Implementation of Baseline Models for the Model Selection Game

Gavin Cawley

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

Saturday 9th December 2006

Introduction

- Provide competitive baseline models.
  - Adapt method from IJCNN performance prediction challenge.
- Minimise test set Balanced Error Rate (BER).
  - Training sets not balanced (some highly skewed).
  - Adjust threshold or weight training patterns.
- Many datasets are high dimensional - use feature selection?
  - `HIVA` and `NOVA` have more features than training patterns.
  - Ignore this problem and hope regularisation will cope.
- `SYLVA` has too many training patterns.
  - Hope to find some trick to eliminate redundant data.
- The validation sets are very small.
  - Especially for `HIVA`, which is highly skewed.
  - Unreliable for model selection *or* performance estimation.

Bias & Variance in Model Selection

- Choose kernel and hyper-parameters to minimise estimate of generalisation performance.
- The error of an estimator can be decomposed into:
  - **Bias** - represents the degree to which the estimator is systematically different to the true value
  - **Variance** - represents the sensitivity of the estimator to the sampling of the data.
- Bias is relatively unimportant.
  - Just need the minimum in the right place.
- Variance permits over-fitting in model selection.
  - Model selection criterion gives a biased estimate of generalisation performance.
  - Problem gets worse as the number of hyper-parameters increases (e.g. feature scaling, ARD).

Bias & Variance in Performance Estimation

- Both bias and variance are important.
- Most re-sampling approaches have a low bias.
  - Leave-one-out cross-validation (Luntz 1969).
- Variance is often more of an issue:
  - Leave-one-out has a high variance (Kohavi 1995).
- Validation set is too small to be a reliable indicator.
  - e.g. `HIVA` validation set has 14 +ve and 370 -ve examples.
- Should not re-use model selection criterion.
  - Over-fitting introduces an optimistic bias.
- Model selection is an integral part of model fitting.
  - Should be performed independently in each fold of the cross-validation procedure to avoid *selection bias*.

Weighted 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:

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

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

$$\begin{bmatrix} \mathbf{K} + \mu\ell\mathbf{W} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{t} \\ 0 \end{bmatrix}, \quad \mathbf{W} = \mathrm{diag}\left(\zeta_i^{-1}, \ldots, \zeta_\ell^{-1}\right).$$

- Weighting factor $\zeta_i = \frac{\ell}{2\ell^+}$ if $t_i = +1$ or $\zeta_i = \frac{\ell}{2\ell^-}$ otherwise.

Kernel Functions

- Kernel models rely on a good choice of kernel function.
- Linear : $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$.
- Polynomial : $\mathcal{K}(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + c)^d$.
- Boolean : $\mathcal{K}(\mathbf{x}, \mathbf{x}') = (1 + \eta)^{\mathbf{x} \cdot \mathbf{x}'}$.
- Radial Basis Function : $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp\left\{-\eta \|\mathbf{x} - \mathbf{x}'\|^2\right\}$.
- ARD : $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp\left\{-\sum_{i=1}^{d} \eta_i \left(\mathbf{x}_i - \mathbf{x}'_i\right)^2\right\}$.
- Must also optimise kernel parameters, c, d, η etc.
- Also try normalised kernels:

$$\widehat{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = \frac{\mathcal{K}(\mathbf{x}, \mathbf{x}')}{\sqrt{\mathcal{K}(\mathbf{x}, \mathbf{x})\mathcal{K}(\mathbf{x}', \mathbf{x}')}}$$

- N.B. Normalised Boolean kernel $\equiv$ RBF kernel.

Virtual Leave-One-Out Cross-Validation

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

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

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

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

Model Selection Criteria

- Predicted residual sum of squares (PRESS)
- Smoothed error rate

$$ERR(\boldsymbol{\theta}) = \frac{1}{\ell}\sum_{i=1}^{\ell} \zeta_i \Psi\left\{t_i r_i^{(-i)} - 1\right\}, \quad \text{where} \quad \Psi\{x\} = \frac{1}{1 + e^{-\gamma x}}$$

- Hinge ($p = 1$) and squared Hinge ($p = 2$) loss

$$HINGE(\boldsymbol{\theta}) = \frac{1}{\ell}\sum_{i=1}^{\ell} \zeta_i \left[t_i r_i^{(-i)}\right]_+^p, \quad \text{where} \quad [x]_+ = \max\{0, x\}.$$

- Smoothed Wilcoxon-Mann-Whitney statistic (AOROC)

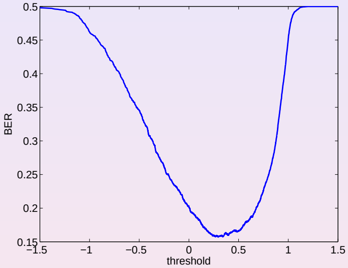
$$WMW(\boldsymbol{\theta}) = \frac{1}{\ell^+\ell^-}\sum_{i:t_i=+1}\sum_{j:t_j=-1} \Psi\left\{y_i^{(-i)} - y_j^{(-j)}\right\}$$

Gridsearch Considered Harmful?

- Easy to compute partial derivatives of e.g. PRESS criterion.
  - Reparameterise for unconstrained optimisation, $\tilde{\boldsymbol{\theta}} = \log_2 \boldsymbol{\theta}$.
  - Scaled conjugate gradient descent works well.
  - Matlab optimization toolbox routine `fminunc`.
- Computational expense:
  - $\mathcal{O}(\ell^3)$ per kernel-parameter.
  - $\mathcal{O}(\ell^2)$ for the regularisation parameter.
- Approximate gradient information using finite differences.
  - Implemented as an option by `fminunc`.
  - Typically only about twice as expensive.
- Alternatively use Nelder-Mead simplex algorithm.
  - Matlab optimization toolbox routine `fminsearch`.
  - Typically around twice as slow as CG using finite differences.
- No real need to use grid-search.

Optimising the Threshold

- Weight training patterns or modify threshold to optimise BER.
- Alternatively could perform correction analytically.
- LS-SVM similar to KFD.
- Relies on assumptions regarding the distribution of patterns.



- Setting threshold to minimise training set BER ineffective.
  - Especially if zero error is achieved on the training set.
- Set threshold to minimise leave-one-out BER.
  - Prevents use of leave-one-out BER for performance estimation.

Basic Strategy

- Perform model selection using virtual leave-one-out cross-validation.
  - Un-weighted training and model selection ctiteria.
  - Different kernel functions and selection criterion.
  - Scaled conjugate gradient descent optimisation.
- Train final model.
- Set threshold so as to minimise the leave-one-out BER.
- Choose best combination of factors by minimising leave-one-out cross-validation BER.
- Estimate performance for the best model using 100 random training/test splits of the data.
  - Perform model selection independently in each fold.
  - Set threshold independently in each trial.

IJCNN Challenge Results : Final Submission

Table: Final submission (joint winner of challenge), model choice via leave-one-out BER, performance estimation via 100-fold validation BER.

Dataset	Balanced Error			Guess	Guess Error	Test Score
	Train	Valid	Test			
ADA	0.1490	0.1542	0.1845	0.1742	0.0103	0.1947
GINA	0.0000	0.0000	0.0461	0.0470	0.0009	0.0466
HIVA	0.0180	0.0216	0.2804	0.2776	0.0028	0.2814
NOVA	0.0000	0.0000	0.0445	0.0470	0.0025	0.0464
SYLVA	0.0028	0.0029	0.0067	0.0065	0.0002	0.0067
Overall	0.0340	0.0357	0.1124	0.1105	0.0034	0.1152

(`ADA` - PRESS, Boolean kernel : `GINA` - PRESS, normalised cubic kernel, weighted training criterion : `HIVA` - PRESS, RBF kernel : `NOVA` - WMW, normalised quadratic kernel : `SYLVA` - WMW, normalised quadratic kernel, weighted training criterion.)

IJCNN Challenge Results : Regression Analysis

Table: Weights obtained by regression analysis of 100-fold validation estimate of the test balanced error rate, note lack of consistent pattern.

Factor	ADA	GINA	HIVA	NOVA	SYLVA
PRESS	-0.4729	+0.0049	-0.1077	-0.2036	-0.0615
HINGE[1]	+0.6871	+0.0375	-0.3774	+1.4446	+0.1203
HINGE[2]	-0.2796	-0.0005	-0.4189	+0.0037	-0.0554
WMW	-0.6645	+0.0082	-0.1184	-0.7283	-0.0830
ERATE	-0.2087	-0.0265	-0.3169	-0.2913	+0.0165
Training	+0.8832	-0.0085	+0.4943	-0.0806	+0.0420
Selection	-0.7922	-0.0132	+0.8001	-0.3271	-0.0236
Linear	+0.5679	+1.9856	+0.1422	-0.7780	-0.5275
Quadratic	-0.6471	-0.4272	+0.0343	-0.5183	-0.9257
Cubic	-0.7513	-0.5911	-0.8821	+0.1824	-0.9274
Boolean	+0.0629	-0.4682	-0.6189	+0.7011	+1.1598
RBF	-0.1711	-0.4754	-0.0149	+0.6379	+1.1577

Results : The Good

- Select models according to 100-fold validation BER.
- Low variance estimator.
- Computationally expensive.
- No results for ARD kernel.

Benchmark	Criterion	Kernel	XVAL	TRAIN	VALID	TEST
ADA	XENT	Quadratic	0.1879	0.1682	0.2127	?.????
GINA	XENT	CUBIC	0.0527	0.0000	0.0285	?.????
HIVA	PRESS	Quadratic	0.2444	0.0212	0.2535	?.????
NOVA	XENT	Linear	0.0483	0.0004	0.0440	?.????
SYLVA	SERATE	Quadratic	0.0076	0.0020	0.0053	?.????
Overall			0.1082	0.0384	0.1088	?.????

Results : The Bad

- Select models according to validation set BER.
- High variance estimator.
  - Especially for HIVA.
  - Probably won't work all that well.
- Computationally inexpensive.

Benchmark	Criterion	Kernel	TRAIN	VALID	TEST
ADA	PRESS	ARD	0.1464	0.1806	?.????
GINA	PRESS	ARD	0.0000	0.0253	?.????
HIVA	PRESS	Cubic	0.0158	0.2467	?.????
NOVA	PRESS	Linear	0.0004	0.0440	?.????
SYLVA	PRESS	Quadratic	0.0023	0.0045	?.????
Overall			0.0330	0.10002	?.????

Results : The Ugly

- Select models according to leave-one-out BER.
- Moderately high-variance estimator.
- Biased as also directly used to set the threshold.
- Computationally inexpensive.
- Likely to be better than "The Bad".

Benchmark	Criterion	Kernel	LOO	TRAIN	VALID	TEST
ADA	PRESS	ARD	0.1732	0.1464	0.1806	?.????
GINA	PRESS	ARD	0.0230	0.0000	0.0253	?.????
HIVA	PRESS	Quadratic	0.2358	0.0212	0.2535	?.????
NOVA	PRESS	Linear	0.0424	0.0004	0.0440	?.????
SYLVA	PRESS	RBF	0.0060	0.0020	0.0049	?.????
Overall			0.0961	0.0340	0.1016	?.????

Summary

- Careful model selection is part of best practice.
- Performance estimation is also important
- Model tuning/selection:
  - Computationally expensive - need something cheap!
  - Virtual leave-one-out cross-validation.
  - Choice of criteria relatively unimportant.
- Performance estimation:
  - Only performed once - cost less important.
  - (Repeated) k-fold cross-validation.
  - Bootstrap.
  - Multiple random test/train splits.
  - Low bias and low variance are both desirable.
  - Perform model selection independently in each fold.
  - Do not re-use the model selection criteria for performance estimation.
- Well worth burning lots of processor cycles to get it right!