

A new Mercer sigmoid kernel for clinical data classification

André Carrington, M.Math, P.Eng.

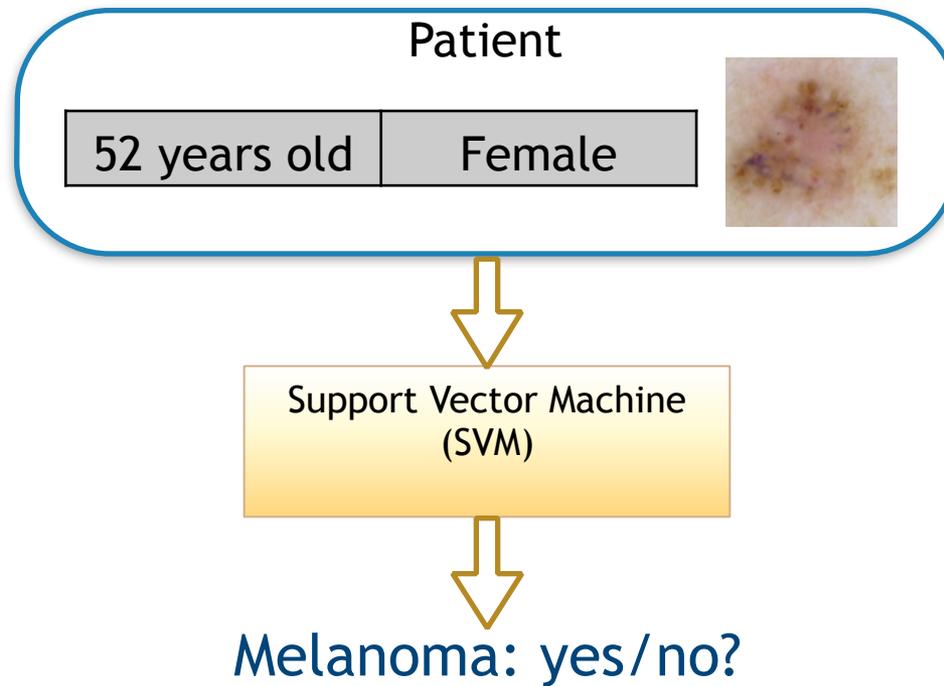
Paul Fieguth, Ph.D.

Helen Chen, Ph.D.

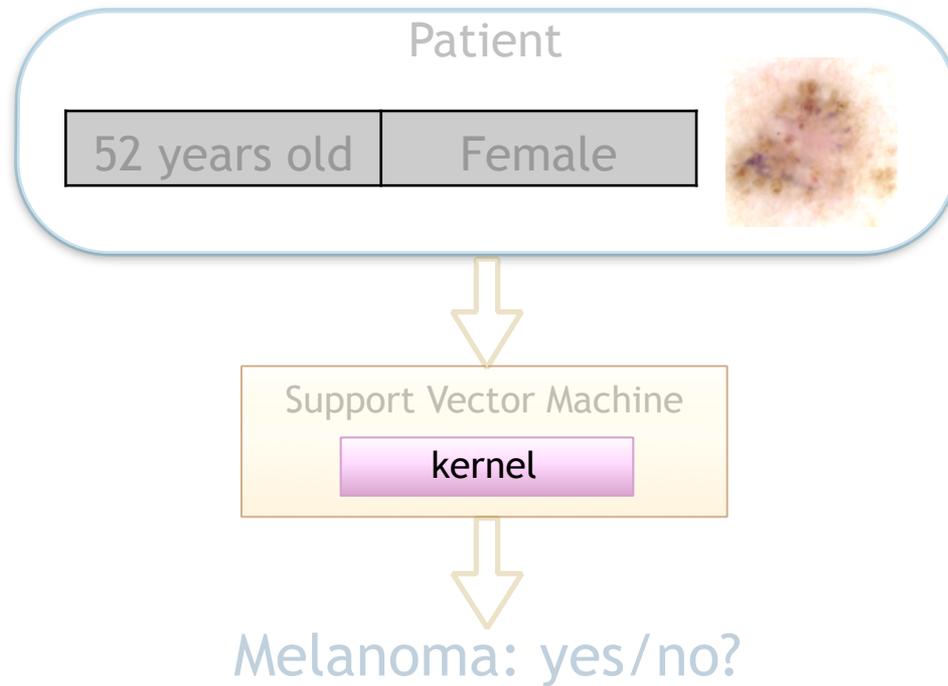
Agenda

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

Clinical data classification



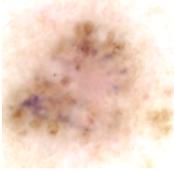
Kernels



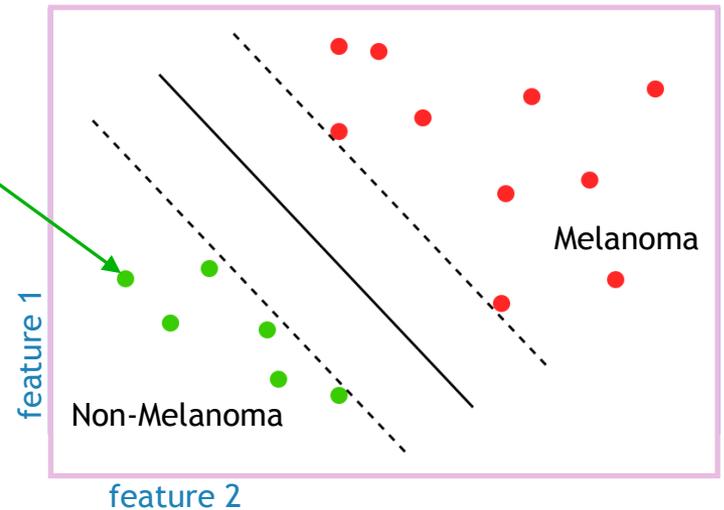
Feature-space geometry

Patient

52 years old	Female
--------------	--------



kernel



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

The problem

The sigmoid kernel:

- is not Mercer, i.e. positive definite (p.d.), and
- thus **not prima facie valid** for Support Vector Machines (SVM)
 - because the derivation of kernel SVM uses matrix algebra that is only valid for p.d. kernels

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.

Kernels

Formally, a kernel k is

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

Mercer's theorem

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

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

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

Mercer or p.d. kernels

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

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

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



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 conditionally positive definite (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

The sigmoid kernel is not p.d.

It is conditionally positive definite (c.p.d.) in a certain range.

- **Bouhonorbel et al:** Show that using a p.d. kernel with the SVM dual $\arg\min_{\alpha} W'(\alpha)$ is equivalent to using its related c.p.d. kernel with $2 * \arg\min_{\alpha} W(\alpha/2)$. Conclude c.p.d. kernels are 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.

valid



The sigmoid kernel is not p.d.

It is conditionally positive definite (c.p.d.) in a certain range.

- **Bouhonorbel et al:** Show that using a p.s.d. kernel with the SVM dual $\arg\min_{\alpha} W'(\alpha)$ is equivalent to using its related c.p.d. kernel with $2 * \arg\min_{\alpha} W(\alpha/2)$. Conclude c.p.d. kernels are valid for SVM. valid
✓
- **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. ?

The sigmoid kernel

If invalid

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

Even if c.p.d. kernels are valid

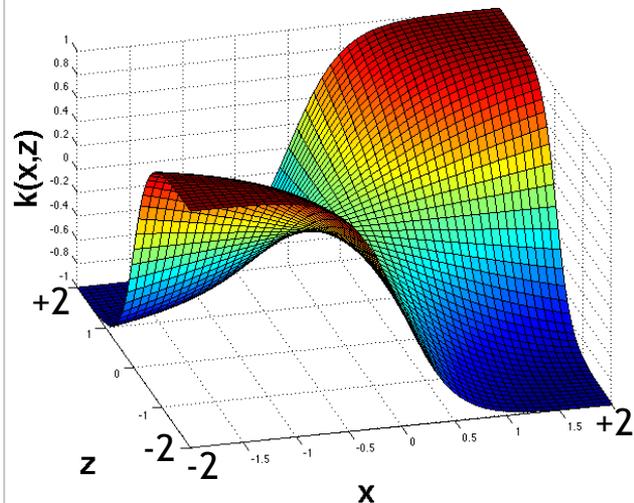
- the sigmoid is only c.p.d. in a certain range (hard to determine & data-dependent)!
- they are only valid for kernel methods with proofs of equivalence

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.

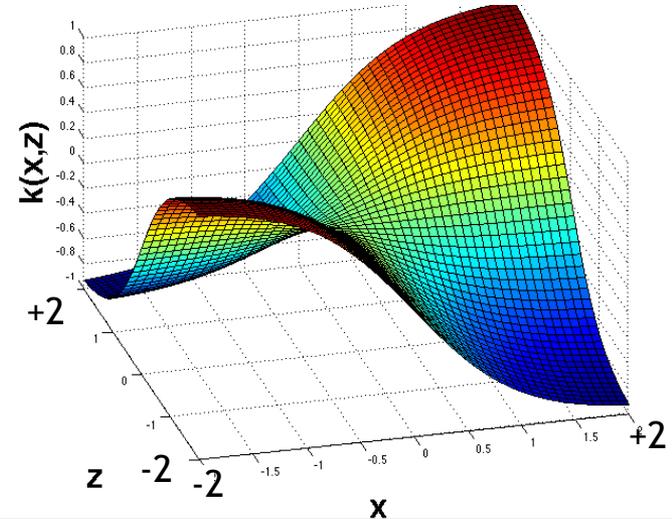
Comparison

Sigmoid kernel (normalized)



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

Mercer sigmoid kernel



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

Comparison

$\tanh(xz)$

form in sigmoid

\approx

$\tanh(x) \cdot \tanh(z)$

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

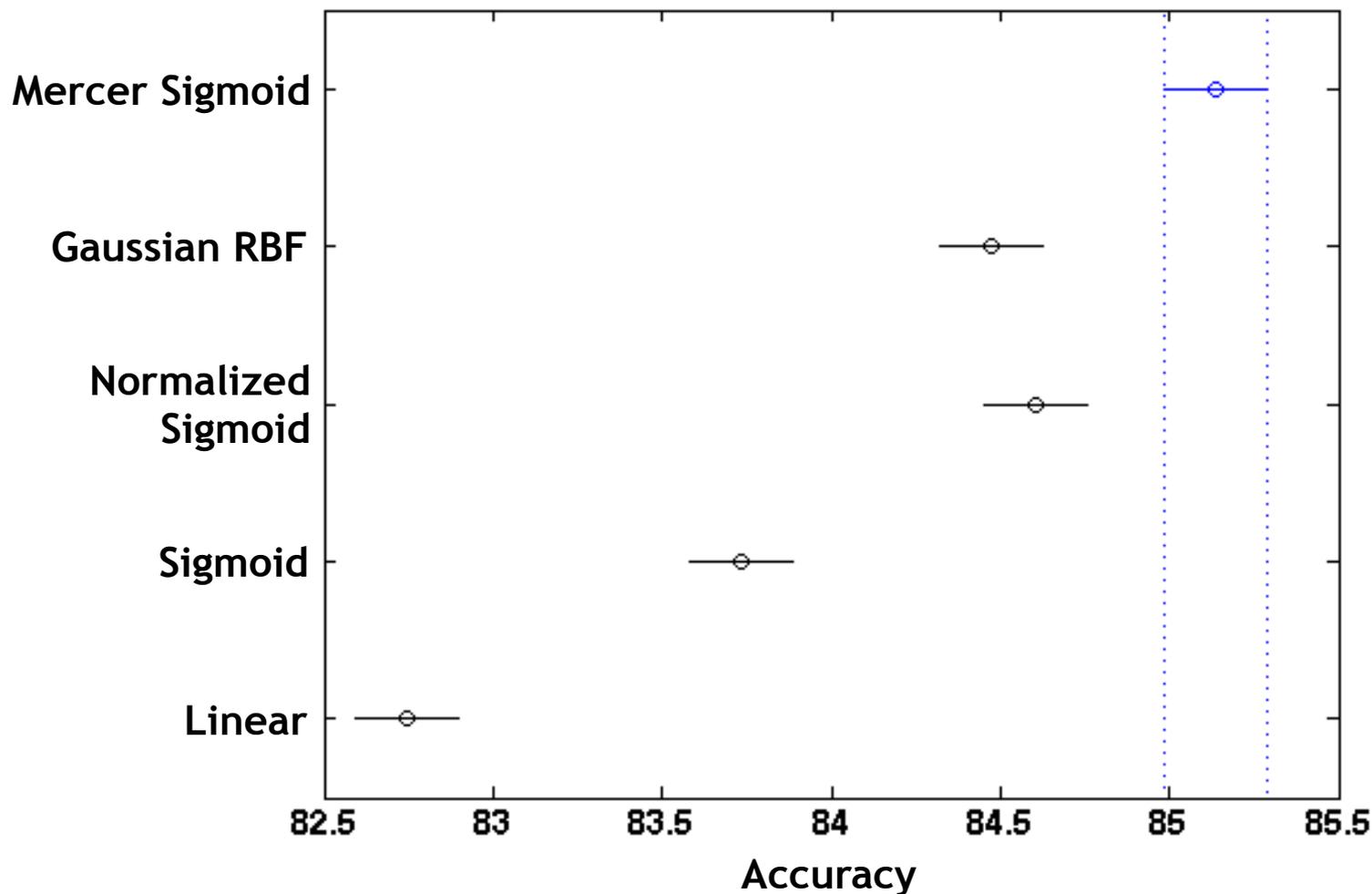
Data Set	Instances		Validation Method	Heterogeneous data types	Source
	Training	Validation			
Skin Lesion	57	57	5× 10-fold cross-validation	yes	Dr. Ehrsam
Heart	270	270	5× 10-fold cross-validation	yes	UCI
Diabetes	512	256	separate validation set	yes	UCI
Mushrooms	200*	200*	separate validation set	yes	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)

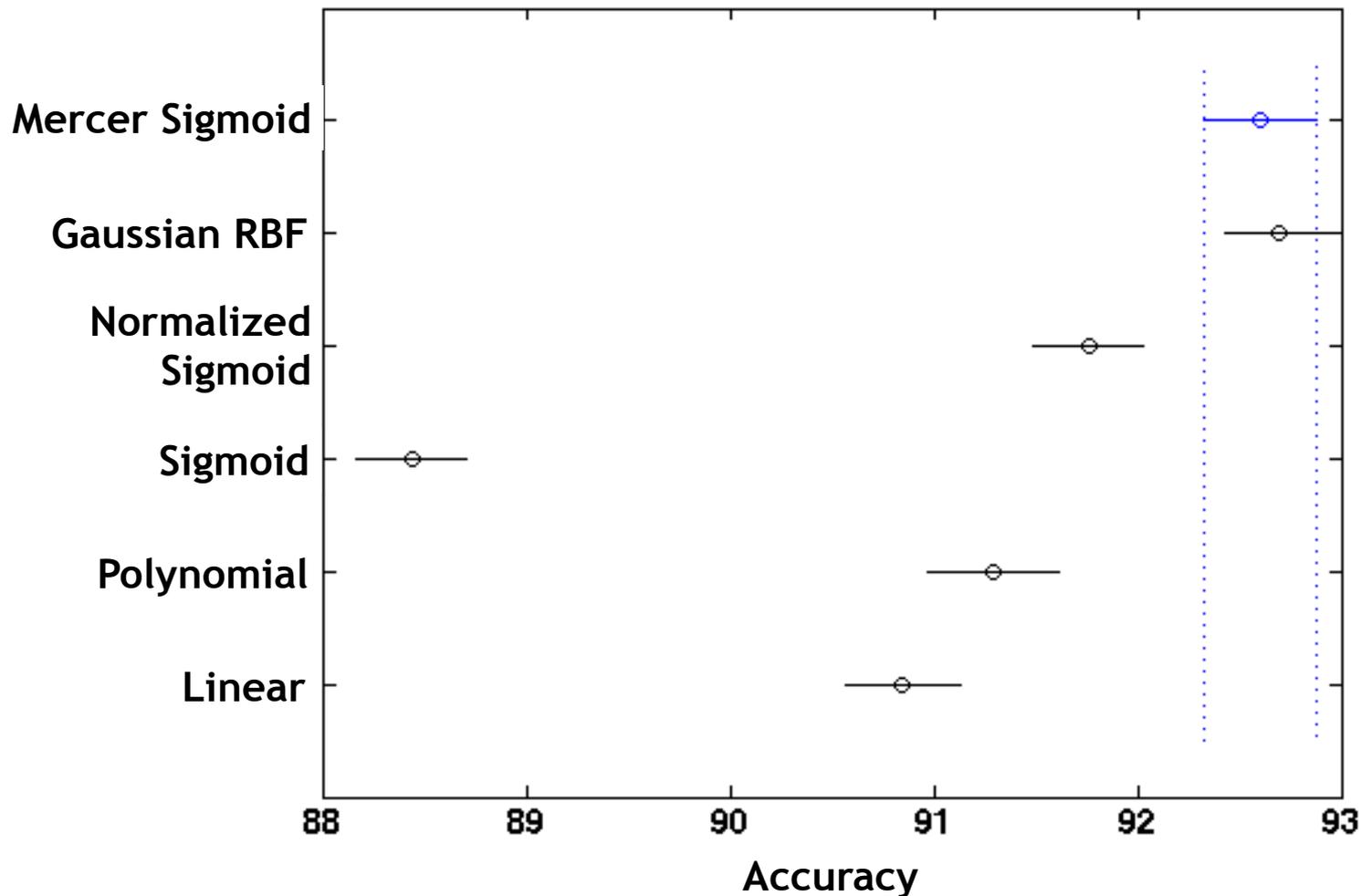
Clinical results

Mean accuracy with 95% confidence intervals



Non-clinical results

Mean accuracy with 95% confidence intervals



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 *and is similar to obscure guidance found in the literature to fix $a=1/n$.*
- 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 the zip file from [my webpage](#).

- 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.
21. Leslie, Christina S., Eleazar Eskin, and William Stafford Noble. "The spectrum kernel: A string kernel for SVM protein classification." Pacific symposium on biocomputing. Vol. 7. 2002.
22. Leslie, Christina S., Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. "Mismatch string kernels for discriminative protein classification." Bioinformatics 20, no. 4 (2004): 467-476.
23. Rätsch, Gunnar, and Sören Sonnenburg. "13 Accurate Splice Site Detection for Caenorhabditis elegans." Kernel Methods in Computational Biology (2004): 277.

Answers to expected questions

- 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 root cause is the subject of future work. The importance of this work is a Mercer version of the sigmoid kernel. Its impressive result with clinical data *may* not generalize further, since confounding factors *may* apply (i.e. Simpson's paradox). That said, it is also best in preliminary classification results with a nephrology data set.
- Q: Have advantages of the pre-existing sigmoid kernel been identified?
 - A: While it has been applied to many problems, we are not aware of any general guidance on its applicability to data or tasks, other than Vapnik's original assertion in 1995 (paraphrased) that in SVM it is a universal approximator.

Answers to expected questions 2

- Q: Are there any disadvantages with using the Mercer sigmoid kernel?
 - A: The Mercer sigmoid kernel has a **higher standard deviation** than the Gaussian RBF kernel. It has a larger hyperparameter search space with three hyperparameters compared to two in the sigmoid kernel or one in the Gaussian RBF and polynomial kernels. We use a random (not grid) search over 60 points in each validation run.
 - A: It partitions data into two regions per feature (dimension). If the data are trimodal within one feature and other features are completely independent, and the data are not separable in other features, then this kernel would be sub-optimal.
 - A: With imbalanced classes, e.g. $P(\text{outcome/event}) = 10\%$, other kernels may appear to have better accuracy if degenerate results are not filtered out, e.g. picking the majority class yields 90% accuracy, with 100% specificity and 0% sensitivity. We filter out results with sensitivity or specificity $< 50\%$.

Answers to expected questions 3

- 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 literature has numerous (13+) explicit and separable kernels including the linear, polynomial, chi-square and wavelet kernels. 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.
 - ☀ The Mercer sigmoid kernel uses 15% less SV than the Gaussian RBF kernel.
 - A: The kernel trick replaces XX^T in the SVM dual with any valid kernel. Implicit basis are often (erroneously) assumed and confounded with that.

Answers to expected questions 4

- 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 human/manual interpretation.
- Q: How does the Gaussian RBF kernel compare theoretically (or analytically) to the sigmoid and Mercer sigmoid kernels?
 - A: Future work may explore how the Mercer sigmoid may be better suited to typical/common structures found in heterogeneous clinical data. Lin and Lin noted a close relationship between the sigmoid kernel and the Gaussian RBF kernel, where the former is c.p.d.

Answers to expected questions 5

- Q: The RMS deviation $< 10.1\%$ is specific to $a=[0.1,10]$ for the normalized sigmoid kernel, where a is the maximum slope – is that range sufficient?
 - A: The literature recommends $a=1/n$ for the sigmoid kernel (although hard to find); whereas we apply that normalization and explore the range $[0.1,10]$ as a factor around it.
 - A: At smaller values of a which approach zero, the deviation grows but the kernels lose their ability to classify data in the feature space.
 - The sigmoid kernel and corresponding ($b=1/\sqrt{a}$) Mercer sigmoid kernel look like near horizontal planes. The Taylor series of the sigmoid kernel responds like a near-horizontal line in each dimension; hence the discriminating boundary between classes becomes subject to round-off error.

Answers to expected questions 6

- Q: If the sigmoid kernel is non-Mercer, and not valid for SVM, why does it seem to work well?
 - A: Ong, Mary, Canu and Smola claim and show that stable (not optimal) results with generalization bounds may be achieved with the sigmoid kernel and other indefinite kernels in kernel methods (like SVM). This is shown in the obscure context of Krein spaces as opposed to Hilbert spaces.

Experimental limitations

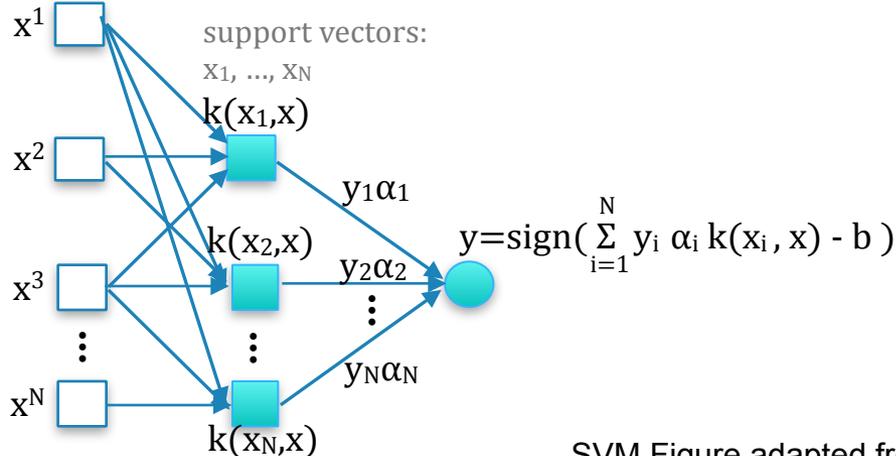
- We allowed Sequential Minimal Optimization (SMO) to violate Karush Kuhn Tucker (KKT) conditions for kernels where it was not needed
- 52% of our skin data used the clinician (vs. pathology) as ground-truth
- We used balanced costs, although melanoma detection should use an imbalanced cost
- None of 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)

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]: $k(\mathbf{x}, \mathbf{z}) = S(\mathbf{a}\mathbf{x}^T\mathbf{z} + r)$.

input vector:
 $\mathbf{x} = (x^1, \dots, x^N)$



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

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 $f(\cdot)$ on X , $X \subseteq \mathbb{R}^p$ (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 $f(x_i) = (1/\sqrt{p}) \tanh((x_i - b)/d)$ where x_i and f are real-valued (3)

- From (1,3): $k(x_i, z_i) = (1/\sqrt{p}) \tanh((x_i - b)/d) * (1/\sqrt{p}) \tanh((z_i - b)/d)$ is valid (4)

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

Mercer or p.d. kernels

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

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

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

↙
k is an inner product

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

↘
positive eigenvalues
↔

↓
convex optimization (SDP); a cone in the vector space of $p \times p$ matrices

Even if the sigmoid is 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 .
we note: the sigmoid kernel does not meet these conditions.

Separable kernels

Separable kernels are explicitly of the form $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})g(\mathbf{z})$. They have **lower space complexity**, since only the vectors \mathbf{x} and \mathbf{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
- Polynomial kernel*
- Hellinger kernel
- Hellinger exponential kernel
- Wavelet kernel
- Generalized histogram intersection kernel
- Chi-square kernel
- Probability product kernel
- Bhattacharyya kernel
- Expected Likelihood kernel
- Spectrum kernel
- Mismatch string kernel
- Weighted-degree kernel

Experimental method

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

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

log = log₁₀

*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
	Sediment	9	no	9	0	0	0

clinical

non-clinical

heterogeneous