Over-fitting in Model Selection and Its Avoidance

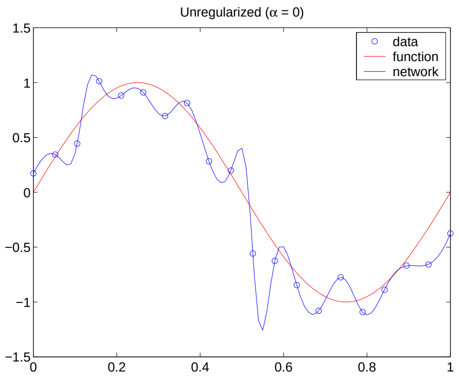
Gavin Cawley

School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ, U.K.

Saturday 27th October 2012

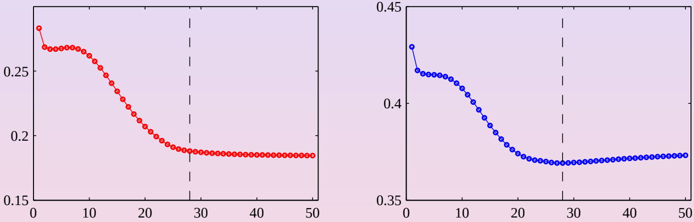
Example of Over-fitting in Training

- Use a large neural network to model a noisy sinusoid
- Small set of training samples
- Network memorizes the noise as well as the function



Classic Hallmark of Overfitting

- The training criterion monotonically decreases.
- After a while generalisation error starts to rise again.



From C. Bishop, "Pattern Recognition and Machine Learning", Springer 2006.

- We can minimise the training criterion too much!
- Exploits peculiarities of the particular sample.

Remedies for Over-fitting

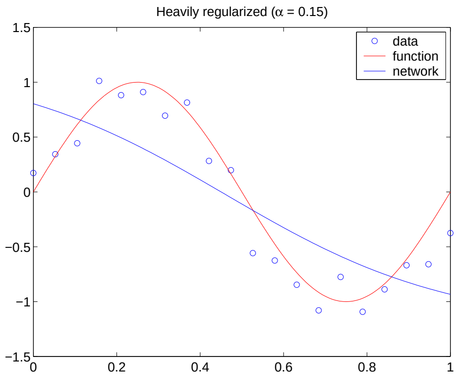
- To perform complex mappings, a neural net needs:
  - A large number of weights and hidden layer neurons
  - Weights with large magnitudes
- There are three main approaches to avoiding over-fitting
  - Early stopping - stop training before test error starts rising.
  - Structural stabilisation - prune redundant parameters from a complex model, or add parameters to a simple model
  - Regularisation - add a penalty term to penalise complex mappings

$$L_{\text{reg}}(f) = L(f) + \lambda\Omega(f)$$

- The aim is to reduce the complexity of the model to the minimum required to solve the problem given the data we have

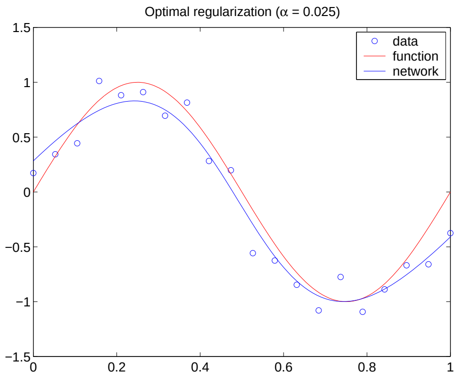
Heavily Regularised Solution

- The network has "underfitted" the training data data
- Ignored the noise, but has also ignored the underlying function
- Generalisation is poor



Optimally Regularised Solution

- Network learns underlying function, but ignores noise
- Ignores the noise, but not the function
- Generalisation is good.



Multi-level Inference

- Most machine learning algorithms involve more than one level of inference
  - First level - optimise the parameters of the model
  - Second level - optimise the hyper-parameters of the model
  - Usually stop there!
- There are many reasons
  - There may be efficient algorithms for level 1 inference
  - Overall model is not theoretically/mathematically tractable
- Second level of inference often called model selection
  - Selection of input features
  - Selection of model architecture
  - Tuning of regularisation parameters
  - Tuning of kernel parameters

Over-fitting In Model Selection

- How do we perform inference at the second level
- Minimise a model selection criterion over a finite sample
  - Often cross-validation
  - Model selection criterion is also prone to over-fitting!
- This is the topic of the talk
  - Normally assumed that model selection criterion is not susceptible to over-fitting
  - Experiments suggest otherwise
  - All rather obvious in hindsight
  - The extent of the problem is interesting
  - Can cause problems for performance evaluation
  - Considerable scope for research

Kernel Ridge Regression Machine

- Data : $\mathcal{D} = \{(x_i, t_i)\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$, $t_i \in \{-1, +1\}$
- Model : $f(x) = w \cdot \phi(x) + b$
- Regularised least-squares loss function:

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + \frac{1}{2\lambda\ell}\sum_{i=1}^{\ell}[t_i - w \cdot \phi(x_i) - b]^2$$

- $\mathcal{K}(x, x') = \phi(x) \cdot \phi(x') \quad \Longrightarrow \quad f(x_i) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(x_i, x) + b$
- System of linear equations (solve via Cholesky factorisation)

$$\begin{bmatrix} K + \lambda\ell I & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} t \\ 0 \end{bmatrix}$$

- Simple and efficient for small(ish) datasets

Kernel Functions

- Kernel models rely on a good choice of kernel function
- Linear : $\mathcal{K}(x, x') = x \cdot x'$
- Polynomial : $\mathcal{K}(x, x') = (x \cdot x' + c)^d$
- Boolean : $\mathcal{K}(x, x') = (1 + \eta)^{x \cdot x'}$
- Radial Basis Function : $\mathcal{K}(x, x') = \exp\left\{-\eta \|x - x'\|^2\right\}$
- Automatic Relevance Determination :

$$\mathcal{K}(x, \mathbf{x'}) = \exp\left\{\sum_{i=1}^{d} \eta_i [x_i - x_i']^2\right\}$$

- Must also optimise kernel parameters, $c, d, \eta, \boldsymbol{\eta}$ etc.
- Use $\boldsymbol{\theta}$ to represent the vector of hyper-parameters (including regularisation parameter, λ)

Virtual Leave-One-Out Cross-Validation

- Can perform leave-one-out cross-validation in closed form
- Let $y_i = f(x_i)$ and $C = \begin{bmatrix} K + \lambda \ell I & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}$
- It can be shown that:

$$r_i^{(-i)} = t_i - y_i^{(-i)} = \frac{\alpha_i}{C_{ii}^{-1}}$$

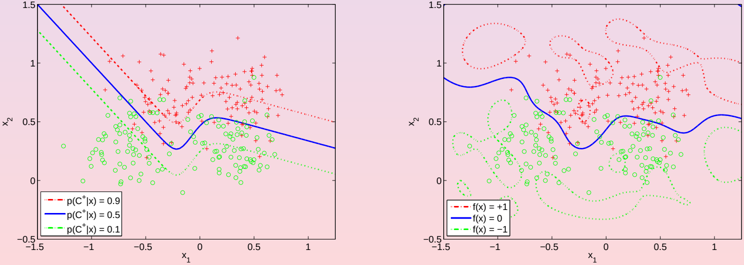
- Uses information available as a by-product of training
- Perform model selection by minimising PRESS

$$PRESS(\boldsymbol{\theta}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left[\frac{\alpha_i}{C_{ii}^{-1}} \right]^2$$

- Use e.g. Nelder-Mead simplex or scaled conjugate gradients

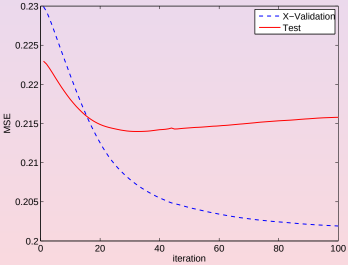
Illustration using a Synthetic Benchmark

- Based on Ripley's famous "synthetic" benchmark
- Data uniformly sampled from four bivariate Gaussians
  - Each class represented by two of the Gaussians
- Kernel ridge regression classifier with RBF kernel
  - Model parameters determined by system of linear equations
  - Two kernel and one regularisation hyper-parameters
  - Leave-one-out cross-validation can be performed for free

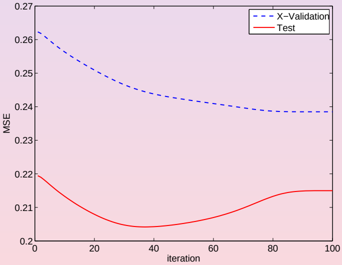


The Hallmark of Over-fitting in Model Selection

- 1000 replications, 4-fold cross-validation based model selection
- Can work out true generalisation performance analytically
- Value of model selection criterion decreases
- Generalisation performance decreases and then increases again
  - Hallmark of over-fitting - but this time at level 2



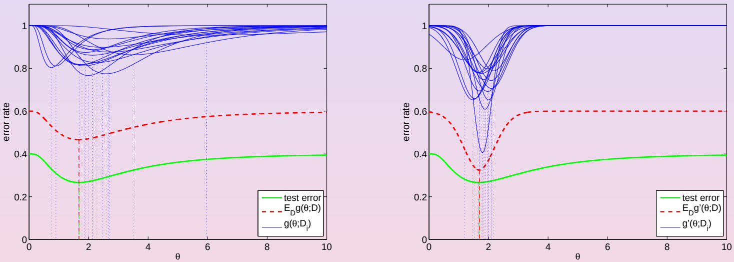
expected



worst case

What makes a Good Model Selection Criterion

- Unbiasedness often cited as beneficial
- Variance not usually mentioned

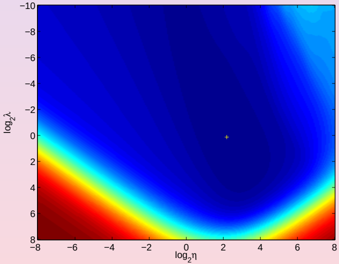


unbiased biased

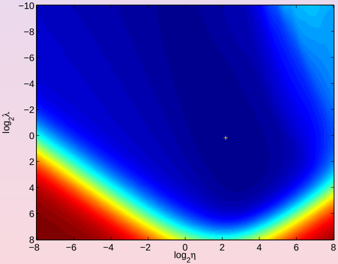
- Minimum reliably in more or less the same place as minimum of generalisation error

Model Selection for Kernel Ridge Regression

- Need to tune regularisation parameter and one kernel parameter
- Fixed training set of 256 patterns
- Disjoint validation set of 64 patterns
- 100 replications with different validation set each time

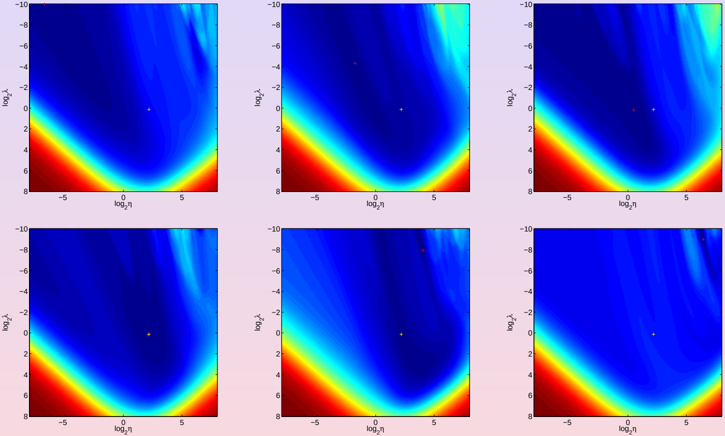


true test error

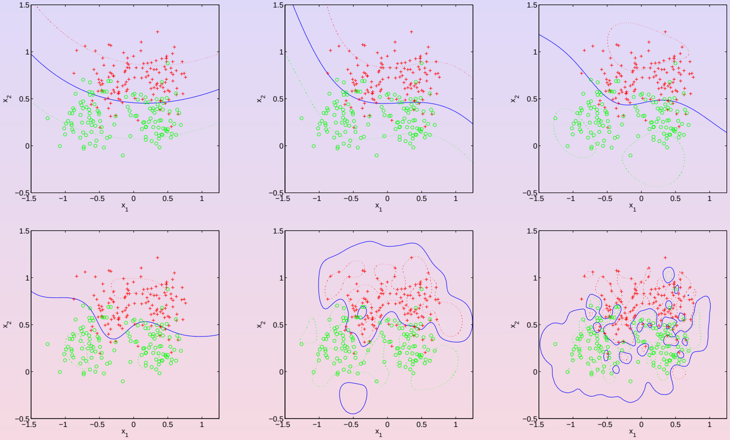


mean validation error

Variability of Validation Set Error



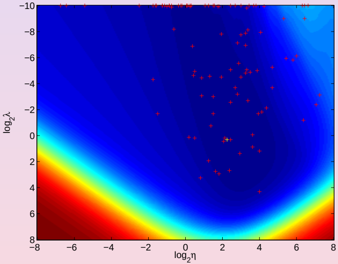
Effect of Over-fitting in Model Selection



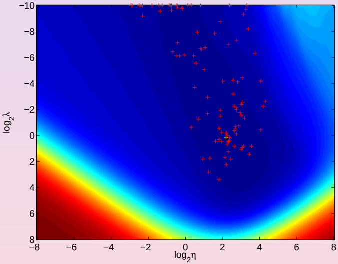
- Can result in models that over-fit or under-fit the training sample!

A Simple Fix

- Use a larger validation set
  - More samples $\Longrightarrow$ lower variance
- Increase validation set to 256 patterns



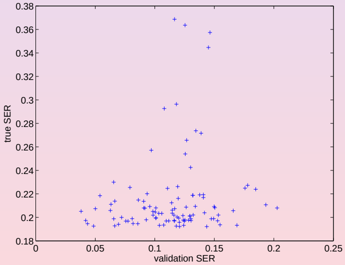
64 samples

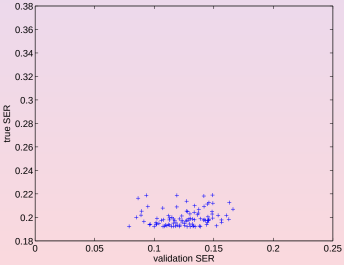


256 samples

A Simple Fix

- Larger validation set gives:
  - Lower variance estimate of generalisation
  - Lower spread of hyper-parameter values
  - Much lower spread of generalisation error
  - Lower average generalisation error
- Additional data are not always available.





Is Over-fitting in Model Selection Genuinely a Problem?

- More hyper-parameters, more degrees of freedom to over-fit the model selection criterion.
- PRESS known to have a high variance

Dataset	Test Error Rate		PRESS	
	RBF	ARD	RBF	ARD
banana	10.610 $\pm$ 0.051	10.638 $\pm$ 0.052	60.808 $\pm$ 0.636	60.957 $\pm$ 0.624
breast cancer	**26.727 $\pm$ 0.466**	28.766 $\pm$ 0.391	70.632 $\pm$ 0.328	**66.789 $\pm$ 0.385**
diabetis	**23.293 $\pm$ 0.169**	24.520 $\pm$ 0.215	146.143 $\pm$ 0.452	**141.465 $\pm$ 0.606**
flare solar	34.140 $\pm$ 0.175	34.375 $\pm$ 0.175	267.332 $\pm$ 0.480	**263.858 $\pm$ 0.550**
german	**23.540 $\pm$ 0.214**	25.847 $\pm$ 0.267	228.256 $\pm$ 0.666	**221.743 $\pm$ 0.822**
heart	**16.730 $\pm$ 0.359**	22.810 $\pm$ 0.411	42.576 $\pm$ 0.356	**37.023 $\pm$ 0.494**
image	2.990 $\pm$ 0.159	**2.188 $\pm$ 0.134**	74.056 $\pm$ 1.685	**44.488 $\pm$ 1.222**
ringnorm	**1.613 $\pm$ 0.015**	2.750 $\pm$ 0.042	28.324 $\pm$ 0.246	**27.680 $\pm$ 0.231**
splice	10.777 $\pm$ 0.144	9.943 $\pm$ 0.520	186.814 $\pm$ 2.174	**130.888 $\pm$ 6.574**
thyroid	4.747 $\pm$ 0.235	4.693 $\pm$ 0.202	9.099 $\pm$ 0.152	**6.816 $\pm$ 0.164**
titanic	22.483 $\pm$ 0.085	22.562 $\pm$ 0.109	48.332 $\pm$ 0.622	47.801 $\pm$ 0.623
twonorm	**2.846 $\pm$ 0.021**	4.292 $\pm$ 0.086	**32.539 $\pm$ 0.279**	35.620 $\pm$ 0.490
waveform	**9.792 $\pm$ 0.045**	11.836 $\pm$ 0.085	61.658 $\pm$ 0.596	**56.424 $\pm$ 0.637**

Not Confined to PRESS Either!

- Bayesian evidence not generally regarded as susceptible
- Same problem occurs for Gaussian Process classifiers.

Dataset	Test Error Rate		-Log Evidence	
	RBF	ARD	RBF	ARD
banana	10.413 ± 0.046	10.459 ± 0.049	116.894 ± 0.917	116.459 ± 0.923
breast cancer	**26.506** ± **0.487**	27.948 ± 0.492	110.628 ± 0.366	**107.181** ± **0.388**
diabetis	**23.280** ± **0.182**	23.853 ± 0.193	230.211 ± 0.553	**222.305** ± **0.581**
flare solar	34.200 ± 0.175	**33.578** ± **0.181**	394.697 ± 0.546	**384.374** ± **0.512**
german	23.363 ± 0.211	23.757 ± 0.217	359.181 ± 0.778	**346.048** ± **0.835**
heart	**16.670** ± **0.290**	19.770 ± 0.365	73.464 ± 0.493	**67.811** ± **0.571**
image	2.817 ± 0.121	**2.188** ± **0.076**	205.061 ± 1.687	**123.896** ± **1.184**
ringnorm	**4.406** ± **0.064**	8.589 ± 0.097	121.260 ± 0.499	**91.356** ± **0.583**
splice	11.609 ± 0.180	**8.618** ± **0.924**	365.208 ± 3.137	**242.464** ± **16.980**
thyroid	4.373 ± 0.219	4.227 ± 0.216	25.461 ± 0.182	**18.867** ± **0.170**
titanic	22.637 ± 0.134	22.725 ± 0.133	78.952 ± 0.670	78.373 ± 0.683
twonorm	**3.060** ± **0.034**	4.025 ± 0.068	45.901 ± 0.577	**42.044** ± **0.610**
waveform	**10.100** ± **0.047**	11.418 ± 0.091	105.925 ± 0.954	**91.239** ± **0.962**

Conclusion #1

Over-fitting in model selection can significantly reduce generalization performance!

(especially where there are many hyper-parameters)

Model Selection Bias in Performance Evaluation

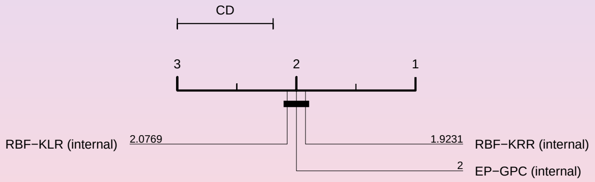
- Compare three classifiers:
  - Kernel Ridge Regression (KRR)
  - Kernel Logistic Regression (KLR)
  - Expectation Propagation (EP) based Gaussian Process Classifier (GPC)
- Suite of thirteen benchmark datasets
  - Different benchmarks present different challenges
  - 100 (20) pre-defined test/training splits
- Begin with an unbiased evaluation protocol
  - Perform model selection independently for each replicate
  - Evaluate the joint performance of the training algorithm and model selection method
  - This is the way it should always be done!

Unbiased Protocol Results

Dataset	GPC (internal)	KLR (internal)	KRR (internal)
banana	10.413 ± 0.046	10.567 ± 0.051	10.610 ± 0.051
breast cancer	26.506 ± 0.487	26.636 ± 0.467	26.727 ± 0.466
diabetis	23.280 ± 0.182	23.387 ± 0.180	23.293 ± 0.169
flare solar	34.200 ± 0.175	34.197 ± 0.170	34.140 ± 0.175
german	23.363 ± 0.211	23.493 ± 0.208	23.540 ± 0.214
heart	16.670 ± 0.290	16.810 ± 0.315	16.730 ± 0.359
image	2.817 ± 0.121	3.094 ± 0.130	2.990 ± 0.159
ringnorm	4.406 ± 0.064	1.681 ± 0.031	1.613 ± 0.015
splice	11.609 ± 0.180	11.248 ± 0.177	10.777 ± 0.144
thyroid	4.373 ± 0.219	4.293 ± 0.222	4.747 ± 0.235
titanic	22.637 ± 0.134	22.473 ± 0.103	22.483 ± 0.085
twonorm	3.060 ± 0.034	2.944 ± 0.042	2.846 ± 0.021
waveform	10.100 ± 0.047	9.918 ± 0.043	9.792 ± 0.045

Statistical (In)Significance

- None of the classifiers are statistically superior to the others
- Friedman test with Nemenyi post-hoc analysis
- Critical difference diagram:



Biased Protocol #1

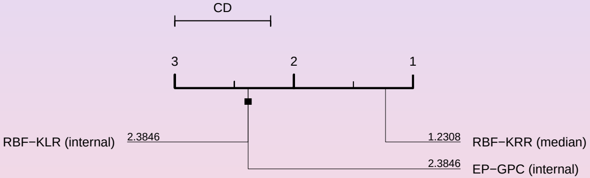
- Perform model selection separately for first five replicates
- Take median hyper-parameter values over five replicates
- Evaluate performance using those median hyper-parameter values
- Problems:
  - Median operation reduces apparent variance
  - Using constant hyper-parameters ameliorates over-fitting
  - Some test data used in fitting hyper-parameters
- Initially used by Rätsch due to computational expense
- Has been widely used in the machine learning community.
  - Over-fitting in model selection perhaps not that obvious!

Biased Protocol #1 Results

Dataset	KRR (internal)	KRR (median)	Bias
banana	10.610 $\pm$ 0.051	10.384 $\pm$ 0.042	0.226 $\pm$ 0.034
breast cancer	26.727 $\pm$ 0.466	26.377 $\pm$ 0.441	0.351 $\pm$ 0.195
diabetis	23.293 $\pm$ 0.169	23.150 $\pm$ 0.157	0.143 $\pm$ 0.074
flare solar	34.140 $\pm$ 0.175	34.013 $\pm$ 0.166	0.128 $\pm$ 0.082
german	23.540 $\pm$ 0.214	23.380 $\pm$ 0.220	0.160 $\pm$ 0.067
heart	16.730 $\pm$ 0.359	15.720 $\pm$ 0.306	1.010 $\pm$ 0.186
image	2.990 $\pm$ 0.159	2.802 $\pm$ 0.129	0.188 $\pm$ 0.095
ringnorm	1.613 $\pm$ 0.015	1.573 $\pm$ 0.010	0.040 $\pm$ 0.010
splice	10.777 $\pm$ 0.144	10.763 $\pm$ 0.137	0.014 $\pm$ 0.055
thyroid	4.747 $\pm$ 0.235	4.560 $\pm$ 0.200	0.187 $\pm$ 0.100
titanic	22.483 $\pm$ 0.085	22.407 $\pm$ 0.102	0.076 $\pm$ 0.077
twonorm	2.846 $\pm$ 0.021	2.868 $\pm$ 0.017	-0.022 $\pm$ 0.014
waveform	9.792 $\pm$ 0.045	9.821 $\pm$ 0.039	-0.029 $\pm$ 0.020

(Spurious) Statistical Significance

- KRR now appears to be significantly superior
- Difference is spurious - due to selection bias



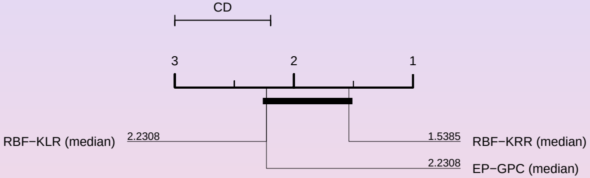
- Cannot directly compare results obtained using biased and unbiased protocols

Biased Protocol #1 - More Results

Dataset	EP-GPC (median)	RBF-KLR (median)	RBF-KRR (median)
banana	10.371 ± 0.045	10.407 ± 0.047	10.384 ± 0.042
breast cancer	26.117 ± 0.472	26.130 ± 0.474	26.377 ± 0.441
diabetis	23.333 ± 0.191	23.300 ± 0.177	23.150 ± 0.157
flare solar	34.150 ± 0.170	34.212 ± 0.176	34.013 ± 0.166
german	23.160 ± 0.216	23.203 ± 0.218	23.380 ± 0.220
heart	16.400 ± 0.273	16.120 ± 0.295	15.720 ± 0.306
image	2.851 ± 0.102	3.030 ± 0.120	2.802 ± 0.129
ringnorm	4.400 ± 0.064	1.574 ± 0.011	1.573 ± 0.010
splice	11.607 ± 0.184	11.172 ± 0.168	10.763 ± 0.137
thyroid	4.307 ± 0.217	4.040 ± 0.221	4.560 ± 0.200
titanic	22.490 ± 0.095	22.591 ± 0.135	22.407 ± 0.102
twonorm	3.241 ± 0.039	3.068 ± 0.033	2.868 ± 0.017
waveform	10.163 ± 0.045	9.888 ± 0.042	9.821 ± 0.039

Statistical Significance

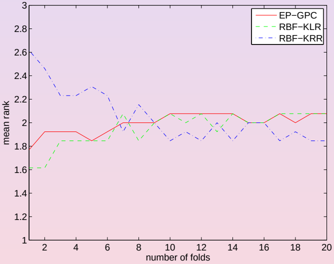
- Difference in ranks approaches statistical significance
- Again any difference is spurious



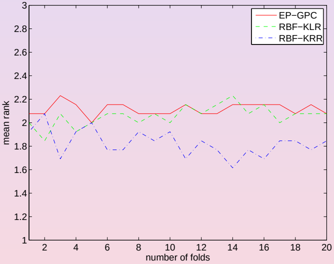
- Median protocol internally inconsistent
- Different algorithms have different susceptibilities
  - More susceptible algorithms actively favoured by the bias
- Bias greater for models with large numbers of hyper-parameters

Is This Really Due To Selection Bias?

- Repeat experiment with repeated split sample model selection
- Variance decreases as number of splits increases
- Only difference is in variance of model selection criterion



internal



median

Conclusion #2

Over-fitting in model selection can significantly bias performance evaluation!

If we don't have a clear picture of where existing algorithms fail, how can we decide how to go about improving them?

- Guidelines:
  - Use lots of data sets and/or lots of re-sampling
  - Always perform model selection independently for each test/train partition of the data
  - Evaluate combinations of training algorithm and model selection procedure
  - Automate - don't become part of the loop!

How Can We Prevent Over-Fitting In Model Selection?

- Regularize the model selection criterion!

$$M(\boldsymbol{\theta}) = \zeta Q(\boldsymbol{\theta}) + \xi \Omega(\boldsymbol{\theta}) \quad \text{where} \quad \Omega = \frac{1}{2}\sum_{i=1}^{d} \eta_i^2$$

- Penalize models with sensitive kernels.
- Marginalise (integrate out) regularization parameters ζ and ξ

$$L(\boldsymbol{\theta}) = \frac{\ell}{2}\log\{Q(\boldsymbol{\theta})\} + \frac{d}{2}\log\{\Omega(\boldsymbol{\theta})\}$$

  - Avoids third level of inference
  - Based on Bayesian ANN due to Buntine and Weigend (1991)
  - Related to evidence maximisation

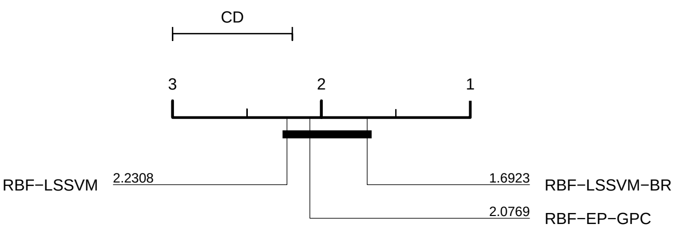
Some Results

Dataset	Radial Basis Function		
	LSSVM	LSSVM-BR	EP-GPC
Banana	10.60 ± 0.052	10.59 ± 0.050	**10.41 ± 0.046**
Breast cancer	*26.73 ± 0.466*	27.08 ± 0.494	**26.52 ± 0.489**
Diabetes	23.34 ± 0.166	**23.14 ± 0.166**	*23.28 ± 0.182*
Flare solar	34.22 ± 0.169	34.07 ± 0.171	34.20 ± 0.175
German	*23.55 ± 0.216*	23.59 ± 0.216	**23.36 ± 0.211**
Heart	*16.64 ± 0.358*	**16.19 ± 0.348**	16.65 ± 0.287
Image	3.00 ± 0.158	2.90 ± 0.154	2.80 ± 0.123
Ringnorm	**1.61 ± 0.015**	**1.61 ± 0.015**	*4.41 ± 0.064*
Splice	10.97 ± 0.158	10.91 ± 0.154	11.61 ± 0.181
Thyroid	4.68 ± 0.232	4.63 ± 0.218	*4.36 ± 0.217*
Titanic	**22.47 ± 0.085**	22.59 ± 0.120	22.64 ± 0.134
Twonorm	**2.84 ± 0.021**	**2.84 ± 0.021**	*3.06 ± 0.034*
Waveform	*9.79 ± 0.045*	**9.78 ± 0.044**	10.10 ± 0.047

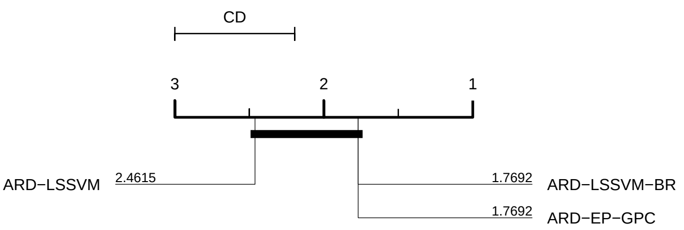
Some More Results

Dataset	Automatic Relevance Determination		
	LSSVM	LSSVM-BR	EP-GPC
Banana	10.79 $\pm$ 0.072	10.73 $\pm$ 0.070	*10.46* $\pm$ *0.049*
Breast cancer	29.08 $\pm$ 0.415	27.81 $\pm$ 0.432	27.97 $\pm$ 0.493
Diabetes	24.35 $\pm$ 0.194	23.42 $\pm$ 0.177	23.86 $\pm$ 0.193
Flare solar	34.39 $\pm$ 0.194	*33.61* $\pm$ *0.151*	**33.58** $\pm$ **0.182**
German	26.10 $\pm$ 0.261	23.88 $\pm$ 0.217	23.77 $\pm$ 0.221
Heart	23.65 $\pm$ 0.355	17.68 $\pm$ 0.623	19.68 $\pm$ 0.366
Image	**1.96** $\pm$ **0.115**	*2.00* $\pm$ *0.113*	2.16 $\pm$ 0.068
Ringnorm	2.11 $\pm$ 0.040	1.98 $\pm$ 0.026	8.58 $\pm$ 0.096
Splice	*5.86* $\pm$ *0.179*	**5.14** $\pm$ **0.145**	7.07 $\pm$ 0.765
Thyroid	4.68 $\pm$ 0.199	4.71 $\pm$ 0.214	**4.24** $\pm$ **0.218**
Titanic	*22.58* $\pm$ *0.108*	22.86 $\pm$ 0.199	22.73 $\pm$ 0.134
Twonorm	5.18 $\pm$ 0.072	4.53 $\pm$ 0.077	4.02 $\pm$ 0.068
Waveform	13.56 $\pm$ 0.141	11.48 $\pm$ 0.177	11.34 $\pm$ 0.195

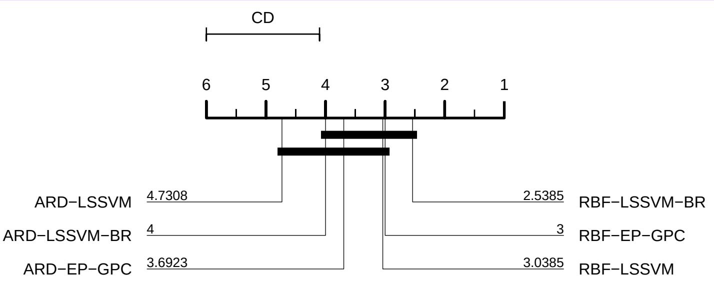
Critical Difference Diagram #1



Critical Difference Diagram #2



Critical Difference Diagram #3



Conclusion #3

Great scope for further research and practical performance gains!

Many methods for avoiding over-fitting have been investigated at the first level of inference; very few have been investigated at the second!

References

Cawley, G. C. and Talbot, N. L. C., “Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters”, Journal of Machine Learning Research, volume 8, pages 841–861, 2007.

Cawley, G. C. and Talbot, N. L. C., “On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation”, Journal of Machine Learning Research, volume 11, pages 2079–2107, 2010.