Selection of Basis Functions in Regression as Search Guided by the Evidence

Ignacio Barrio Enrique Romero Lluís Belanche

Soft Computing Group
Universitat Politècnica de Catalunya

Learning'06

Outline

Outline

Outline

Outline

Outline

Problem Definition

Data Set: $\{x_n, t_n\}_{n=1}^{N} \quad t_n \in \mathcal{R}$
Use a generalized linear model:

$$y(x; w) = \sum_{i=1}^{M} w_i \phi_i(x)$$

- Compute the parameters w_i
- Select M basis functions from a dictionary (M is unknown)

Outline

Review of Methods for the Selection of Basis

Implicit Selection

- Basis Pursuit
- Relevance Vector Machine
- Least Absolute Shrinkage and Selection Operator
- Support Vector Machines

Explicit Selection

- Matching Pursuits
- Orthogonal Least Squares
- Regularized OLS
- Kernel Matching Pursuit
- Gaussian Process Approximations

Outline

Objectives of this work

Within the Bayesian framework, the evidence for the model has been suggested to compare different models.

Objective: Find a model with high evidence.

How: Use an explicit search process guided by the evidence.

Use different search strategies to assess the effect of evidence maximization.

Outline

Bayesian Interpolation

D. J. Mackay. Bayesian Interpolation. *Neural Computation*, 1992.

Three levels of inference:

- first level: Posterior distribution over the parameters
- second level: Adaptation of hyperparameters
- third level: Comparison of different models

First Level of Inference

Assume i.i.d. additive Gaussian noise with variance σ^2

$$t = y(x; w) + \nu = \Phi w + \nu$$

$$P(t|w, \beta) = (2\pi\beta^{-1})^{-N/2} \exp\left\{-\frac{\beta}{2}\|t - \Phi w\|^2\right\}$$

where $\beta = 1/\sigma^2$.

Assume zero-mean Gaussian prior over w:

$$P(w|\alpha) = (2\pi\alpha^{-1})^{-M/2} \exp\left(-\frac{\alpha\|w\|^2}{2}\right)$$

where α is the inverse variance.

First Level of Inference

Find the posterior over the parameters:

$$P(w|t,\alpha,\beta) = \frac{P(t|w,\beta)P(w|\alpha)}{P(t|\alpha,\beta)}$$

which is Gaussian

$$P(w|t,\alpha,\beta) = (2\pi)^{-M/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\right\},$$

where

$$\Sigma = (\beta\Phi^T\Phi + \alpha I)^{-1} \qquad \text{and} \qquad \mu = \beta\Sigma\Phi^T t.$$

Second Level of Inference

Marginal likelihood: $P(t|\alpha,\beta) = \int P(t|w,\beta)P(w|\alpha)dw$

$$P(t|\alpha,\beta) = (2\pi)^{-N/2}|\beta^{-1}I+\alpha^{-1}\Phi\Phi^T|^{-1/2}\exp\left\{-\frac{1}{2}t^T(\beta^{-1}I+\alpha^{-1}\Phi\Phi^T)^{-1}t\right\}.$$

Find the most probable hyperparameters α and β:

$$P(\alpha,\beta|t) = \frac{P(t|\alpha,\beta)P(\alpha,\beta)}{P(t)}. \quad (1)$$

Maximize the marginal likelihood:

$$\alpha_{new} = \frac{\gamma}{\|w\|^2} \quad \text{and} \quad \beta_{new} = \frac{N-\gamma}{\|t-\Phi w\|^2}, \quad (2)$$

where

$$\gamma = M - \alpha\,\mathrm{tr}\Sigma \quad (3)$$

Third Level of Inference

Find the most probable model:

$$P(\mathcal{H}_i|t) = \frac{P(t|\mathcal{H}_i)P(\mathcal{H}_i)}{P(t)},$$

where $P(\mathcal{H}_i)$ is the prior probability of model $\mathcal{H}_i$.
The evidence is an intractable integral

$$P(t|\mathcal{H}_i) = \int P(t|\alpha, \beta, \mathcal{H}_i)P(\alpha, \beta|\mathcal{H}_i)d\alpha d\beta,$$

which Mackay approximates with a separable Gaussian around $P(t|\alpha_{MP}, \beta_{MP}, \mathcal{H}_i)$:

$$P(t|\mathcal{H}_i) \simeq P(t|\alpha_{MP}, \beta_{MP}, \mathcal{H}_i)P(\alpha_{MP}, \beta_{MP}|\mathcal{H}_i)2\pi\sqrt{\sigma^2_{\log\alpha}\sigma^2_{\log\beta}},$$

Outline

Search Strategies

Commonly used in feature selection. Some examples:

PTA(l,r) Plus *l* and Take Away *r*. Forward Selection is equivalent to PTA(1,0)

SFFS Sequential Forward Floating Selection.

Oscillating(c) Oscillate around a fixed number of elements.

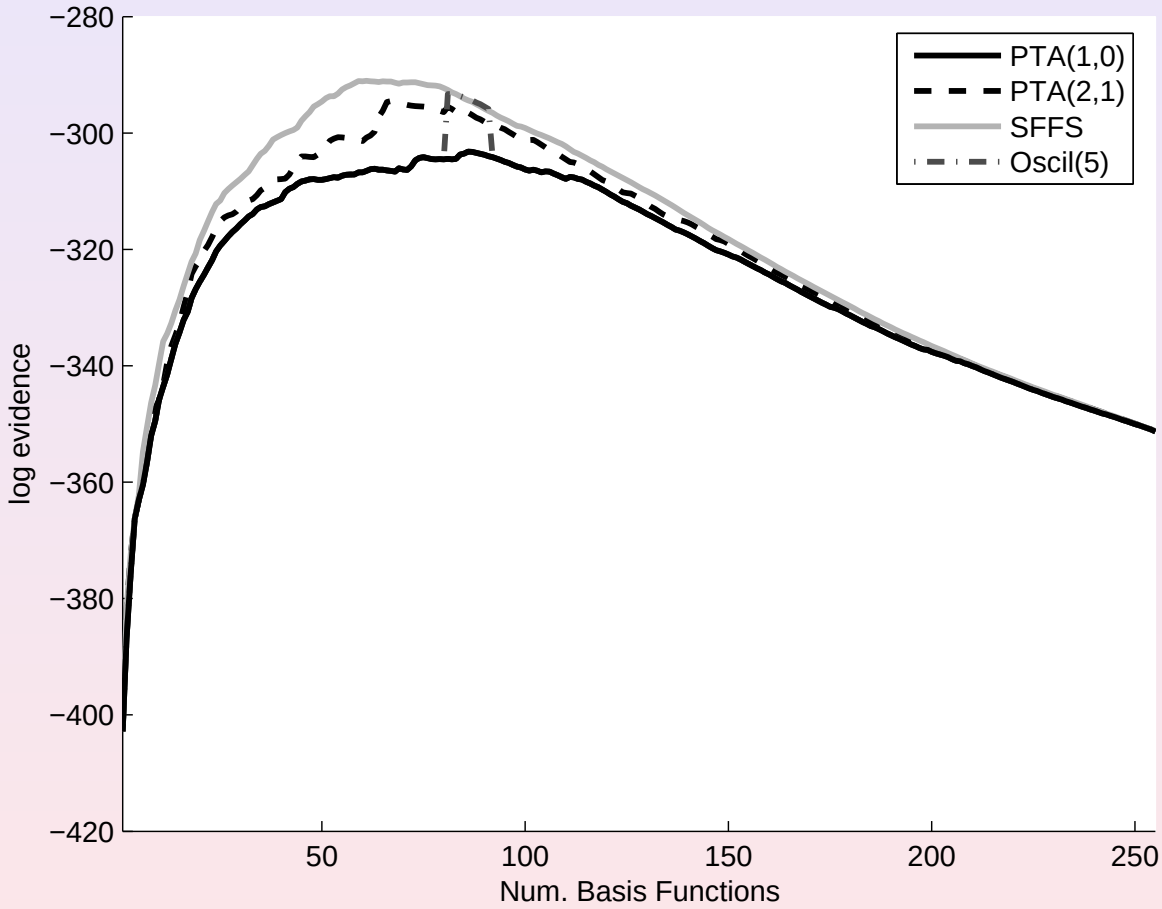
Outline

Implementation

- Operations to add the best basis and to remove the worst basis (according to the evidence).
- Create a dictionay of basis functions (e.g., RBF centered at the input data).
- Select the search strategy.

Outline

Experiments



SFFS:

- highest evidence
- lowest number of basis functions

Experiments

Method	pumadyn-8				kin-8			
	fh	fm	nh	nm	fh	fm	nh	nm
PTA(1,0)	39	80	59	68	48	71	120	264
PTA(2,1)	11	59	24	34	39	46	103	203
SFFS	2.5	4.5	8.5	13	10	22	76	159
Oscillating(5)	39	80	59	68	48	71	120	264
RVM	3.5	6.5	9.0	13	9.2	24	90	185
SVM	593	726	668	608	708	667	682	836
OLS	11	14	15	26	18	35	49	169

Table: Mean number of basis functions of the resulting models on different tasks.

Experiments

PTA(1,0)	-	0	0	6	3	6	4	0
PTA(2,1)	3	-	0	3	1	9	4	0
SFFS	9	5	-	11	3	13	6	0
Oscillating(5)	4	0	0	-	0	7	1	0
RVM	8	3	1	6	-	12	3	0
ABF	1	1	0	2	2	-	2	0
SVM	5	2	0	6	2	5	-	0
OLS	21	17	8	19	13	19	11	-

Table: The left-to-right order is the same as the top-to-bottom. Each cell shows the number of tasks where the column method performed better (p-value lower than 0.05) than the row method.

Discussion

- SSGE are competitive with RVM and SVM.
- The evidence prefers simpler models.
- Highest evidence does not correspond with lowest error.
- $P(\mathcal{H}_i)$ should not be the same for all the models. $P(\mathcal{H}_i|t) \propto P(\mathcal{H}_i)P(t|\mathcal{H}_i)$. But the assumption is practical.

Training Time	Model Size	Generalization error
PTA(1,0)	SFFS	Oscillating(5)
PTA(2,1)	PTA(2,1)	PTA(1,0)
Oscillating(5)	PTA(1,0) / Oscillating(5)	PTA(2,1)
SFFS	PTA(1,0) / Oscillating(5)	SFFS

Table: The methods ordered for each preference.

Thank you!