Regularization

August, 2013

Lernziele

- Regularisierung
- *Capacity Control*
- Regularisierung für Lineare Regression und Logische Regression

Intuition: Regularization

Bekannt:

- Wenn $\mathcal{H}$ zu komplex ist, ist *Overfitting* wahrscheinlich.
- Polynome höheren Grades haben eine höhere Modellkomplexität als Polynome niedrigeren Grades.

Polynom 4-Grades:

$$h_\Theta(x) = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \Theta_3 x^3 + \Theta_4 x^4$$

- Zwangsbedingung (*Constaint*) $\Theta_3 = \Theta_4 = 0$ erniedrigt die Modellkomplexität.
- Zwangsbedingungen: $\Theta_3 \leq \epsilon$ und $\Theta_4 \leq \epsilon$ mit ϵ sehr klein $\rightarrow$ Modellkomplexität sollte nur wenig zugenommen haben.

Idee: Kostenfunktion Regularization

Kostenfunktion für Polynom 4-Grades, die kleine Werte für Θ_3 und Θ_4 erzwingt:

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)})^2 + \lambda' \Theta_3^2 + \lambda' \Theta_4^2$$

mit großem Hyperparameter λ.

Regularization bei Linear Regression

- Kleinere Werte für $\Theta_0, \Theta_1, \ldots, \Theta_n$ führen zu "einfacheren" Hypothesen $\rightarrow$ weniger "Overfitting"
- weiteres Beispiel "Hauspreise" mit vielen Features $\rightarrow$ einige Features tragen wenig Information und führen so zu schlechterer Generalisierung. Regularisierung $\Rightarrow$ Beschränkung der Θ-Werte unterdrückt diese Feature.

Kostenfunktion bei der Regularization

$$J(\Theta) = \frac{1}{m}\left[\sum_{i=1}^{m} loss(h_\Theta(x^{(i)}), y^{(i)}) + \frac{\lambda}{2}\sum_{j=1}^{n}\Theta_j^2\right]$$

mit

- λ: Regularization (Hyper-)Parameter - Kontrolliert die Komplexität des Modells
  - großes $\lambda \rightarrow$ niedrige Komplexität
  - kleines $\lambda \rightarrow$ höhere Komplexität
- Typischerweise ist Θ_0 nicht reguliert.

Augmented Error

statt J_{train} (Trainingsfehler) zu minimieren, wird ein erweiteter Fehler (*augmented Error*) minimiert:

$$J_{aug} = J_{train}(\Theta) + \frac{\lambda}{m}\Omega(\Theta) = J_{train}(\Theta) + overfitpenalty$$

Gradient Descent mit Regularization bei der linaren und logistischer Regression

Kostenfunktion

$$J(\Theta) = \frac{1}{m}\left[\sum_{i=1}^{m} loss(h_\Theta(x^{(i)}), y^{(i)}) + \frac{\lambda}{2}\sum_{k=1}^{n}\Theta_k^2\right]$$

ergibt mit der *Update Rule*

$$\Theta_j \leftarrow \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta)$$

bei der logistischen und lineare Regression:

für $j = 0$ (keine Veränderung)

$$\Theta_0 \leftarrow \Theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\Theta(\vec{x}^{(i)}) - y^{(i)})$$

und $j \neq 0$

$$\Theta_j \leftarrow \Theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \frac{\lambda}{m}\Theta_j\right]$$

Interpretation

Umformen der *Update Rule* für $j \neq 0$ ergibt

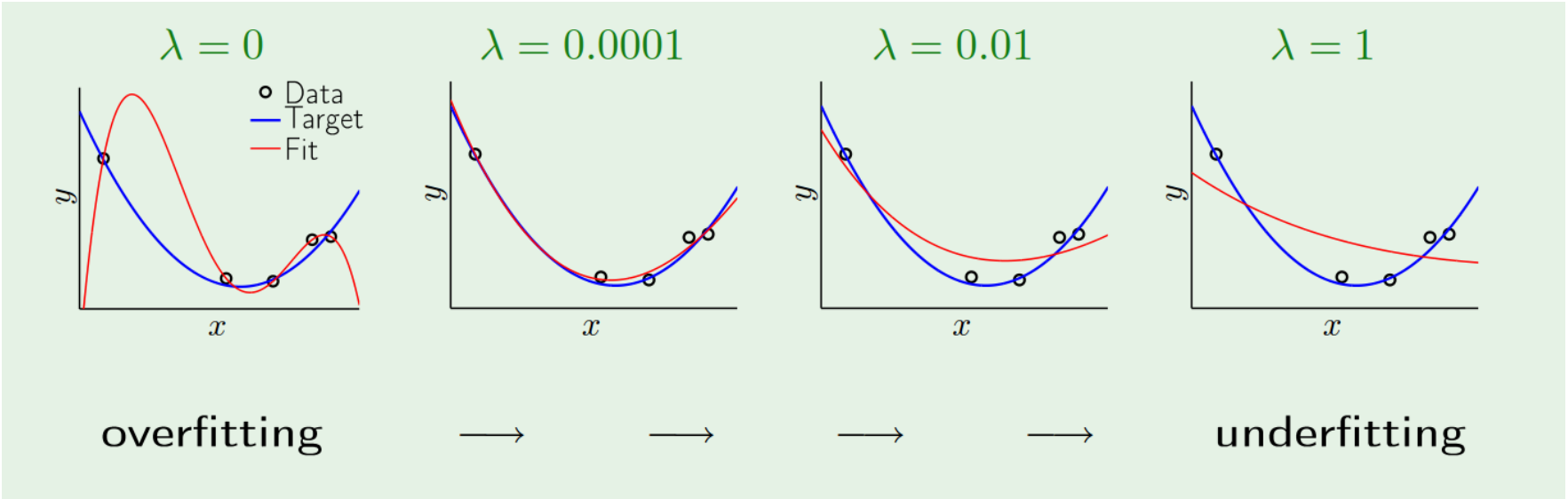
$$\Theta_j \leftarrow \Theta_j(1 - \alpha\frac{\lambda}{m}) - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\Theta(\vec{x}^{(i)}) - y^{(i)})x_j^{(i)}$$

Vergleich mit unregulierter *Update Rule* ergibt, dass Θ_j mit einem "Schrumpffaktor"

$$(1 - \alpha\frac{\lambda}{m}) < 1$$

multipliziert wird.

Gift oder Medizin

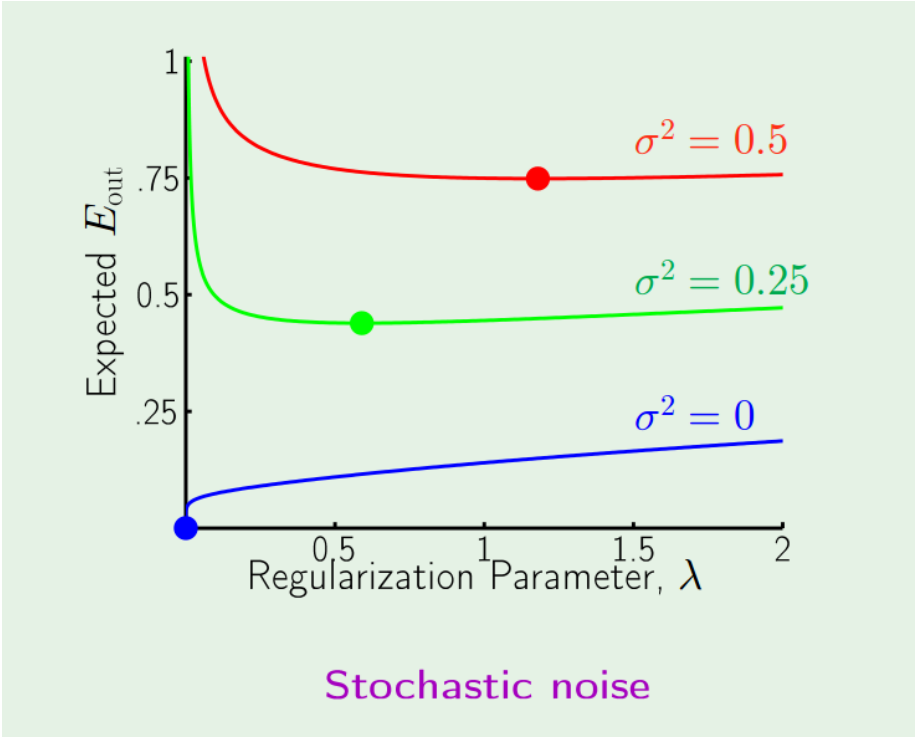


Quelle: [Yas]

Hypothesenmenge $\mathcal{H}$: Polynome 4. Grades;

5-Freiheitsgrade (effektive Parameter) erlauben ohne Regularisierung die 5. Trainingspunkte (leicht verrauscht) perfekt "zu treffen".

λ und Rauschen



Performance of the uniform regularizer at differnt levels of stochastic noise σ. Both target and model are polynominals of order 15. Quelle: [Yas]

Literaturangabe

- [Yas] Yaser Abu-Mostafa: Learning from Data, Caltech Machine Learning bzw.
  Yaser Abu-Mostafa et all.: Learning from Data, AMLBook 2012; siehe auch Foliensatz Lektion 12
- Andrew Ng: Machine Learning (Cousera Online Kurs), 2013

Weiterführende Literatur:

- Trevor Hastie,Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning, insb.: Kapitel 7, Springer Verlag 2009