustworthy results!

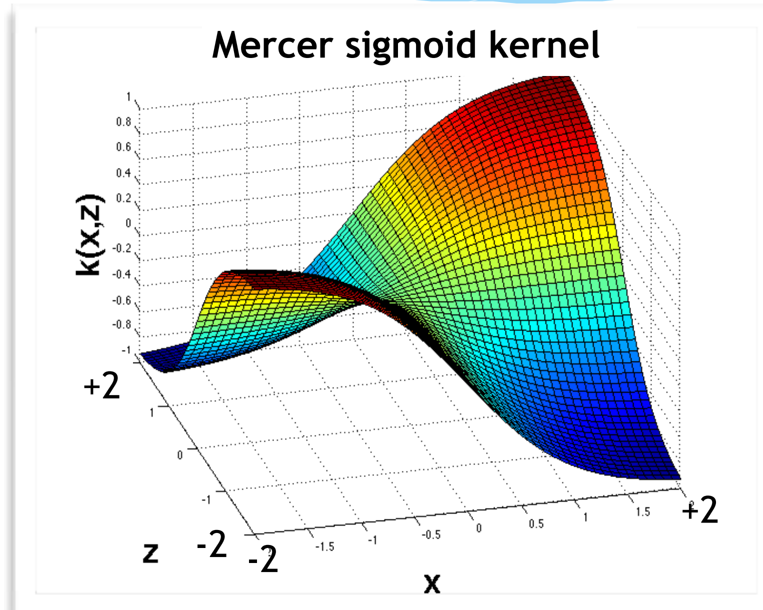
Even if valid
- it is only valid when *conditionally positive definite*, c.p.d. — hard to determine & data-dependent!
- it is only valid for specific kernel methods

Our new Mercer sigmoid kernel

- Is **similar to the sigmoid kernel**, Mercer-compliant, p.d., and **always valid for any kernel method.**
- Performs clinical data classification **significantly better** on 3 data sets vs. Gaussian RBF, linear, polynomial, sigmoid.
- Performs non-clinical data classification **about the same** as the Gaussian RBF, and better than others.

Comparison





$$k_S\left(\frac{\mathbf{x}}{\sqrt{p}}, \frac{\mathbf{z}}{\sqrt{p}}\right) = \tanh\left(\frac{a}{p} \cdot \mathbf{x}^T\mathbf{z} + r\right)$$

$$k_M\left(\mathbf{x}, \mathbf{z}\right) \triangleq \frac{1}{p}\sum_{i=1}^{p} \tanh\left(\frac{x_i - d}{b}\right) \cdot \tanh\left(\frac{z_i - d}{b}\right)$$

Comparison

$$\underbrace{\tanh(xz)}_{\text{form in sigmoid}} \approx \tanh(x)\cdot\underbrace{\tanh(z)}_{\text{form in Mercer sigmoid}}$$

dot-product kernel	vs.	separable kernel
non-Mercer	vs.	Mercer
infinite/implicit	vs.	finite/explicit feature space

< 10.1% RMS deviation

We tested kernels on six data sets

Experiment

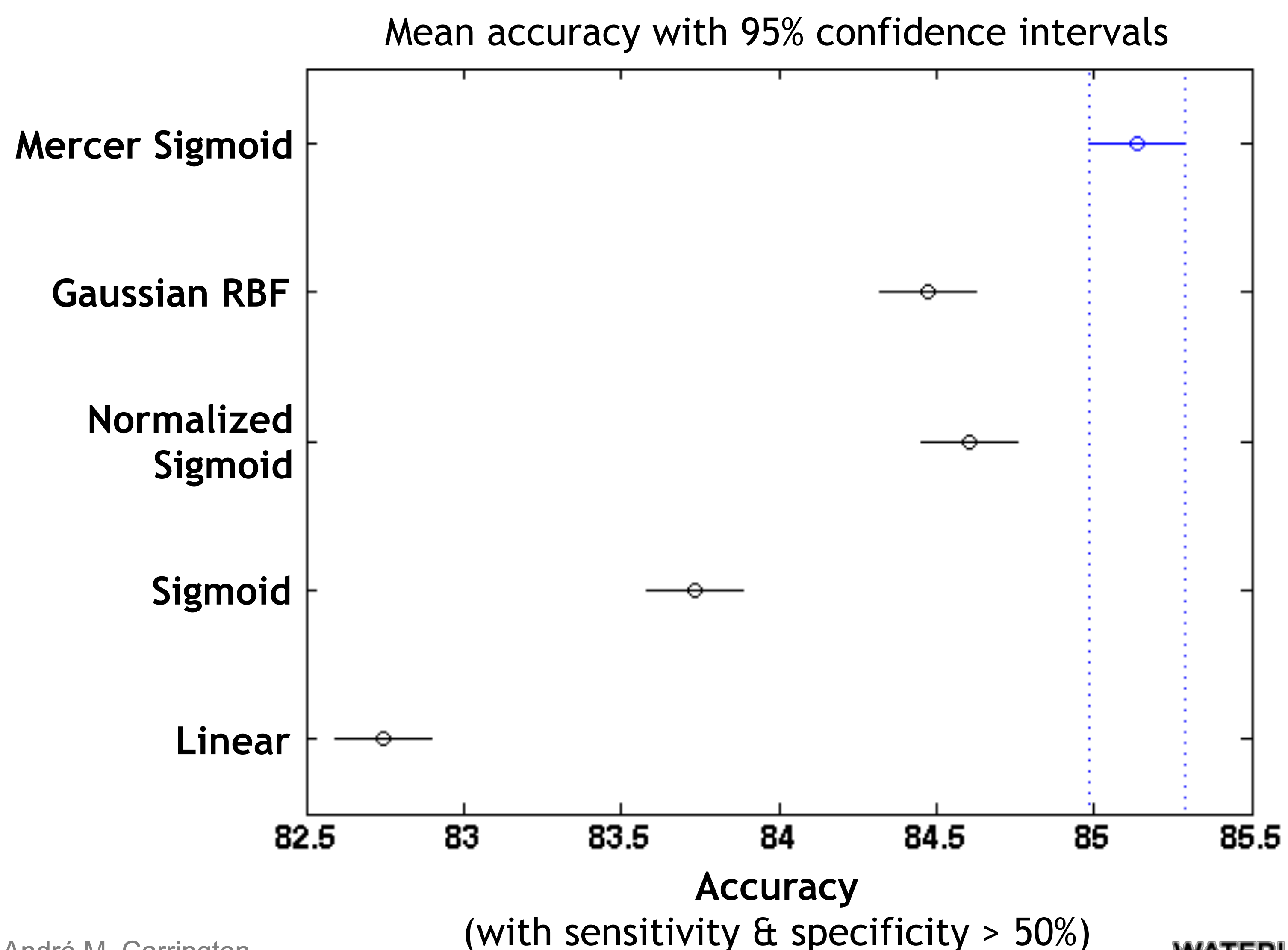
DATA SET SUMMARY

Data Set	Instances		Validation Method	Heterogeneous data types	Source
	Training	Validation			
Skin Lesion	57	57	5× 10-fold cross-validation	yes	Dr. Ehrsam
Heart	270	270	5× 10-fold cross-validation	yes	UCI
Diabetes	512	256	separate validation set	yes	UCI
Mushrooms	200*	200*	separate validation set	yes	UCI
Ionosphere	176	175	separate validation set	no	UCI
Sediment	1413	471	separate validation set	no	UCI

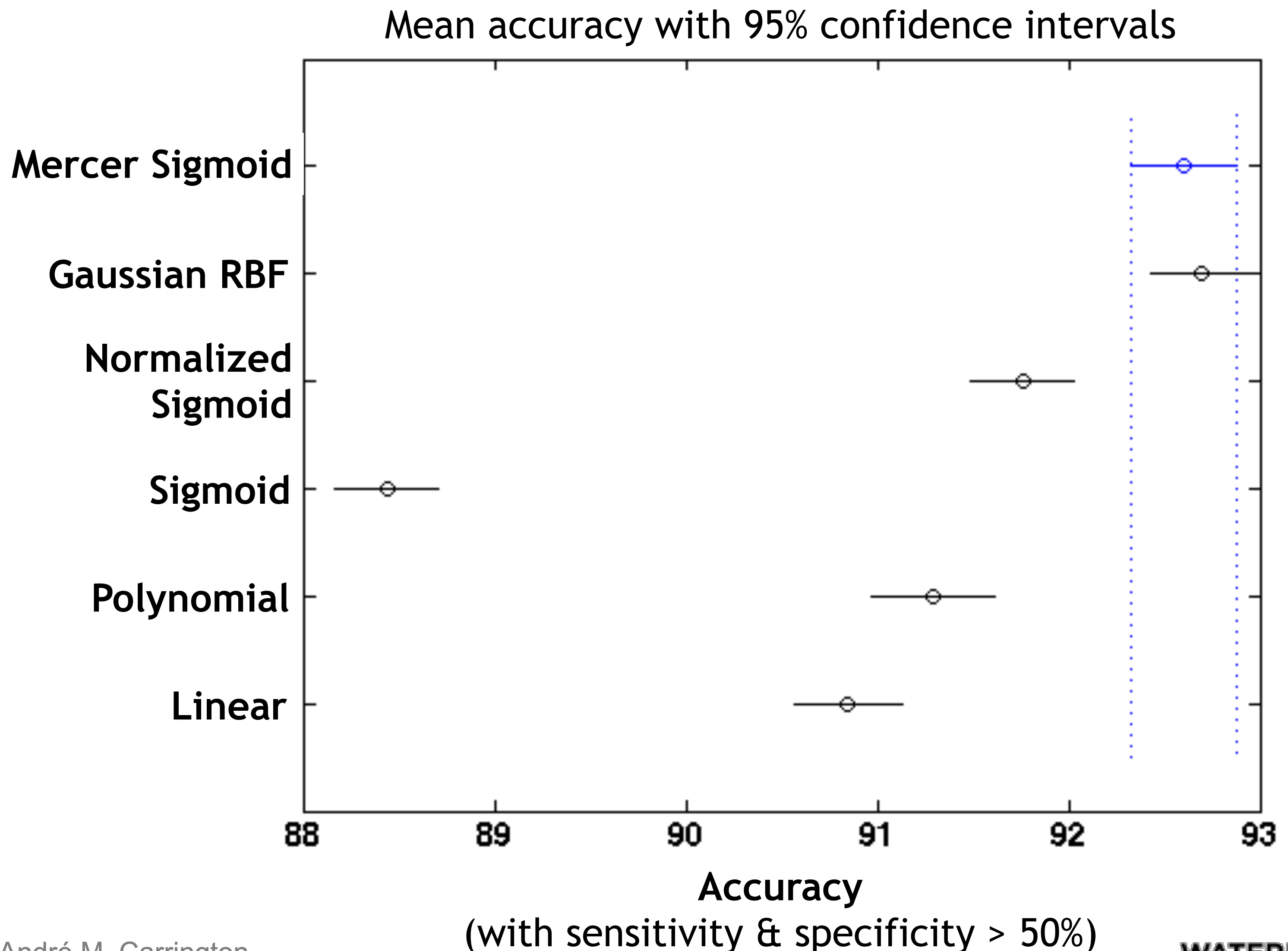
x 60 hyperparameters using random search (best result per kernel)

x 29 experiments (average of best results per kernel)

Clinical results



Non-clinical results



Conclusion

Since:

- The Mercer sigmoid kernel outperforms other kernels on 3 ***clinical*** data sets
- The existing sigmoid kernel lacks assurance
- SVM classifiers are developed with multiple candidate kernels

We recommend including the Mercer sigmoid kernel as a candidate for SVM classification of clinical data

Other benefits

- Our dot-product normalization of the existing sigmoid kernel is novel and significantly improves accuracy.
- The Mercer sigmoid kernel, as a separable kernel, is theoretically advantageous for big data. Platforms or tools made specifically for big data are required to exploit this.

Try it!

Download it from the VIP lab at Univ. Waterloo.

- Matlab script: example.m
  MSig.m
- Matlab C/MEX code: mexMSig.c

And please tell us about your results!

- amcarrin@uwaterloo.ca

Thank-you

amcarrin@uwaterloo.ca

Appendix A: references, FAQ, experimental limitations

References

1. Bishop, Christopher M. Pattern recognition and machine learning, volume 4. Springer New York, 2006.
2. Boughorbel, Sabri, J-P Tarel, and Nozha Boujemaa. Conditionally positive definite kernels for SVM based image recognition. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pages 113–116. IEEE, 2005.
3. Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2, no. 2 (1998): 121-167.
4. Camps-Valls, Gustavo, José David Martín-Guerrero, José Luis Rojo-Álvarez, and Emilio Soria-Olivas. "Fuzzy sigmoid kernel for support vector classifiers". Neurocomputing (2004): 62:501–506.
5. Cristianini, Nello and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
6. Delaitre, Vincent, et al. "Scene semantics from long-term observation of people." Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012. 284-298.
7. Doloc-Mihu, Anca. "Kernel method for improving image retrieval performance: a survey." International Journal of Data Mining, Modelling and Management 3.1 (2011): 42-74.
8. Genton, Marc G. "Classes of kernels for machine learning: a statistics perspective." The Journal of Machine Learning Research 2 (2002): 299-312.
9. Han, Liu, Liu Ding, and Deng Ling-Feng. Chaotic time series prediction using fuzzy sigmoid kernel-based support vector machines. Chinese Physics, 15(6):1196, 2006.
10. Hein, Matthias, and Olivier Bousquet. "Hilbertian metrics and positive definite kernels on probability measures." (2004).
11. Jebara, Tony, Risi Kondor, and Andrew Howard. "Probability product kernels." The Journal of Machine Learning Research 5 (2004): 819-844.

References

12. Jebara, Tony, and Risi Kondor. "Bhattacharyya and expected likelihood kernels." Learning Theory and Kernel Machines. Springer Berlin Heidelberg, 2003. 57-71.
13. Lin, Hsuan-Tien and Chih-Jen Lin. “A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods”. Neural Computation (2003): 1-32.
14. Maji, Subhransu, Alexander C. Berg, and Jitendra Malik. "Classification using intersection kernel support vector machines is efficient." In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1-8. IEEE, 2008.
15. Scholkopf, Bernhard. "The kernel trick for distances". Advances in neural information processing systems, pages 301–307, 2001.
16. Shawe-Taylor, John, and Nello Cristianini. Kernel methods for pattern analysis. Cambridge university press, 2004.
17. Tran, Huy Dat, and Haizhou Li. "Probabilistic distance SVM with Hellinger-exponential kernel for sound event classification." Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.
18. Vandebril, Raf, Marc Van Barel, and Nicola Mastronardi. Matrix computations and semiseparable matrices: linear systems. Vol. 1. JHU Press, 2007.
19. Vedaldi, Andrea and Andrew Zisserman. "Efficient additive kernels via explicit feature maps." Pattern Analysis and Machine Intelligence, IEEE Transactions on 34.3 (2012): 480-492.
20. Zhang, Li, Weida Zhou, and Licheng Jiao. "Wavelet support vector machine." IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34.1 (2004): 34-39.

Frequently Asked Questions (FAQ)

- Q: Why is the Mercer sigmoid kernel significantly more accurate in classification with the 3 clinical data sets? What is the cause? What differentiates them from non-clinical data?
  - A: The result may not generalize to other data sets since confounding factors may apply (i.e. Simpson's paradox). Note: it is also best in preliminary results with real-life nephrology data. We have a hypothesis regarding the cause in other work underway.
- Q: Does 1/p in the Mercer sigmoid kernel have any effect in SVM?
  - A: No, it has no effect with SVM. 1/p is included so that comparison (RMS deviation) with the sigmoid kernel is meaningful. The output is also more intuitive for users.

Mercer sigmoid FAQ 2

- Q: The Mercer sigmoid kernel has an explicit basis function and finite dimensional feature space. Aren't kernels usually defined with an implicit basis function and an infinite dimensional feature space? Is that the kernel trick?
  - A: The kernel trick is the ability to replace XX^T in the SVM dual with any kernel (explicit or implicit). There are 9 other explicit and separable kernels in the literature.
  - A: Also, infinite dimensional feature spaces are overrated for SVM:
    - The significance of terms in a Taylor series taper off very quickly. The effect of the latter terms in optimization (re error) is negligible.
    - Classifiers are based on a finite number of support vectors or SV (less than the number of instances) — i.e. finite complexity, as appropriate for generalization.
    - When the Mercer sigmoid performs better, it uses 35% less SV.

Mercer sigmoid FAQ 3

- Q: Does the kernel only partition data into two regions per dimension?
  - A: Yes. If data is trimodal within one dimension, other dimensions are completely independent, and the data is not separable in other dimensions, then this kernel is sub-optimal.
- Q: How can it perform better than the Gaussian RBF kernel?
  - A: Future work may investigate this. Lin and Lin noted a close relationship between the sigmoid kernel and the Gaussian RBF kernel, where the former is c.p.d.
- Q: Have advantages of (i.e. specific applications for) the *pre-existing sigmoid kernel* been identified?
  - A: No.

Mercer sigmoid FAQ 4

- Q: The RMS deviation < 10.1% is specific to a=[0.1,10] for the normalized sigmoid kernel, where a is the maximum slope. Some literature recommends a=1/N for the sigmoid kernel.
  - A: While we learned about a=1/N after-the-fact, I posit that smaller values of a would have a negligible effect. At a=0.1, both a sigmoid kernel and a normalized sigmoid kernel look like a near-horizontal plane.
- Q: Are there any disadvantages with using the Mercer sigmoid kernel?
  - When using random search for hyperparameters in our experiment, the Mercer sigmoid kernel had a higher standard deviation in its accuracy than the Gaussian RBF kernel. As a result one has to use more hyperparameters than the Gaussian RBF kernel for best results.

Experimental limitations

- We allowed Sequential Minimal Optimization (SMO) to violate Karush Kuhn Tucker (KKT) conditions for kernels where it was not needed
- 52% of our skin data used the clinician (vs. pathology) as ground-truth
- We used balanced costs, although melanoma detection should use an imbalanced cost
- None of the 6 data sets had a large number of instances (i.e. big data)
- We tested the Mercer sigmoid kernel with SMO not QP.
- We did not use a final test set after cross-validation

Appendix B: other details (following the presentation order)

Mercer or p.d. kernels

A kernel function is Mercer or positive definite (p.d.) if

$$\iint_{\mathcal{X}\times\mathcal{X}} \kappa(\mathbf{x},\mathbf{z})\, f(\mathbf{x})\; f(\mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z} \geq 0$$

for all $\mathbf{x},\mathbf{z} \in \mathcal{X},\; f \in L_2(\mathcal{X}),\; \kappa$ symmetric

$$\Updownarrow$$

A kernel matrix is p.s.d. if

$$\sum_{j,k=1}^{n} \kappa(x_j, x_k)\, c_j\, c_k \geq 0$$

for all $\mathbf{x}_j, \mathbf{x}_k \in \mathcal{X},\; c_j\, c_k \in \mathbb{R}$,
κ symmetric, x_j, $n \geq 1$, $\mathcal{X}$ non-empty

c.p.d. kernels

A kernel matrix is c.p.d. if

$$\sum_{j,k=1}^{n} \kappa(x_j, x_k)\, c_j\, c_k \ \geq 0$$

for all $\mathbf{x}_j, \mathbf{x}_k \in \mathcal{X},\ c_j\, c_k \in \mathbb{R}, \sum_{j=1}^{n} c_j = 0$

κ symmetric, x_j, $n \geq 1$, $\mathcal{X}$ non-empty

Proof of Mercer compliance for our Mercer sigmoid kernel

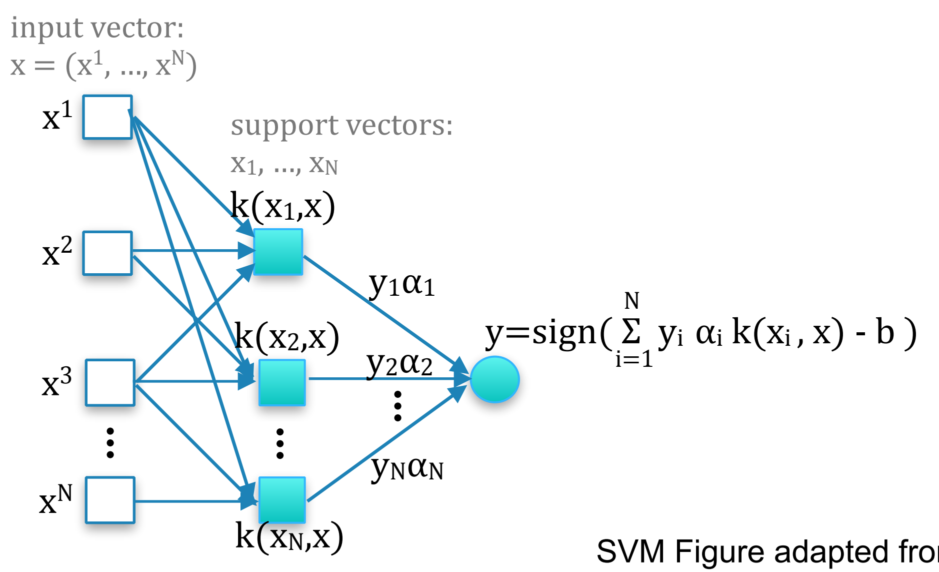
A valid kernel, a kernel that is positive semidefinite and symmetric, and a Mercer kernel are equivalent conditions and terms. We therefore seek a valid kernel to ensure Mercer compliance.

- $k(\mathbf{x},\mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$ is a valid kernel for real-valued f(.) on X, $X \subseteq R^p$ (1)
- $k(\mathbf{x},\mathbf{z}) = k_1(\mathbf{x},\mathbf{z}) + k_2(\mathbf{x},\mathbf{z})$ is a valid kernel if k_1 and k_2 are valid kernels (2)
- Let $f(x_i)=(1/\surd p)\tanh((x_i-b)/d)$ where x_i and f are real-valued (3)
- From (1,3): $k(x_i,z_i)=(1/\surd p)\tanh((x_i-b)/d)*(1/\surd p)\tanh((z_i-b)/d)$ is valid (4)

From (2,4): $k_M(\mathbf{x},\mathbf{z})=\Sigma_i\, k(x_i,z_i) = \Sigma_i\, (1/\surd p)\tanh((x_i-b)/d) * (1/\surd p)\tanh((z_i-b)/d)$ is a valid kernel, which is our proposed kernel.

Why is the sigmoid kernel widely implemented?

- Current SVMs (i.e. soft-margin) were introduced in 1995 by Boser, Guyon and Vapnik.
- Vapnik, the creator of SVM, published a book in 1995 discussing 3 kernels (or SVM types): polynomial, radial basis function, or two layer neural networks [i.e. sigmoid]: $k(\boldsymbol{x},\boldsymbol{z}) = S(a\boldsymbol{x}^T\boldsymbol{z} + r)$.



SVM Figure adapted from Vapnik

- He asserts that all 3 can approximate a continuous function to any degree of accuracy (p.155).

The sigmoid kernel dispute

- **Boughorbel et al**: Show that using a p.d. kernel with the SVM dual `argmin`$_\alpha$ `W'(`α`)` is equivalent to using its related c.p.d. kernel with `2*argmin`$_\alpha$ `W(`α`/2)`. Conclude c.p.d. kernels valid for SVM.
- **Scholkopf**: Argues (does not show) that a c.p.d. kernel is valid for SVM; shows it is valid for kernel PCA.

The sigmoid kernel dispute

- **Boughorbel et al:** Show that using a p.s.d. kernel with the SVM dual `argmin`$_\alpha$ `W'(α)` is equivalent to using its related c.p.d. kernel with `2*argmin`$_\alpha$ `W(α/2)`. Conclude c.p.d. kernels valid for SVM.
- **Scholkopf:** Argues (does not show) that a c.p.d. kernel is valid for SVM; shows it is valid for kernel PCA.
- **Smola et al:** Impossible to use the kernel (with *a*=1) for SVM. Does not satisfy Mercer's condition for any parameter values.
- **Most literature:** Only p.d. (Mercer) kernels are valid for SVM.

Even if valid…

- ***It is only valid for a specific range of parameters that is difficult to determine & data-dependent!***
  - A range that is proven to exist. Solving *r* for your data, is analytically difficult, but you can check a proposed *r*.
  - Lin et al: r must be small. $\{a>0, r<0\}$ is the most suitable quadrant (corroborated by implementations).
  - Burges: There are 3 conditions for dot-product kernels to be p.d. (not c.p.d.) including a data-dependent range of *a* & *r*.

Separable kernels

Separable kernels are explicitly of the form $k(\boldsymbol{x},\boldsymbol{z})=f(\boldsymbol{x})g(\boldsymbol{z})$. They have **lower space complexity**, since only the vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ need to be stored instead of the kernel matrix. Genton reduces some kernels to separable kernels and asserts benefits for big data.

Other separable kernels include:

- Linear kernel
- Hellinger kernel
- Hellinger exponential kernel
- Wavelet kernel
- Generalized histogram intersection kernel
- Chi-square kernel
- Probability product kernel
- Bhattacharyya kernel
- Expected Likelihood kernel

Experimental method

- We randomly generate hyperparameters from a uniform distribution (i.e. random search)

	Kernel						SVM	
	Poly	RBF	Sig		MSig			
Limit	d	$\log \sigma$	a	r	b	d	$\log C$	kkt
Lower	2	-1	ε*	-5	$\frac{1}{\sqrt{a}}$	-2	-1	0
Upper	7	3	10	$-\varepsilon$		+2	3	1

log = $\log_{10}$

*a lower = 0.1 for RMS Δ

Features

Dataset	Features	Images	Real	Count	Binominal	Nominal	Ordinal
Skin Lesion	101	no	1	1	97	2	0
	126	no	1	1	122	2	0
	133	yes	33	1	97	2	0
	158	yes	33	1	122	2	0
Heart	20	no	4	2	3	10	1
Diabetes	8	no	7	1	0	0	0
Mushrooms	112	no	0	1	4	107	0
Ionosphere	33	no	33	0	0	0	0
Sediment	9	no	9	0	0	0	0

clinical

non-clinical

hetero-geneous