Regularisierung

M. Gruber

KW 49

Wir folgen [1], Lec 12.

Bei der Regularisierung wird ein modifizierter *in-sample error* minimiert. Regularisierung ist ein Mittel zur Vermeidung von Überanpassung.

Lösung eines einfachen Lernproblems mit Regularisierung

- Zielfunktion mit stochastischem Rauschen

$$y = f(x) + \varepsilon(x) \text{ mit } \mathbf{E}\,\varepsilon(x) = 0,\ \mathbf{Var}\,\varepsilon(x) = \sigma^2.$$

- Lernmenge

$$D = \left\{ \begin{bmatrix} x^{(i)} \\ y^{(i)} \end{bmatrix} \,\middle|\, i = 1, \dots, N \right\} \subset [-1, 1] \times \mathbf{R}.$$

- Hypothesenraum: Polynome Q-ter Ordnung,

$$H_Q = \left\{ \sum_{0 \le q \le Q} w_q \tilde{L}_q \,\middle|\, \tilde{L}_q \text{ normierte Legendre-Polynome}, w \in \mathbf{R}^{Q+1} \right\}.$$

  Es besteht eine 1-zu-1-Beziehung zwischen den $h \in H_Q$ und den Gewichten $w \in \mathbf{R}^{n+1}$.

- *Erweiterter (augmented) in-sample error*

$$\mathrm{Err}_{\mathrm{aug}}(w) = \tfrac{1}{N}\|Zw - y\|^2 + \tfrac{\lambda}{N}\|w\|^2 = \tfrac{1}{N}(Zw - y)^T(Zw - y) + \tfrac{\lambda}{N} w^T w.$$

  Dabei ist $Z = \begin{bmatrix} \tilde{L}_0(x^{(1)}) & \dots & \tilde{L}_Q(x^{(1)}) \\ \vdots & \ddots & \vdots \\ \tilde{L}_0(x^{(N)}) & \dots & \tilde{L}_Q(x^{(N)}) \end{bmatrix}$ und $\lambda > 0$ der "geeignet gewählte" Regularisierungsparameter.

- Der Lernalgorithmus minimiert $\mathrm{Err}_{\mathrm{aug}}(w)$, d.h. hier: er löst

$$\tfrac{N}{2} \nabla \mathrm{Err}_{\mathrm{aug}}(w) = Z^T(Zw - y) + \lambda w = 0.$$

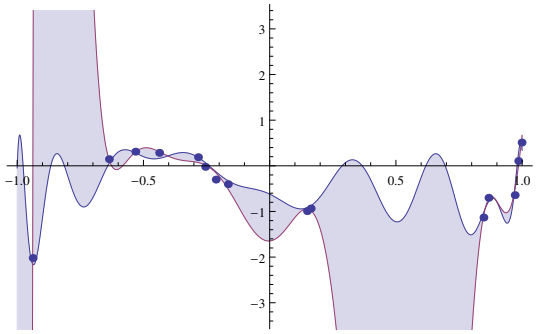
Die Lösung ist

$$w_{\text{aug}} = (Z^T Z + \lambda I)^{-1} Z^T y.$$

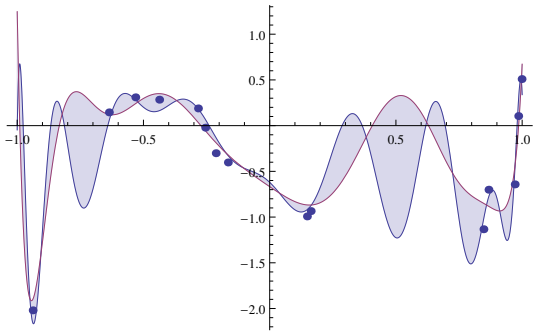
Diskussion

- Der Regularisierungsparameter $\lambda > 0$ entspricht einer Beschränkung $\|w\| \leq C$ des Hypothesenraums. Eine Zunahme von λ entspricht einer Abnahme von C.
- Mögliche Gefahren: λ zu klein—*overfitting*, λ zu groß—*underfitting*.
- Welcher λ-Wert ist gut? Nächste Vorlesung (über Validierung) abwarten!

Beispiel 1 *Zielfunktion (blau) ist ein Polynom 19.Ordnung. Die Lernmenge* ($N = 15$) *ist stochastisch verrauscht* ($\sigma = 0.1$). *Der Hypothesenraum besteht aus den Polynomen 11.Ordnung. Der out-of-sample-error der finalen Hypothese (rot) ist das Quadratintegral des schattierten Bereichs. Es beträgt* 1130.13.



Zum Vergleich dasselbe Problem, gelöst mit Regularisierung. Der Regularisierungsparameter $\lambda = 0.8$ *bewirkt eine Beschränkung der Gewichte* w_k. *Der out-of-sample-error beträgt hier* 0.274898.



Referenzen

[1] Y. S. Abu-Mostafa. Learning from Data.