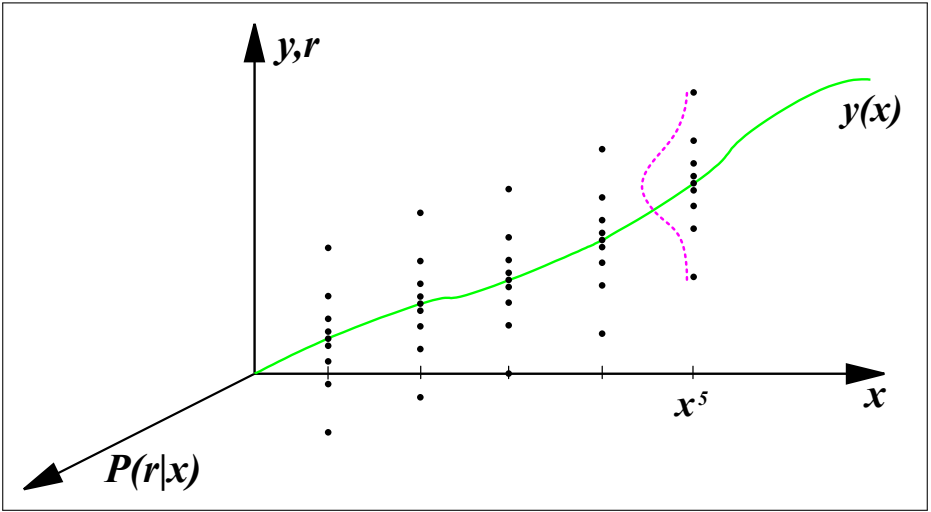
7. Statistische Analysen (Bias, Varianz)

7.1 Statistische Interpretation der Fehlerfunktion

Überblick:

- Varianz in den Solldaten
- Umstrukturierung der Fehlerfunktion
- Gewichtungsabhängiger/-unabhängiger Fehleranteil
- Annahmen/Forderungen zur Fehlerminimierung
- Rekonstruktion einer deterministischen Funktion
- Interpretation gewichtsunabhängiger Fehleranteil

Varianz in den Solldaten



Umstrukturierung der Fehlerfunktion

$$
\begin{aligned}
D(w) &:= \lim_{M\to\infty} \frac{1}{2M} \sum_{m=1}^{M} \sum_{i=1}^{N_L} (y_i(x^m, w) - r_i^m)^2 \\
&= \frac{1}{2} \sum_{i=1}^{N_L} \iint (y_i(x, w) - r_i)^2 P(r_i, x) dr_i dx \\
&= \frac{1}{2} \sum_{i=1}^{N_L} \iint (y_i(x, w) - r_i)^2 P(r_i|x) P(x) dr_i dx \\
&= \frac{1}{2} \sum_{i=1}^{N_L} \iint (y_i^2(x, w) - 2y_i(x, w) r_i + r_i^2) P(r_i|x) P(x) dr_i dx
\end{aligned}
$$

Umstrukturierung der Fehlerfunktion

$$= \frac{1}{2} \sum_{i=1}^{N_L} \underbrace{\iint y_i^2(x,w) P(r_i|x) P(x) dr_i dx}_{h_1} -$$

$$\frac{1}{2} \sum_{i=1}^{N_L} \underbrace{\iint 2 y_i(x,w) r_i P(r_i|x) P(x) dr_i dx}_{h_2} +$$

$$\frac{1}{2} \sum_{i=1}^{N_L} \underbrace{\iint r_i^2 P(r_i|x) P(x) dr_i dx}_{h_3}$$

Umstrukturierung der Fehlerfunktion

Es gilt für $\boldsymbol{h_1}$:

$$\begin{aligned} \boldsymbol{h_1} &= \int \boldsymbol{y_i^2}(\boldsymbol{x}, \boldsymbol{w}) \underbrace{\left(\int \boldsymbol{P}(\boldsymbol{r_i}|\boldsymbol{x}) \boldsymbol{dr_i} \right)}_{=1} \boldsymbol{P}(\boldsymbol{x}) \boldsymbol{dx} \\ &= \int \boldsymbol{y_i^2}(\boldsymbol{x}, \boldsymbol{w}) \boldsymbol{P}(\boldsymbol{x}) \boldsymbol{dx} \end{aligned}$$

Umstrukturierung der Fehlerfunktion

Es gilt für h_2:

$$h_2 = \int 2y_i(x,w) \underbrace{\left(\int r_i P(r_i|x) dr_i \right)}_{E(r_i|x) = \langle r_i|x \rangle} P(x) dx$$

$$= \int 2y_i(x,w) \langle r_i|x \rangle P(x) dx$$

Umstrukturierung der Fehlerfunktion

Es gilt für $\boldsymbol{h_3}$:

$$\begin{aligned} \boldsymbol{h_3} &= \int \underbrace{\left(\int \boldsymbol{r_i^2 P(r_i|x) dr_i} \right)}_{\boldsymbol{E(r_i^2|x) = \langle r_i^2|x \rangle}} \boldsymbol{P(x) dx} \\ &= \int \boldsymbol{\langle r_i^2|x \rangle P(x) dx} \end{aligned}$$

Umstrukturierung der Fehlerfunktion

Daraus folgt:

$$
\begin{aligned}
\boldsymbol{D(w)} &= \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_L}} (\boldsymbol{h_1 - h_2 + h_3}) \\
&= \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_L}} \Big(\int \boldsymbol{y_i^2(x,w) P(x) dx} - \int \boldsymbol{2 y_i(x,w) \langle r_i | x \rangle P(x) dx} \\
&\qquad + \int \boldsymbol{\langle r_i | x \rangle^2 P(x) dx} \Big) + \\
&\qquad \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_L}} \Big(\int \boldsymbol{\langle r_i^2 | x \rangle P(x) dx} - \int \boldsymbol{\langle r_i | x \rangle^2 P(x) dx} \Big)
\end{aligned}
$$

Umstrukturierung der Fehlerfunktion

Wir erhalten zwei Terme in der Fehlerfunktion:

$$\boldsymbol{D(w)} = \underbrace{\frac{1}{2}\sum_{i=1}^{N_L}\int(y_i(x,w) - \langle r_i|x\rangle)^2 P(x)dx}_{\text{Term I}} +$$

$$\underbrace{\frac{1}{2}\sum_{i=1}^{N_L}\int(\langle r_i^2|x\rangle - \langle r_i|x\rangle^2)P(x)dx}_{\text{Term II}}$$

Gewichtsabhängiger/-unabhängiger Fehleranteil

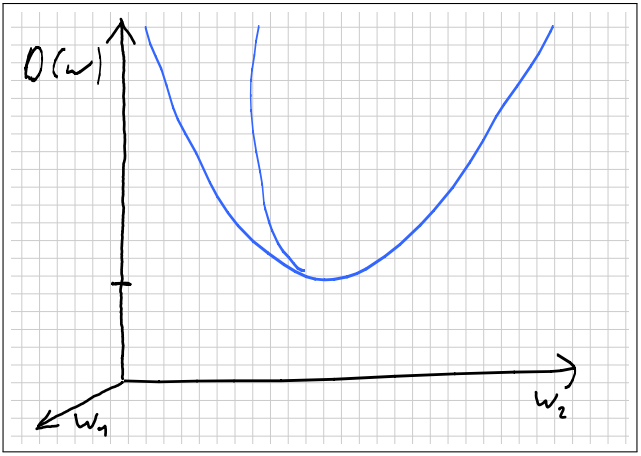
Term II:

Unabhängig von Abbildungsfunktion des Netzes, d.h. unabhängig von den gelernten Gewichten.

Term I:

Wenn $\boldsymbol{y_i(x, w^*) = \langle r_i | x \rangle}$, $\forall \boldsymbol{i \in \{1, \ldots, N_L\}}$, d.h. Integrand verschwindet, dann absolutes Minimum der Fehlerfunktion erreicht.

Gewichtsabhängiger/-unabhängiger Fehleranteil



Annahmen/Forderungen zur Fehlerminimierung

- Optimierung der Netzparameter, um Minimum der Fehlerfunktion zu finden.
- Abbildungsfunktion des neuronalen Netzes muß hinreichend flexibel sein, damit der minimale Fehler einen niedrigen Wert erreicht.
- Die Solldaten müssen möglichst zuverlässig und exakt sein, damit der minimale Fehler einen niedrigen Wert erreicht.

Rekonstruktion einer deterministischen Funktion

Annahmen: Solldaten r_i^m wurden durch verrauschte Funktion gebildet und x^m waren exakt meßbar, $\quad r_i^m = g_i(x^m) + \delta_i^m$;

Deterministische, unbekannte Funktion g_i;
Erwartungswert des Rauschens sei null, d.h $\langle \delta_i \rangle = 0$.

Nach optimalem Lernen gilt $\forall i \in \{1, \ldots, N_L\}$:

$$y_i(x, w^*) = \langle r_i | x \rangle = \langle (g_i(x) + \delta_i) | x \rangle = \underbrace{\langle g_i(x) | x \rangle}_{= g_i(x)} + \underbrace{\langle \delta_i | x \rangle}_{= 0} = g_i(x)$$

Das optimale Netz liefert bedingte Erwartungswerte der Zielfunktion, d.h. rekonstruiert den deterministischen Anteil dieser Funktion.

Interpretation gewichtsunabhängiger Fehleranteil

Term II ist unabhängig von Netzgewichten.

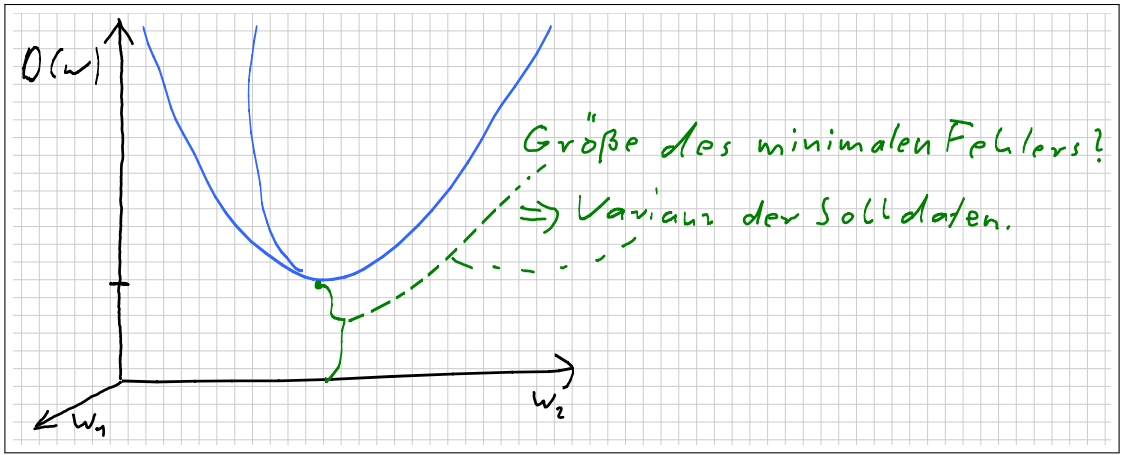
$$\langle r_i^2|x\rangle - \langle r_i|x\rangle^2 \stackrel{!}{=} \langle (r_i - \langle r_i|x\rangle)^2|x\rangle = \\ E\big((r_i - E(r_i|x))^2|x\big) = \sigma^2_{r_i}(x)$$

Untere Grenze des erreichbaren Fehlers ist stochastische Varianz (Rauschen) der Solldaten, bezogen auf Input-Trainingselement x, und i-te Komponente der Solldaten.

Summe der Erwartungswerte der Varianz der Solldaten:

$$\frac{1}{2}\sum_{i=1}^{N_L}\int(\langle r_i^2|x\rangle - \langle r_i|x\rangle^2)P(x)dx$$

Interpretation gewichtsunabhängiger Fehleranteil



7.2 Plastizitäts-/Stabilitätsdilemma

Überblick:

- Speicherung versus Generalisierung
- Komplexität des Modells
- Aufspalten der Lernmenge
- Formale Definition von Bias und Varianz
- Beispiele für Bias und Varianz
- Minimierung von Bias und Varianz

Speicherung versus Generalisierung

Das Neuronale Netz soll den realen Sachverhalt modellieren.

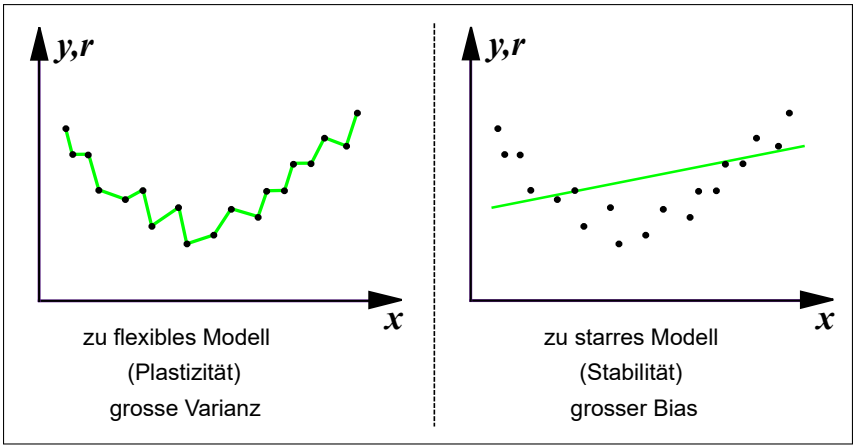
Das zu lernende Modell muß sowohl hinreichend gut die Daten widerspiegeln, als auch hinreichend stark von ihnen abstrahieren.

Speicherungsaspekt und Generalisierungsaspekt im Konflikt.

Die Netzstruktur muß hinreichend viele Freiheitsgrade für die Anpassung haben. Diese Freiheitsgrade müssen durch geeignete Zwänge eingeschränkt werden.

Speicherung versus Generalisierung

Balance zwischen Anpassung (des Modells an Trainingsdaten) und Abstraktion erforderlich.



Komplexität des Modells

Zwei Beispiele für Möglichkeiten zur Kontrolle der Komplexität des Modells (synonym Regularisierung).

Strukturelle Regularisierung:

- Änderung der Zahl der Schichten/Knoten des Netzes, sowie Art der Propagierungs- und Aktivierungsfunktion.

Belohnung/Bestrafung Parameterwerte:

- Zur Fehlerfunktion wird ein Regularisierungsterm hinzugezogen.

Aufspalten der Lernmenge

Problembereich Ω ist Vereinigung von Lernmenge Ω_L, Evaluationsmenge Ω_E und Anwendungsmenge Ω_A.

Naive, einmaliges Aufspalten der Lernmenge Ω_L in Trainingsmenge Ω_T und Validierungsmenge Ω_V, d.h. $\Omega_L := \Omega_T \cup \Omega_V$.

Naive Lernmethodik: Beim Trainieren des NN wird Ω_T verwendet, und beim Validieren das Ω_V.

Speicherkapazität (Gedächtnis): Erfolgreich nur für Ω_T und Ω_V, aber nicht für Ω_E? Dann hat das NN lediglich auswendig gelernt.

Generalisierungsfähigkeit (Abstraktion): Erfolgreich auch für Ω_E?

Aufspalten der Lernmenge

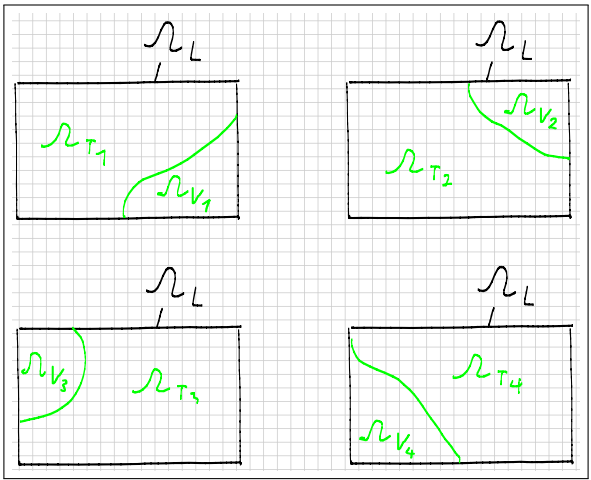
Bessere Lernmethodik:

- Variables Aufspalten der Lernmenge,
  $$\Omega_L := \Omega_{T_j} \cup \Omega_{V_j},\ j \in \{1, \cdots\}$$
- Systematisches oder zufälliges Verändern der Zusammensetzung von Trainings- und Validierungsmenge.
- Jeweils Lernen mit aktueller Trainings- und Validierungsmenge.
- Zusammenführen aller Ergebnisse.

Dieser sog. *Cross Validation* Ansatz verspricht eine bessere Generalisierungsfähigkeit.

Aufspalten der Lernmenge

Beispiel:



Formale Definition von Bias und Varianz

Bezug zu Term I der Fehlerfunktion:

Im folgenden wird Index i als Komponente eines möglicherweise mehrdimensionalen Outputvektors zur Vereinfachung weggelassen (aber ohne Beschränkung der Allgemeinheit).

Lernergebnis $y(x)$ hängt von der speziellen Wahl der Trainingsmenge Ω_{T_j} ab.

Unabhängigkeit erzielt man durch Bildung des Erwartungswertes des inneren Teils von Term I, durch Heranziehen mehrerer Ω_{T_j}, d.h. $E_T((y(x) - \langle r|x\rangle)^2)$.

Formale Definition von Bias und Varianz

Expandieren des inneren Teils von Term I:

$$
\begin{aligned}
(\boldsymbol{y}(\boldsymbol{x}) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)^2 &= (\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x})) + \boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x})) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)^2 \\
&= (\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x})))^2 + (\boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x})) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)^2 + \\
&\quad \underbrace{\boldsymbol{2}(\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x})))(\boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x})) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)}_{=: \boldsymbol{h}}
\end{aligned}
$$

Dabei ist $\boldsymbol{E}_T(\boldsymbol{y}(\boldsymbol{x}))$ der Erwartungswert von $\boldsymbol{y}(\boldsymbol{x})$, ermittelt über mehrere $\boldsymbol{\Omega}_{\boldsymbol{T}_j}$.

Formale Definition von Bias und Varianz

Bildung des Erwartungswerts der expandierten Teilformel:

$$\boldsymbol{E_T}((\boldsymbol{y}(\boldsymbol{x}) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)^2)$$

Es gilt: $\boldsymbol{E_T}(\boldsymbol{h}) \stackrel{!}{=} \boldsymbol{0}$

Begründung:

$$\boldsymbol{E_T}(\boldsymbol{y} \cdot \boldsymbol{E_T}(\boldsymbol{y}) - \boldsymbol{E_T^2}(\boldsymbol{y}) - \boldsymbol{y}\langle \boldsymbol{r}|\boldsymbol{x}\rangle + \boldsymbol{E_T}(\boldsymbol{y})\langle \boldsymbol{r}|\boldsymbol{x}\rangle) =$$

$$\boldsymbol{E_T}(\boldsymbol{y}) \cdot \boldsymbol{E_T}(\boldsymbol{y}) - \boldsymbol{E_T^2}(\boldsymbol{y}) - \boldsymbol{E_T}(\boldsymbol{y})\langle \boldsymbol{r}|\boldsymbol{x}\rangle + \boldsymbol{E_T}(\boldsymbol{y})\langle \boldsymbol{r}|\boldsymbol{x}\rangle = \boldsymbol{0}$$

Formale Definition von Bias und Varianz

Daraus folgt:

$$\boldsymbol{E_T}\big((\boldsymbol{y}(\boldsymbol{x}) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)^2\big) = \underbrace{\boldsymbol{E_T}\big((\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{E_T}(\boldsymbol{y}(\boldsymbol{x})))^2\big)}_{\text{Varianz}_y(\boldsymbol{x})} +$$

$$\boldsymbol{E_T}\big(\underbrace{(\boldsymbol{E_T}(\boldsymbol{y}(\boldsymbol{x})) - \langle \boldsymbol{r}|\boldsymbol{x}\rangle)^2}_{\text{Bias}_y^2(\boldsymbol{x})}\big)$$

Mittelung über alle $\boldsymbol{x}$:

$$(\text{Bias}_y)^2 \;:=\; \frac{1}{2}\int \text{Bias}_y^2(\boldsymbol{x})\boldsymbol{P}(\boldsymbol{x})d\boldsymbol{x}$$

$$\text{Varianz}_y \;:=\; \frac{1}{2}\int \text{Varianz}_y(\boldsymbol{x})\boldsymbol{P}(\boldsymbol{x})d\boldsymbol{x}$$

Formale Definition von Bias und Varianz

Bias: Abweichung zwischen dem Erwartungswert der Netzwerkfunktion und dem Erwartungswert der gefordeten Regressionsfunktion.

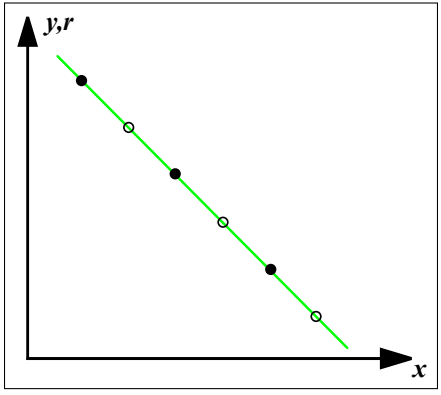
Varianz: Abhängigkeit der Netzwerkfunktion von spezieller Wahl der Trainingsmenge Ω_{T_j}.

Ziel: Minimierung von Bias und Varianz.

Bias-Varianz-Dilemma: Bias und Varianz sind oft im Konflikt zueinander. Das Lernverfahren muß eine Balance zwischen Bias und Varianz erreichen.

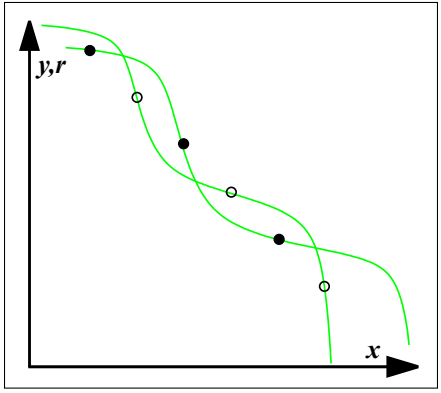
Beispiele für Bias und Varianz

a) Beispiel mit linearer Modellfunktion: keine Varianz und kein Bias, wenn Daten tatsächlich linear verteilt sind.

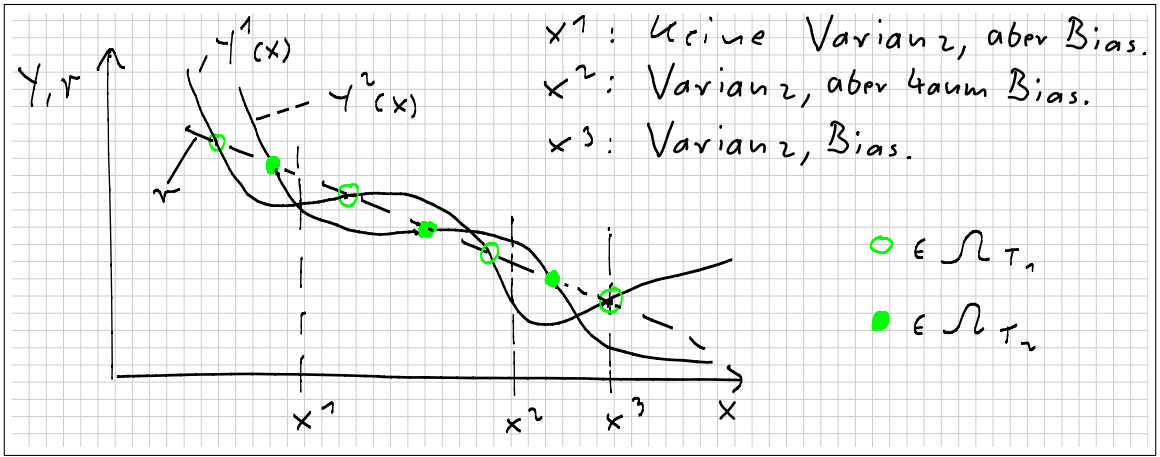


Beispiele für Bias und Varianz

b) Bsp. mit polynominaler Modellfunktion: Varianz und Bias abhängig von tatsächl. Verteilung der Daten und verwendetem Grad des Polynoms.



Beispiele für Bias und Varianz



Beispiele für Bias und Varianz

c) Beispiel mit einer festen Funktion $\overline{\boldsymbol{f}}(\boldsymbol{x})$ als Abbildungsfunktion $\boldsymbol{y}(\boldsymbol{x})$, wobei $\overline{\boldsymbol{f}}(\boldsymbol{x})$ unabhängig von $\boldsymbol{\Omega_T}$ sei. $\overline{\boldsymbol{f}}$ wird nicht gelernt, sondern a priori vorgegeben.

Varianz verschwindet, da $\boldsymbol{y}(\boldsymbol{x}) = \overline{\boldsymbol{f}}(\boldsymbol{x})$, und somit $\boldsymbol{E_T}(\boldsymbol{y}(\boldsymbol{x})) = \overline{\boldsymbol{f}}(\boldsymbol{x})$.

Bias ist i.A. hoch, da Abhängigkeit von Trainingsmenge nicht berücksichtigt wurde.

Beispiele für Bias und Varianz

d) Beispiel mit auswendig gelernten Funktionen $\boldsymbol{f_j(x)}$, die die Trainingsmengen $\boldsymbol{\Omega_{T_j}}$ jeweils perfekt widerspiegeln.

Für die Trainingselemente, die in der Schnittmenge der herangezogenen $\boldsymbol{\Omega_{T_j}}$ liegen, $\boldsymbol{j \in \{1, \cdots\}}$, ist Bias und Varianz gleich 0.

Für die übrigen Trainingselemente ist Bias und Varianz i.A. sehr hoch.

Minimierung von Bias und Varianz

- Komplexe Modelle verwenden, um Bias zu reduzieren, eventuell Vorwissen bezüglich unbekannter Modellfunktion einbringen.
- Größere Trainingsmengen $\boldsymbol{\Omega_{T_j}}$ führen zu geringerer Varianz.

Literatur

- C. Bishop: Neural Networks for Pattern Recognition; Kapitel 9, Oxford Press, 1995.