# *Baseline Models using Kernel Methods*

**Gavin Cawley** & Nicola L. C. Talbot

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

Friday 17$^{\mathrm{th}}$ August 2007

# *Introduction*

- Aim: To produce competitive agnostic track baseline models.
- Method: Least-squares support vector machine.
  - Simple to implement.
  - Reasonably efficient for small datasets.
  - Model selection via leave-one-out cross-validation.
  - Performed well on the previous challenge.
- Issues:
  - Minimise Balanced Error Rate (BER) on the test set.
  - Many datasets are high dimensional.
  - `SYLVA` has too many training patterns.
  - The validation sets are very small.
  - Limited computing power available.
- Had a go at the prior knowledge track as well.

# *Bias & Variance in Model Selection*

- Choose hyper-parameters to minimise estimate of generalisation error.
- 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*.

## *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 (solve via Cholesky factorisation)

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

- Simple and efficient for small(ish) datasets.

## *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 : $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp\left\{-\eta \|\mathbf{x} - \mathbf{x}'\|^2\right\}$.
- Must also optimise kernel parameters, $c, d, \eta$ 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 $\mathbf{C} = \begin{bmatrix} \mathbf{K} + \mu\ell\mathbf{I} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}$.
- 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

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

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

## *Basic Strategy*

- Perform model selection using virtual leave-one-out cross-validation.
  - Weighted training and/or weighted model selection criteria.
  - Different kernel functions and selection criterion.
  - Nelder-Mead simplex optimisation.
- Train final models on training + validation sets (agnostic).
- Set threshold for estimating BER:
  - Set threshold to minimise the leave-one-out BER.
- Choose best combination of factors by minimising LOO BER.
- Performance estimation:
  - 100 random training/test splits (agnostic).
  - 10-fold cross-validation (prior knowledge).
  - Perform model selection independently in each fold.

# *Results: ADA*

- Prior knowledge track encoding quite good already.
- Box-Tidwell transformation of `age`, `capital-gain` & `capital loss`, e.g.

$$x_i^{\mathrm{age}} = \sqrt[10]{x_i^{\mathrm{age}}}$$

| model | kernel | cross-validation | | validation set | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | linear | 0.2004 | 0.8838 | 0.2206 | 0.8644 |
| KRR | poly ($p = 2$) | 0.1909 | 0.8948 | 0.2143 | 0.8745 |
| KRR | poly ($p = 3$) | 0.1920 | 0.8941 | 0.2094 | 0.8727 |
| KRR | RBF | 0.1949 | 0.8941 | 0.2095 | 0.8729 |
| KRR | ARD | **0.1653**† | **0.9180**† | **0.1740** | **0.8910** |

† biased leave-one-out estimate from the model selection process.

## *Results : GINA - Agnostic Track*

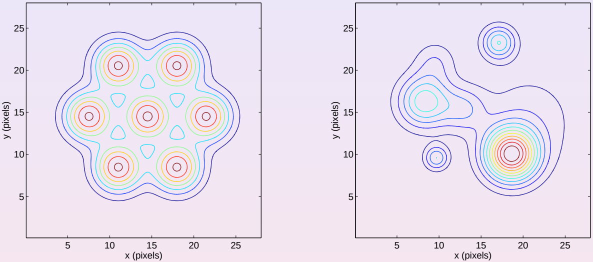
- Optical character recognition.
- Many distractors:
  - Features represent bit-map for two adjacent digits.
  - Target is one if second digit is even.
- Normalise input features.

| model | kernel | 100-fold validation | | validation set | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | linear | 0.1324 | 0.9364 | 0.1273 | 0.9461 |
| KRR | poly ($p = 2$) | 0.0578 | 0.9848 | 0.0317 | 0.9940 |
| KRR | poly ($p = 3$) | 0.0532 | 0.9870 | 0.0285 | 0.9955 |
| KRR | RBF | 0.0571 | 0.9853 | 0.0442 | 0.9955 |
| KRR | PCA-ARD | **0.0297**† | **0.9950**† | **0.0253** | **0.9968** |

† biased leave-one-out estimate from the model selection process.

# Results: *GINA* - Prior Knowledge Track

- Use RBF kernel with tunable Gaussian receptive fields.




- Target is a composite concept {1,3,5,7,9} vs {0,2,4,6,8}
  - Train 25 models to distinguish between odd-even pairs.
  - Train model to combine the output of the experts.
  - Train combiner with LOO output of the experts.

## *Results: GINA - Prior Knowledge Track*

- Getting rid of the distractors seems to help.
- MRF and hierarchical models make less difference.

| model | kernel | cross validation | | validation set | |
|---|---|---|---|---|---|
| | | BER | AUC | BER | AUC |
| KRR | linear | 0.1297 | 0.9416 | 0.1270 | 0.9525 |
| KRR | poly ($p = 2$) | 0.0365 | 0.9914 | 0.0158 | 0.9998 |
| KRR | poly ($p = 3$) | 0.0310 | 0.9938 | 0.0095 | **0.9999** |
| KRR | poly ($p = 4$) | 0.0284 | 0.9948 | **0.0064** | **0.9999** |
| KRR | poly ($p = 5$) | 0.0279 | 0.9949 | **0.0064** | **0.9999** |
| KRR | poly ($p = 6$) | 0.0256 | 0.9949 | 0.0126 | **0.9999** |
| KRR | RBF | 0.0290 | 0.9945 | 0.0095 | 0.9998 |
| KRR | MRF | 0.0315 | 0.9948 | 0.0157 | 0.9996 |
| KRR+KRR | RBF+RBF | 0.0263 | 0.9956 | 0.0128 | 0.9996 |
| KRR+KRR | RBF+ARD | 0.0253 | 0.9959 | 0.0192 | 0.9994 |

# *Results: HIVA*

- Agnostic track

| model | kernel | 100-fold validation | | validation set | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | linear | 0.2547 | 0.8071 | 0.3311 | 0.6990 |
| KRR | poly ($d = 2$) | **0.2444** | 0.7991 | 0.2535 | 0.7253 |
| KRR | poly ($d = 3$) | 0.2523 | 0.8051 | **0.2467** | 0.7486 |
| KRR | RBF | 0.2495 | **0.8092** | 0.2819 | **0.7604** |

- Prior knowledge track - ChemTK chemical fingerprint

| model | kernel | 100-fold validation | | validation set | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | linear | 0.2957 | **0.7988** | **0.2548** | 0.7486 |
| KRR | poly ($d = 2$) | 0.2914 | 0.7411 | 0.2476 | 0.6786 |
| KRR | poly ($d = 3$) | **0.2888** | 0.7406 | 0.2629 | **0.7741** |
| KRR | poly ($d = 4$) | 0.2989 | 0.7365 | 0.3444 | 0.7384 |
| KRR | RBF | 0.4889 | 0.4573 | 0.5000 | 0.4519 |

# *Results: NOVA - Agnostic Track*

- Text classification problem
  - Distinguish between usenet groups by content.
  - Short words deleted.
  - 2000 very common words deleted.
  - Words truncated to first seven letters.
  - 16,969 features - far more features than patterns.

| **model** | **kernel** | **100-fold validation** | | **validation set** | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | linear | **0.0491** | **0.9878** | **0.0440** | **0.9968** |
| KRR | poly ($d = 2$) | 0.0550 | 0.9862 | 0.0640 | 0.9955 |
| KRR | poly ($d = 3$) | 0.0569 | 0.9854 | **0.0044** | 0.9947 |
| KRR | RBF | 0.0635 | 0.9828 | 0.0480 | 0.9942 |

## *Results: NOVA - Prior Knowledge Track*

- Stemming - remove suffixes and affixes to leave root.
  - E.g. "fisher", "fishing" & "fished" become "fish"
- Spell checking - USENET messages often posted in haste.
- Term frequency-inverse document frequency (TF-IDF) coding scheme

$$tf = \frac{n_i}{\sum_k n_k}, \qquad \& \qquad idf = \log\left\{\frac{|D|}{|d_k \supset t_i|}\right\}$$

| **model** | **pre-processing** | **cross validation** | | **validation set** | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | none | **0.0432** | **0.9894** | 0.0540 | **0.9886** |
| KRR | stemming | 0.0504 | 0.9890 | **0.0360** | 0.9878 |
| KRR | spell+stem | 0.0626 | 0.9817 | 0.0540 | 0.9782 |

## *Results:* ***SYLVA*** *- Agnostic Track*

- Based on Forest Cover benchmark.
  - Distinguish Ponderosa Pine from all other species.
- Many distractors!
- Two features can be used to pre-classify the data.
  - Remaining "awkward" patterns classified via KRR.

| model | kernel | 100-fold validation | | validation set | |
|---|---|---|---|---|---|
| | | **BER** | **AUC** | **BER** | **AUC** |
| KRR | linear | 0.0149 | 0.9982 | 0.0069 | 0.9980 |
| KRR | poly ($d = 2$) | **0.0077** | **0.9991** | **0.0045** | 0.0990 |
| KRR | poly ($d = 3$) | 0.0078 | 0.9990 | **0.0045** | **0.9991** |
| KRR | RBF | 0.0079 | 0.9990 | 0.0049 | **0.9991** |

## *Results: SYLVA - Prior Knowledge Track*

- Separate the two sub-patterns (26,172 records).
- No ponderosa pine in Rahwa or Neotah.
- Only found in 13 of the 40 soil types.
- This leaves only 1,335 *difficult* patterns.
- validation set BER of 0.0041 & an AUC of 0.9992.

| Cover Type | Rawah | Neota | Comanche Peak | Cache la Poudre |
|---|---|---|---|---|
| Spruce-Fir | 4779 | 796 | 3919 | 0 |
| Lodgepole Pine | 6635 | 410 | 5609 | 135 |
| Ponderosa Pine | 0 | 0 | 663 | 947 |
| Cottonwood/Willow | 0 | 0 | 0 | 137 |
| Aspen | 174 | 0 | 245 | 0 |
| Douglas-Fir | 0 | 0 | 373 | 453 |
| Krummholz | 228 | 104 | 565 | 0 |
| **Total** | 11816 | 1310 | 11374 | 1672 |

## *Summary*

- Don't re-use the model selection criteria for performance estimation.
- Model tuning/selection:
  - Computationally expensive - need something cheap!
  - Virtual leave-one-out cross-validation.
- Performance estimation:
  - Only performed once - cost less important.
  - $k$-fold cross-validation.
  - Low bias and low variance are both desirable.
  - Use as many iterations as are feasible.
  - Perform model selection independently in each fold.
- Prior knowledge track solutions only slightly better.
  - Is that a good thing?