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**: \( \mathcal{D} = \{(x_i, t_i)\} \), \( x_i \in \mathcal{X} \subset \mathbb{R}^d \), \( t_i \in \{-1, +1\} \).
- **Model**: \( f(x) = w \cdot \phi(x) + b \),
- **Regularised least-squares loss function**: 
  \[
  \mathcal{L} = \frac{1}{2} \|w\|^2 + \frac{1}{2\mu\ell} \sum_{i=1}^{\ell} \zeta_i [t_i - w \cdot \phi(x_i) - b]^2.
  \]
- **Kernel**: \( \mathcal{K}(x, x') = \phi(x) \cdot \phi(x') \implies f(x_i) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(x_i, x) + b. \)
- **System of linear equations (solve via Cholesky factorisation)**
  \[
  \begin{bmatrix}
  K + \mu\ell W & 1 \\
  1^T & 0
  \end{bmatrix}
  \begin{bmatrix}
  \alpha \\
  b
  \end{bmatrix}
  =
  \begin{bmatrix}
  t \\
  0
  \end{bmatrix},
  \quad
  W = \text{diag} \left( \zeta_i^{-1}, \ldots, \zeta_\ell^{-1} \right).
  \]
- **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**: \( \mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp \left\{ -\sum_{i=1}^{d} \eta_i (\mathbf{x}_i - \mathbf{x}'_i)^2 \right\} \).
- Must also optimise kernel parameters, \( c, d, \eta \) etc.
- Also try normalised kernels:

\[
\hat{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = \frac{\mathcal{K}(\mathbf{x}, \mathbf{x}')}{\sqrt{\mathcal{K}(\mathbf{x}, \mathbf{x})\mathcal{K}(\mathbf{x}', \mathbf{x}')}}
\]

- 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 $y_i = f(x_i)$ and $C = \begin{bmatrix} K + \mu \ell W & 1 \\ 1^T & 0 \end{bmatrix}$.
- It can be shown that:
  \[ r_i^{(-1)} = t_i - y_i^{(-i)} = \frac{\alpha_i}{C_{ii}^{-1}}. \]
- Uses information available as a by-product of training.
- Perform model selection by minimising (weighted) PRESS
  \[ PRESS(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} \zeta_i \left( \frac{\alpha_i}{C_{ii}^{-1}} \right)^2. \]
Model Selection Criteria

- Predicted residual sum of squares (PRESS)

\[ ERR(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i \psi \left\{ t_i r_i^{(-i)} - 1 \right\}, \text{ where } \psi\{x\} = \frac{1}{1 + e^{-\gamma x}} \]

- Smoothed error rate

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

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

- Smoothed Wilcoxon-Mann-Whitney statistic (AOROC)

\[ WMW(\theta) = \frac{1}{\ell^+ \ell^-} \sum_{i:t_i=+1} \sum_{j:t_j=-1} \psi \left\{ y_i^{(-i)} - y_j^{(-j)} \right\} \]
Gridsearch Considered Harmful?

- Easy to compute partial derivatives of e.g. PRESS criterion.
  - Reparameterise for unconstrained optimisation, $\tilde{\theta} = \log_2 \theta$.
  - Scaled conjugate gradient descent works well.
  - Matlab optimization toolbox routine `fminunc`.

- Computational expense:
  - $O(\ell^3)$ per kernel-parameter.
  - $O(\ell^2)$ 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 criteria.
  - 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.

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

### 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.

<table>
<thead>
<tr>
<th>Factor</th>
<th>ADA</th>
<th>GINA</th>
<th>HIVA</th>
<th>NOVA</th>
<th>SYLVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRESS</td>
<td>-0.4729</td>
<td>+0.0049</td>
<td>-0.1077</td>
<td>-0.2036</td>
<td>-0.0615</td>
</tr>
<tr>
<td>HINGE¹</td>
<td>+0.6871</td>
<td>+0.0375</td>
<td>-0.3774</td>
<td>+1.4446</td>
<td>+0.1203</td>
</tr>
<tr>
<td>HINGE²</td>
<td>-0.2796</td>
<td>-0.0005</td>
<td>-0.4189</td>
<td>+0.0037</td>
<td>-0.0554</td>
</tr>
<tr>
<td>WMW</td>
<td>-0.6645</td>
<td>+0.0082</td>
<td>-0.1184</td>
<td>-0.7283</td>
<td>-0.0830</td>
</tr>
<tr>
<td>ERATE</td>
<td>-0.2087</td>
<td>-0.0265</td>
<td>-0.3169</td>
<td>-0.2913</td>
<td>+0.0165</td>
</tr>
<tr>
<td>Training</td>
<td>+0.8832</td>
<td>-0.0085</td>
<td>+0.4943</td>
<td>-0.0806</td>
<td>+0.0420</td>
</tr>
<tr>
<td>Selection</td>
<td>-0.7922</td>
<td>-0.0132</td>
<td>+0.8001</td>
<td>-0.3271</td>
<td>-0.0236</td>
</tr>
<tr>
<td>Linear</td>
<td>+0.5679</td>
<td>+1.9856</td>
<td>+0.1422</td>
<td>-0.7780</td>
<td>-0.5275</td>
</tr>
<tr>
<td>Quadratic</td>
<td>-0.6471</td>
<td>-0.4272</td>
<td>+0.0343</td>
<td>-0.5183</td>
<td>-0.9257</td>
</tr>
<tr>
<td>Cubic</td>
<td>-0.7513</td>
<td>-0.5911</td>
<td>-0.8821</td>
<td>+0.1824</td>
<td>-0.9274</td>
</tr>
<tr>
<td>Boolean</td>
<td>+0.0629</td>
<td>-0.4682</td>
<td>-0.6189</td>
<td>+0.7011</td>
<td>+1.1598</td>
</tr>
<tr>
<td>RBF</td>
<td>-0.1711</td>
<td>-0.4754</td>
<td>-0.0149</td>
<td>+0.6379</td>
<td>+1.1577</td>
</tr>
</tbody>
</table>
Results: The Good

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

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Criterion</th>
<th>Kernel</th>
<th>XVAL</th>
<th>TRAIN</th>
<th>VALID</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA</td>
<td>XENT</td>
<td>Quadratic</td>
<td>0.1879</td>
<td>0.1682</td>
<td>0.2127</td>
<td>???</td>
</tr>
<tr>
<td>GINA</td>
<td>XENT</td>
<td>CUBIC</td>
<td>0.0527</td>
<td>0.0000</td>
<td>0.0285</td>
<td>???</td>
</tr>
<tr>
<td>HIVA</td>
<td>PRESS</td>
<td>Quadratic</td>
<td>0.2444</td>
<td>0.0212</td>
<td>0.2535</td>
<td>???</td>
</tr>
<tr>
<td>NOVA</td>
<td>XENT</td>
<td>Linear</td>
<td>0.0483</td>
<td>0.0004</td>
<td>0.0440</td>
<td>???</td>
</tr>
<tr>
<td>SYLVA</td>
<td>SERATE</td>
<td>Quadratic</td>
<td>0.0076</td>
<td>0.0020</td>
<td>0.0053</td>
<td>???</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>0.1082</td>
<td>0.0384</td>
<td>0.1088</td>
<td>???</td>
</tr>
</tbody>
</table>
**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.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Criterion</th>
<th>Kernel</th>
<th>TRAIN</th>
<th>VALID</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA</td>
<td>PRESS</td>
<td>ARD</td>
<td>0.1464</td>
<td>0.1806</td>
<td>?.????</td>
</tr>
<tr>
<td>GINA</td>
<td>PRESS</td>
<td>ARD</td>
<td>0.0000</td>
<td>0.0253</td>
<td>?.????</td>
</tr>
<tr>
<td>HIVA</td>
<td>PRESS</td>
<td>Cubic</td>
<td>0.0158</td>
<td>0.2467</td>
<td>?.????</td>
</tr>
<tr>
<td>NOVA</td>
<td>PRESS</td>
<td>Linear</td>
<td>0.0004</td>
<td>0.0440</td>
<td>?.????</td>
</tr>
<tr>
<td>SYLVA</td>
<td>PRESS</td>
<td>Quadratic</td>
<td>0.0023</td>
<td>0.0045</td>
<td>?.????</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>0.0330</td>
<td>0.10002</td>
<td>?.????</td>
</tr>
</tbody>
</table>
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”.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Criterion</th>
<th>Kernel</th>
<th>LOO</th>
<th>TRAIN</th>
<th>VALID</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA</td>
<td>PRESS</td>
<td>ARD</td>
<td>0.1732</td>
<td>0.1464</td>
<td>0.1806</td>
<td>????</td>
</tr>
<tr>
<td>GINA</td>
<td>PRESS</td>
<td>ARD</td>
<td>0.0230</td>
<td>0.0000</td>
<td>0.0253</td>
<td>????</td>
</tr>
<tr>
<td>HIVA</td>
<td>PRESS</td>
<td>Quadratic</td>
<td>0.2358</td>
<td>0.0212</td>
<td>0.2535</td>
<td>????</td>
</tr>
<tr>
<td>NOVA</td>
<td>PRESS</td>
<td>Linear</td>
<td>0.0424</td>
<td>0.0004</td>
<td>0.0440</td>
<td>????</td>
</tr>
<tr>
<td>SYLVA</td>
<td>PRESS</td>
<td>RBF</td>
<td>0.0060</td>
<td>0.0020</td>
<td>0.0049</td>
<td>????</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>0.0961</td>
<td>0.0340</td>
<td>0.1016</td>
<td>????</td>
</tr>
</tbody>
</table>
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!