Bayesian Regularisation in Model Selection

Gavin Cawley

School of Computing Sciences
University of East Anglia
Norwich, United Kingdom
gcc@cmp.uea.ac.uk

Saturday 9th December 2006

Introduction

- Model selection is a fundamental component of best practice in applications of kernel learning methods.
- Decomposition of the error of a model selection criterion
  - **Bias** - how much the predictions depart from the true value on average.
  - **Variance** - the average squared distance of predictions from their mean.
- Leave-one-out cross-validation
  - Very efficient for many kernel machines.
  - Low bias, but relatively high variance.
- Use Bayesian regularisation in model selection
  - Prevent over-fitting of the model selection criteria.
  - No more computationally expensive than before.

Least-Squares Support Vector Machine

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

$$\mathcal{L} = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2\mu\ell}\sum_{i=1}^{\ell}\left[t_i - \mathbf{w} \cdot \phi(\mathbf{x}_i) - b\right]^2 .$$

- $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}') \quad \Longrightarrow \quad f(\mathbf{x}_i) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b.$
- System of linear equations

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

- Can be solved efficiently via Cholesky factorisation.

Kernel Functions

- Kernel models rely on a good choice of kernel function.
- Radial Basis Function

$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp\left\{-\eta\|\mathbf{x} - \mathbf{x}'\|^2\right\}.$$

- RBF with feature scaling a.k.a. Automatic Relevance Determination

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

- Must optimise kernel parameters, $\boldsymbol{\eta}$, as well as regularisation parameter
- Model selection should provide a means of choosing the kernel function as well.

Virtual Leave-One-Out Cross-Validation

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

$$r_i^{(-i)} = t_i - y_i^{(-i)} = \frac{\alpha_i}{\mathbf{C}_{ii}^{-1}}.$$

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

$$Q(\boldsymbol{\theta}) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left[\frac{\alpha_i}{\mathbf{C}_{ii}^{-1}}\right]^2 \quad \text{where} \quad \boldsymbol{\theta} = \{\mu, \eta_1, \ldots, \eta_d\}.$$

- Use conjugate gradient descent or Nelder-Mead simplex.

Regularisation in Model Selection

- Problem : The high variance of the PRESS criterion allows over-fitting given sufficient degrees of Freedom.
- Solution : Add a regularisation term to the PRESS criterion

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

- $\Omega(\boldsymbol{\theta})$ is intended to discourage hyper-parameter values giving rise to complex models.
- Only kernel parameters are currently regularised.
- Corresponds to the use of a hyper-prior in Bayesian methods
  - Has been used in Gaussian Process Classifiers (GPC).
- Problem : we now have two hyper-hyper-parameters to set :-(

Eliminating the Regularisation Parameters

- Let $Q(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$ represent the negative logarithms of the *likelihood* and *prior*,

$$p(\mathcal{D}|\boldsymbol{\theta}) = \frac{1}{Z_Q}\exp\{-\zeta Q(\boldsymbol{\theta})\} \quad \text{and} \quad p(\boldsymbol{\theta}) = \frac{1}{Z_\Omega}\exp\{-\xi\Omega(\boldsymbol{\theta})\}$$

  where $Z_Q = (2\pi/\zeta)^{\ell/2}$ and $Z_\Omega = (2\pi/\xi)^{d/2}$.

- $M(\theta)$ is then the negative logarithm of the posterior

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- We aim to integrate out ξ using a suitable hyper-prior

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|\xi)p(\xi)d\xi$$

  c.f. Buntine and Weigend (1991).

Eliminating the Regularisation Parameters

- Using the Jeffrey's prior $p(\xi) \propto 1/\xi$, and noting ξ is strictly positive

$$p(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}} \int_0^\infty \xi^{d/2-1} \exp\{-\xi \Omega(\boldsymbol{\theta}) d\xi\}$$

- Using the Gamma integral $\int_0^\infty x^{\nu-1} e^{\mu x} dx = \Gamma(\nu)/\mu^\nu$,

$$p(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}} \frac{\Gamma(d/2)}{\Omega^{d/2}} \quad \Longrightarrow \quad -\log p(\boldsymbol{\theta}) \propto \frac{d}{2} \log \Omega(\boldsymbol{\theta})$$

- Adopting the same approach to $Q(\boldsymbol{\theta})$,

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

- Regularisation parameters have been integrated out.

Relationship with The Evidence Framework

- Maximise the marginal likelihood w.r.t. ζ and ξ.
- Efficient update formulae:

$$\xi = \frac{\gamma}{2\Omega(\boldsymbol{\theta})} \quad \text{and} \quad \zeta = \frac{\ell - \gamma}{2Q(\boldsymbol{\theta})}, \quad \text{where} \quad \gamma = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \xi}$$

  $\lambda_1, \ldots, \lambda_d$ represent the eigenvalues of the Hessian of L with respect to the kernel parameters.
- From a gradient descent perspective, minimising $L \equiv$ minimising M, subject to

$$\xi^{\text{eff}} = \frac{d}{2\Omega(\boldsymbol{\theta})} \qquad \text{and} \qquad \zeta^{\text{eff}} = \frac{\ell}{2Q(\boldsymbol{\theta})}.$$

- Mildly over-regularised relative to the Evidence framework.
- No need to compute the Hessian.

What have we achieved so far?

- We have regularised the model selection criterion without introducing hyper-hyper-parameters to select.
  - Simple to optimise using e.g. scaled conjugate gradients.
  - Implementation only slightly more complcated.
  - No more computationally expensive than PRESS.
- Not as elegant as the fully Bayesian approach.
  - Cross-validation may be more robust.
  - Fewer modelling assumptions.
- Integrate-out approach likely to result in mild under-fitting
  - Model should be less sensitive to assumptions at higher levels of the hierarchy.
- Pragmatic combination of approaches
  - But does it actually work?

Empirical Evaluation

- Use multiple benchmark datasets.
  - See how classifier performs in different situations.
- Use multiple realisations of the datasets.
  - Allow estimation of statistical significance.
- Compare performance with a state-of-the-art method.
  - Expectation Propagation based Gaussian Process classifier.
  - Choose hyper-parameters to maximise the marginal likelihood.
- Use Gunnar Rätsch's suite of thirteen benchmarks.
- Must perform model selection separately in each trial.
  - More representative of actual practice.
  - Standard error reflects variance of model selection criterion.
  - Avoids selection bias (don't average hyper-parameters over the first 5 replicates!!!).

Benchmark Datasets

Dataset	Training Patterns	Testing Patterns	Number of Replications	Input Features
Banana	400	4900	100	2
Breast cancer	200	77	100	9
Diabetis	468	300	100	8
Flare solar	666	400	100	9
German	700	300	100	20
Heart	170	100	100	13
Image	1300	1010	20	18
Ringnorm	400	7000	100	20
Splice	1000	2175	20	60
Thyroid	140	75	100	5
Titanic	150	2051	100	3
Twonorm	400	7000	100	20
Waveform	400	4600	100	21

Results on Benchmark Datasets

Dataset	Radial Basis Function		
	LSSVM	**LS-SVM-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

Statistical Significance

- Compute z-score, means, $\mu_{\{a,b\}}$ and standard errors, $\sigma_{\{a,b\}}$,

$$z = \frac{\mu_a - \mu_b}{\sqrt{\sigma_a^2 + \sigma_b^2}}$$

  $z \geq 1.64$ corresponds to a 95% significance level.

- LS-SVM-BR versus LS-SVM
  - Neither model significantly better on any benchmark.
  - Too few degrees of freedom to significantly over-fit PRESS.
- LS-SVM-BR versus EP-GPC
  - Significantly better (4) : `ringnorm`, `splice`, `twonorm`, `waveform`.
  - Significantly worse (1) : `banana`.
  - Cross-validation may be more robust (fewer assumptions).

Results on Benchmark Datasets

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

Automatic Relevance Determination

- The ARD kernel often degrades predictive performance.
- LS-SVM:
  - Significantly better (2) : `image`, `splice`.
  - Significantly worse (8) : `banana`, `breast cancer`, `diabetis`, `german`, `heart`, `ringnorm`, `twonorm`, `waveform`.
- LS-SVM-BR:
  - Significantly better (3) : `flare solar`, `image`, `splice`
  - Significantly worse (4) : `heart`, `ringnorm`, `twonorm`, `waveform`.
- EP-GPC:
  - Significantly better (3) : `flare solar`, `image`, `splice`.
  - Significantly worse (6) : `breast cancer`, `diabetis`, `heart`, `ringnorm`, `twonorm`, `waveform`.
- Use ARD if identifying informative inputs is itself of interest.

Automatic Relevance Determination

- Many degrees of freedom makes it easier to over-fit the PRESS criterion.
- Bayesian regularisation is highly effective in this case.
- LS-SVM-BR versus LS-SVM:
  - Significantly better (9) : `breast cancer`, `diabetis`, `flare solar`, `german`, `heart`, `ringnorm`, `splice`, `twonorm`, `waveform`.
  - Significantly worse (0) : none.
- LS-SVM-BR versus EP-GPC:
  - Significantly better (4) : `diabetis`, `heart`, `ringnorm`, `splice`.
  - Significantly worse (2) : `banana`, `twonorm`.
- Performance of LS-SVM-BR is comparable (slightly better?) with EP-GPC.

Summary

- Virtual leave-one-out provides an efficient means for model selection for a variety of kernel learning methods.
  - High variance gives possibility of over-fitting.
  - Bayesian regularisation effective solution.
- Combination of strategies
  - Cross-validation potentially more robust.
  - Bayesian approach is good for handling nuisance parameters.
  - Model should be less sensitive to choices made at higher levels in the hierarchy.
- Pragmatic rather than principled
  - Not as elegant as the fully Bayesian approach.
  - Very easily implemented - minimal computational cost.
- Performance comparable with Gaussian Process methods.