Ludwig-Maximilians-Universitaet Muenchen 26.06.2014
Institute for Informatics
Prof. Dr. Volker Tresp
Gregor Jossé

Machine Learning and Data Mining

Summer 2014

Exercise Sheet 7

Presentation of Solutions to the Exercise Sheet on the 26.06.2014

Aufgabe 7-1 Model Comparison

Compare the models of regression and basis functions. Let the prediction for a data point $\mathbf{x}_i \in \mathbb{R}$ be given as:

$$f(\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^{M_\Phi} w_j \phi_j(\mathbf{x}_i)$$

Employ the PLS-solution $\hat{\mathbf{w}} = \left(\Phi^T\Phi + \lambda I\right)^{-1} \Phi^T \mathbf{y}$ mit $\Phi_{i,j} = \phi_j(\mathbf{x}_i) = \mathbf{x}_i^{j-1}$. The following dataset $\mathbf{X}, \mathbf{y}$ of size $N = 10$ with variance $\sigma^2 = 0.25$ is given:

$\mathbf{X}$	0.3	0.4	0.8	1.5	1.8	3.6	4	4.3	4.6	5
$\mathbf{y}$	7	4.7	0.6	−1.1	−0.3	4.6	5.5	5.7	3.1	−0.3

We want to find the optimal model with basis functions $M_\Phi \in \{1, \ldots, 6\}$. Employ the mean squared error (MSE) as loss-function.

a) Find the best model using cross-validation (5 and 10 times). Do the pairwise tests introduced in the lecture support the decision of the MSE? What influence does the λ-parameter have?

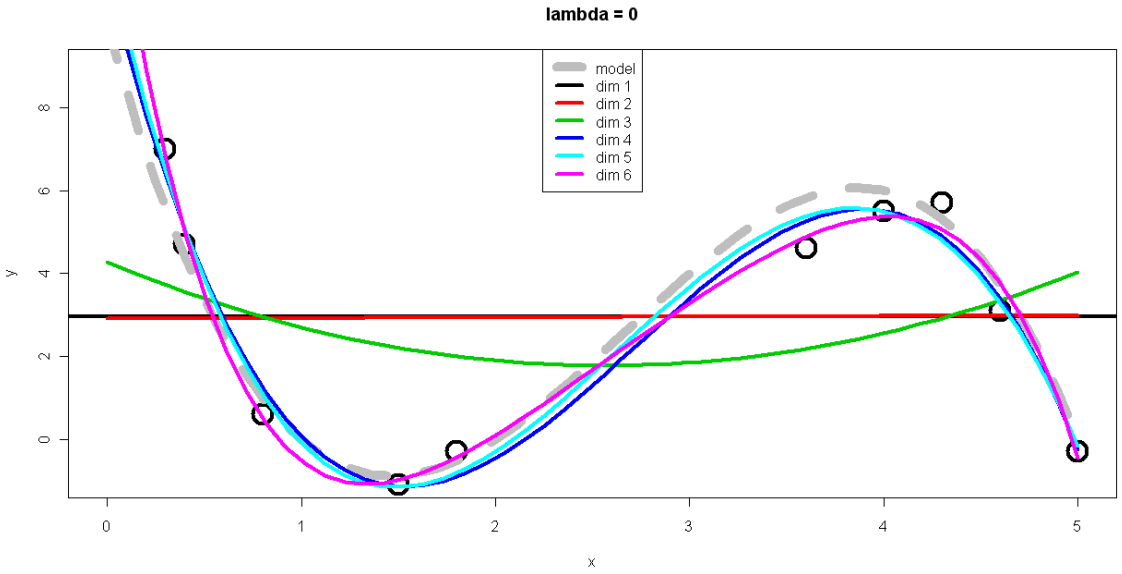
b) Which result do the frequentistic (C_p statistic and AIC) and the bayesian approach produce?

c) Which influence does the data size N have, if you were to simulate a comparable data set with $N = \{100, 1000\}$?

Lösungsvorschlag:

Keep in mind: We optimize the possible models $\mathcal{M}_i$, not weight vectors!

$$\text{MSE}(\mathbf{X},\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i,\mathbf{w}) - \mathbf{y}_i)^2$$



a) Small reminder about Cross-Validation (CV):

$$J_k^{\text{Test}}(\mathcal{M}_i) = \text{MSE}(\mathbf{X}(k),\mathbf{w}) = \frac{1}{N_k}\sum_{i\in\text{Test }(\mathbf{X},k)}(f(\mathbf{x}_i,\mathbf{w}) - \mathbf{y}_i)^2$$

where the k-fold holds N_k objects.
Note that the lecture notes denote $J_k^{\text{Test}}(\mathcal{M}_i)$ by $\text{cost}_{\text{test}_k}[\hat{w} \mid \text{train}_k,\mathcal{M}_i]$, we use the first option for reasons of brevity.
Remember:

$$\textbf{mean}(\mathcal{M}_i) = \frac{1}{K}\sum_{k=1}^{K} J_k^{\text{Test}}(\mathcal{M}_i) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i\in\text{Test }(\mathbf{X},k)}(f(\mathbf{x}_i,\mathbf{w}) - \mathbf{y}_i)^2$$

$$\widehat{\text{Var}}(\textbf{mean}(\mathcal{M}_i)) = \frac{1}{K(K-1)}\sum_{k=1}^{K}(J_k^{\text{Test}}(\mathcal{M}_i) - \textbf{mean}(\mathcal{M}_i))^2$$

We restrict ourselves to the 10-fold CV, because its fragmentation is unique. In contrast, the non-unique 5-fold CV fragmentations yield variable results, which are – when averaged – slightly worse than the 10-fold results.

First, we consider the unregularized case, where $\lambda = 0$.

M_Φ	1	2	3	4	5	6
$\textbf{mean}(\mathcal{M}_i)$	9.79	12.69	19.51	0.66	1.21	3.48
$\widehat{\text{Var}}(\textbf{mean}(\mathcal{M}_i))$	2.17	3.07	7.62	0.20	0.34	2.85

Lösungsvorschlag:

zu a): Pairwise testing: $\mathcal{M}_i$ better than $\mathcal{M}_j$, if

$$\textbf{mean}(\mathcal{M}_i) + \widehat{\text{Var}}(\textbf{mean}(\mathcal{M}_i)) < \textbf{mean}(\mathcal{M}_j) + \widehat{\text{Var}}(\textbf{mean}(\mathcal{M}_j))$$

Test if standard deviations overlap:

M_Φ	1	2	3	4	5	6
1		F	F	F	F	F
2	F		F	F	F	F
3	F	F		F	F	F
4	T	T	T		T	F
5	T	T	T	F		F
6	T	T	T	F	F	

$$\text{MeanDiff}_{i,j} = \frac{1}{K}\sum_{k=1}^{K}\left(J_k^{\text{Test}}(\mathcal{M}_i) - J_k^{\text{Test}}(\mathcal{M}_j)\right)$$

Test if standard deviations overlap:

M_Φ	1	2	3	4	5
2	2.9				
3	9.7	6.8			
4	−9.1	−12.0	−18.8		
5	−8.6	−11.5	−18.3	0.5	
6	−6.3	−9.2	−16.0	2.8	2.3

Pairwise T-tests w.r.t. MSE :

M_Φ	1	2	3	4	5
2	0.980				
3	0.905	0.848			
4	0.001	0.002	0.018		
5	0.002	0.002	0.019	0.969	
6	0.047	0.015	0.006	0.820	0.785

Reminder (?) regarding idea of the pairwise T-test: Compute pairwise differences between input vectors and test if the expected value of these differences complies with some hypothesis. In our case that hypothesis is: "$\mathcal{M}_i$ is better than $\mathcal{M}_j$", i.e. "the errors that $\mathcal{M}_i$ produces are smaller than the errors $\mathcal{M}_j$ produces, i.e. "MSE$_i$− MSE$_j < 0$".
This hypothesis is tested by means of the gaussian distribution (P-value = probability that $P(X <= \text{avg}(\text{MSE}_i - \text{MSE}_j))$.

$\Rightarrow$ the best model w.r.t. all quality measures is $M_\Phi = 4$, i.e. the basis transformation $(1, x, x^2, x^3)$.

Now let us investigate the influece of different λ (Reminder: $\lambda = \frac{\sigma^2}{\alpha^2}$ where α=variance of $\mathbf{w}$):

$\lambda = .01$**:** stabilizes $M_\Phi > 4$; all other models are degraded.

$\lambda = .05$**:** as above but with a stronger effect: now, $M_\Phi = 5$ is the best model.

$\lambda = .25$**:** as above, still $M_\Phi = 5$ is the best model..

In no case are the globally best results better than those without regularization.

Lösungsvorschlag:

b) Now, we refrain from splitting into trainings- and testset, but rely on the application of frequentist and bayesian measures.

Mallot's C_P statistic (Slide 37): $J^{\text{Train}} + 2\frac{M}{N}\sigma^2 \approx \frac{M+N}{N-M} J^{\text{Train}} = \frac{M+N}{N-M} \cdot \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i,\mathbf{w}) - \mathbf{y}_i)^2$

Akaikes Information Criterion (AIC, Slide 43): $\text{AIC} = \frac{1}{\sigma^2} C_P$

Bayesian Information Criterion (BIC): $\text{BIC} = N \cdot \text{AIC} - 2\,M + M \log N$, because:

$$\text{BIC (Sl. 52)} = -2\,\log L + M \log N$$
$$\text{AIC (Sl. 36)} = 2\left(-\frac{1}{N}\log L + \frac{M}{N}\right) \qquad | \cdot N \pm M \log N$$
$$N \cdot \text{AIC} = -2\,\log L + 2\,M + M \log N - M \log N$$
$$\Rightarrow \text{BIC} = N \cdot \text{AIC} - 2\,M + M \log N$$

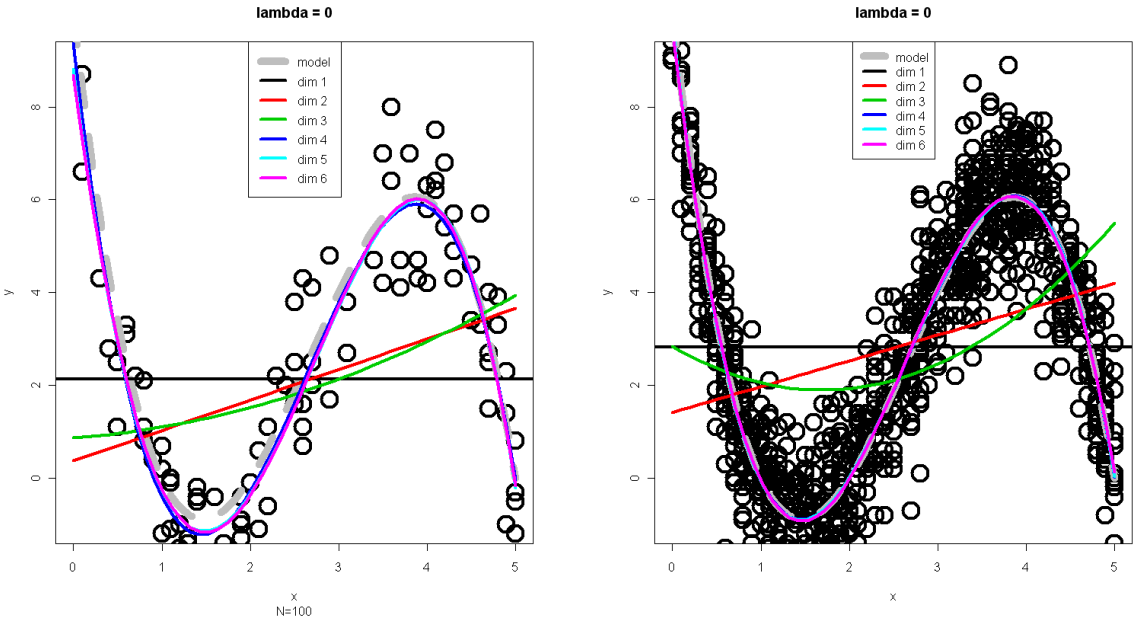
Results:

M_Φ	1	2	3	4	5	6
J^{Train}	7.93	7.93	7.48	0.25	0.24	0.11
C_P	11.90	14.73	17.46	0.76	0.97	0.65
AIC	74.60	58.92	69.85	3.03	3.89	2.59
BIC	746.6	590.1	699.7	31.8	40.7	28.0

$\Rightarrow C_P$ prefers $M_\Phi = 6$ over 4, i.e. a polynomial of degree 5. AIC and BIC continue along with this recommendation.

However, with the slightest regularization ($\lambda = 0.01$), $M_\Phi = 4$ is favored again, because even small regularization terms have big impact on the training error of complex models. Stronger regularization ($\lambda > 0.05$) shifts the decision in the direction of more complex models.

c) $N \in \{100, 1000\}$: $y = (1-x) \cdot (2-x) \cdot (5-x)$



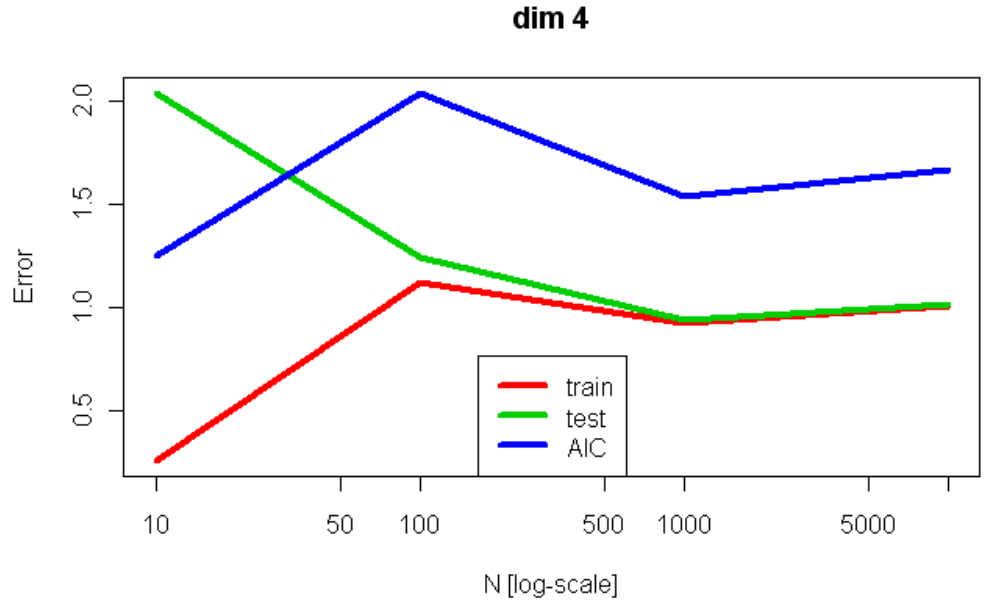
$\Rightarrow$ Better approximation to the original model (grey lines).

Lösungsvorschlag:

zu c) The insights from CV are practically the same, however, the P-values are significantly better. $N = 1000$ is superior to $N = 100$ (and worse than $N = 10000$).

Without CV: Training error is almost the same for $M_\Phi >= 4$, $\Rightarrow 4$ is favored. The same goes for regularization, where also $M_\Phi = 4$ is favored.

Generally: The bigger N, the smaller the test errors, but the bigger the training errors.



Concludingly:

	J^{Train}	J^{Test}
$N \uparrow$	$\uparrow$	$\downarrow$
$M \uparrow$	$\downarrow$	$\uparrow$