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

18.06.2015

Machine Learning and Data Mining

Summer 2015

Exercise Sheet 9

Presentation of Solutions to the Exercise Sheet on the 24.06.2015

Exercise 9-1 Modellvergleich

Vergleichen Sie Modelle der Regression mit Basisfunktionen. Die Vorhersage für einen Datenpunkt $\mathbf{x}_i \in \mathbb{R}$ sei gegeben durch:

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

Verwenden Sie die PLS-Lösung $\hat{\mathbf{w}} = \left(\Phi^T\Phi + \lambda I\right)^{-1}\Phi^T\mathbf{y}$ mit $\Phi_{i,j} = \phi_j(\mathbf{x}_i) = \mathbf{x}_i^{j-1}$. Gegeben sei der Datensatz $\mathbf{X}, \mathbf{y}$ der Größe $N = 10$, basierend auf einer Rauschvarianz von $\sigma^2 = 0.25$:

$\mathbf{X}$	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

Es soll das beste Modell für Basisfunktionen mit $M_\Phi \in \{1, \ldots, 6\}$ bestimmt werden. Als Loss-Funktion nehmen Sie im folgenden den mittleren quadratischen Fehler (MSE).

a) Bestimmen Sie das beste Modell durch Kreuzvalidierung (5-fach und 10-fach). Unterstützen die paarweisen Tests aus der Vorlesung die Entscheidung des MSE? Welchen Einfluss hat der λ-Parameter?

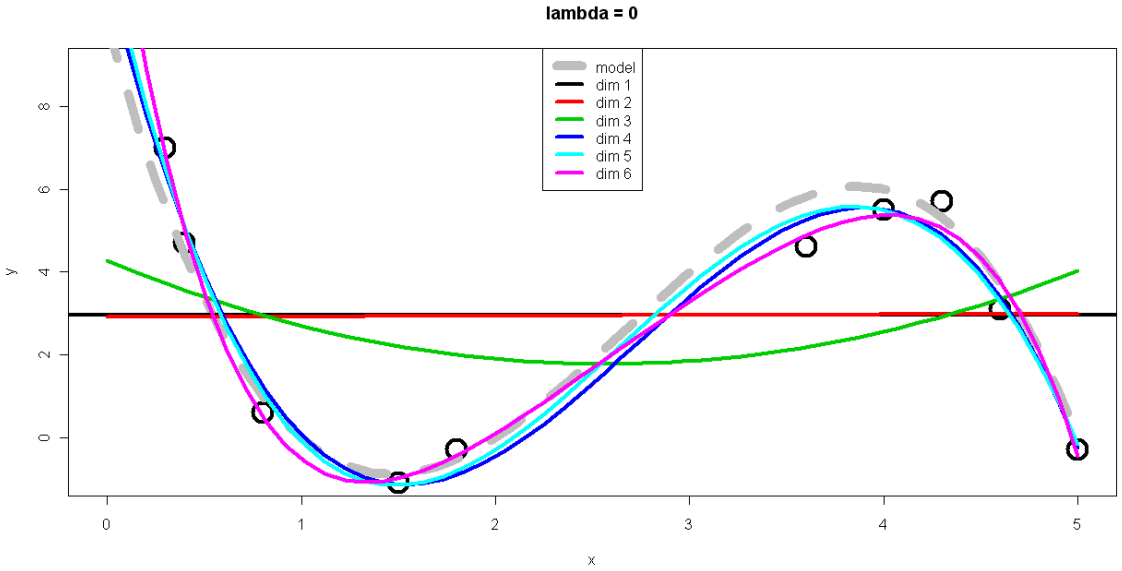
b) Kommen Sie mit den frequentistischen Verfahren (C_p Statistik und AIC) und dem Bayes'schen Verfahren (BIC) zu den gleichen Schlüssen?

c) Welchen Einfluss hat die Datengröße N, wenn Sie einen vergleichbaren Datensatz für $N = \{100, 1000\}$ simulieren?

Possible Solution:

Achtung: i für $\mathcal{M}_i$ hat hier eine andere Bedeutung als in Bezug auf $\mathbf{x}_i$!

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



a)

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

mit N_k Objekten in Fold k.

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

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

10-fold CV hier eindeutig, da es nur 10 Beispiele gibt. Die Ergebnisse von 5-fold variieren stark pro Durchlauf und sind generell etwas schlechter.

Wir testen zunächst mit $\lambda = 0$, also der nicht-regularisierten Variante.

M_Φ	1	2	3	4	5	6
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

b.w.

Possible Solution:

zu a): Paarweise Entscheidungsregel: $\mathcal{M}_i$ besser als $\mathcal{M}_j$ wenn

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

Test auf Überlappung der Standardabweichung:

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 auf Überlappung der Standardabweichung:

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

T-Tests auf die MSEs pro CV-fold:

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

Zum gepaarten T-Test: Bildet paarweise Differenzen der Eingabevektoren und testet ob deren Erwartungswert einer Hypothese entspricht – in unserem Fall: "$\mathcal{M}_i$ ist besser als $\mathcal{M}_j$", also "die Fehler von $\mathcal{M}_i$ sind kleiner als die Fehler von $\mathcal{M}_j$", also "$\text{MSE}_i - \text{MSE}_j < 0$". Diese Hypothese wird anhand der Normalverteilung überprüft (P-value: Zufallswahrscheinlichkeit, dass $P(X <= \text{avg}(\text{MSE}_i - \text{MSE}_j))$; entsprechende Methode in `R`: `t.test`.

$\Rightarrow$ der Gewinner ist nach alle Qualitätsmaßen $M_\Phi = 4$, also die Basistransformation $(1, x, x^2, x^3)$.

Effekt von alternativen λ (zur Erinnerung aus Blatt 3: $\lambda = \frac{\sigma^2}{\alpha^2}$ mit α=Streuung von $\mathbf{w}$):

$\lambda = .01$**:** stabilisert $M_\Phi > 4$; restliche Modelle werden eher verschlechtert.

$\lambda = .05$**:** s.o. nur verstärkter Effekt; jetzt ist $M_\Phi = 5$ am besten.

$\lambda = .25$**:** s.o. nur deutlicher: $M_\Phi = 5$ am besten.

Allerdings sind in keinem Fall die global besten Ergebnisse besser als ohne Regularisierung.

Zusammenhang Varianz/Bias:	hohe k (Anzahl von Folds):	hohe Varianz, kleiner Bias
	niedrige k:	kleine Varianz, großer Bias

b.w.

Possible Solution:

b) Hier soll nun auf den Split in Trainings- und Testset verzichtet werden – die zugrundeliegende Likelihoodmodellierung (Bayes'sches Rauschmodell), sowie die Loss-Funktion (MSE) sind für die folgende Anwendung der frequ. und bayes. Maße Vorraussetzung.

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

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

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

$$
\begin{aligned}
\text{BIC} &= -2\,\log L + M \log N \\
\text{AIC} &= 2\left(-\frac{1}{N}\log L + \frac{M}{N}\right) \qquad | \cdot N \pm M \log N \\
N \cdot \text{AIC} &= -2\,\log L + 2\,M + M \log N - M \log N \\
\Rightarrow \text{BIC} &= N \cdot \text{AIC} - 2\,M + M \log N
\end{aligned}
$$

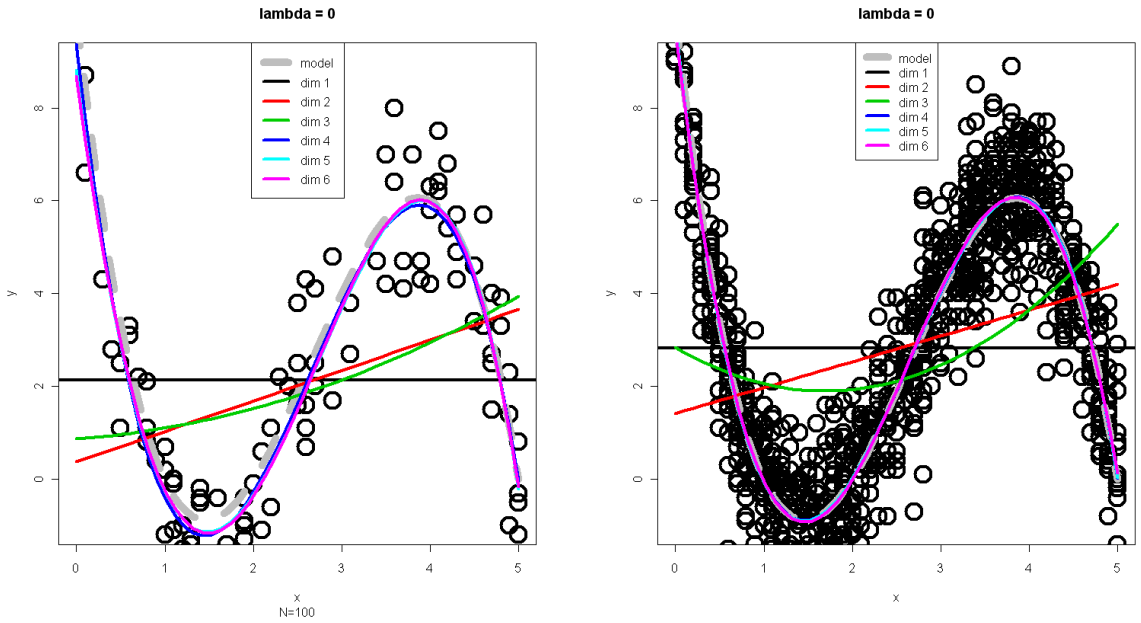
Ergebnisse:

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

$\Rightarrow$ Bereits C_P favorisiert eher $M_\Phi = 6$ als 4, also ein Polynom fünfter Ordnung. AIC und BIC folgen dieser Empfehlung im wesentlichen.

Kaum wird ein wenig regularisiert ($\lambda = 0.01$), gewinnt aber wieder $M_\Phi = 4$, da sich das fatal auf den Traininsfehler der komplexeren Modelle auswirkt. Bei stärkerer Regularisierung ($\lambda > 0.05$) verschiebt sich die Entscheidung wieder zugunsten der komplexeren Modelle.

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



$\Rightarrow$ Bessere Annäherung an das ursprüngliche Modell (graue Linien).

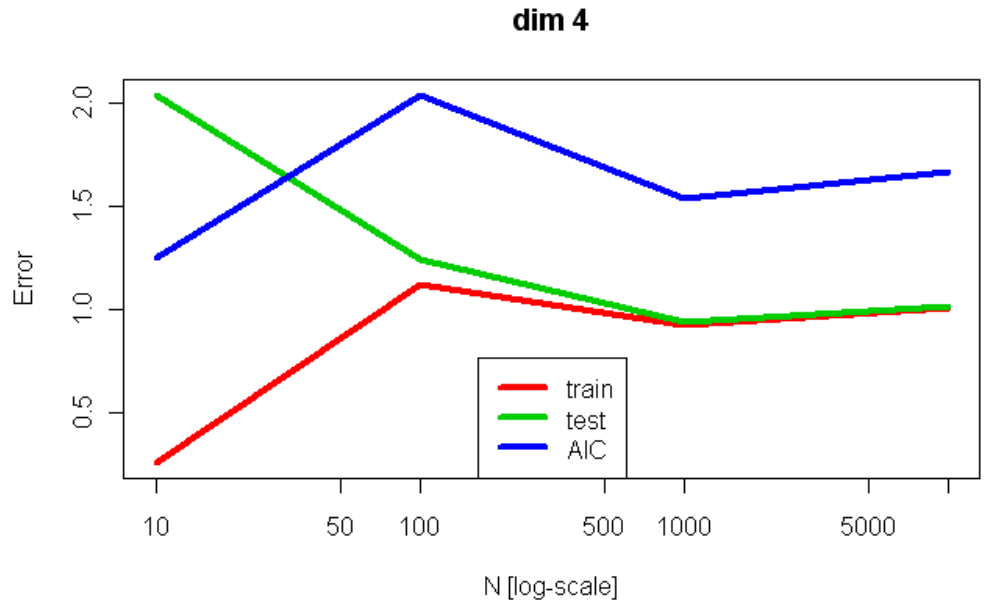
b.w.

Possible Solution:

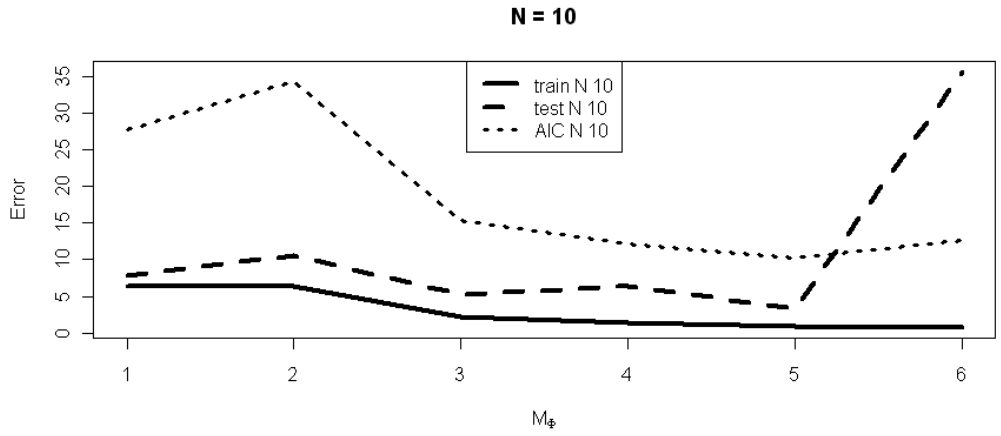
zu c) Die Erkenntnisse aus der CV sind im wesentlichen gleich – sogar mit schlechteren MSEs (das Sample der Größe 10 ist nicht sonderlich repräsentativ) – bei deutlich verbesserten P-Values im T-Test. $N = 1000$ ist besser als $N = 100$ (und schlechter als $N = 10000$). Regularisierung geht lange ($\lambda > .1$) zugunsten von $M_\Phi = 4$.

Ohne CV: Trainingsfehler quasi gleich für $M_\Phi >= 4$ (und sehr nahe am Testfehler aus der CV), $\Rightarrow 4$ gewinnt. Auch Regularisierung bringt am ehesten einen Profit für $M_\Phi = 4$.

Generell gilt: Größeres N, kleinere Testfehler, dafür größere Traininsfehler.



Sowie natürlich: Größeres D, kleinere Traininsfehler, dafür größere Testfehler. (Bild gehört zu anderem 10-er Sample)



Also:

	JTrain	JTest
$N \uparrow$	$\uparrow$	$\downarrow$
$M \uparrow$	$\downarrow$	$\uparrow$

Hauptgrund, warum die Zufalls-Samples sich nicht perfekt an die Theorie halten? Der Sample-Bereich ist sehr charakteristisch für die Verteilung $\Rightarrow$ Modell kann ungewöhnlich gut gefittet werden.