Was macht gute Vorhersagen aus?

Tobias Fissler

ETH Zürich & Gymnasium / FMS Lerbermatt

10. September 2025

35. Schweizerischer Tag über Mathematik und Unterricht

Kantonsschule am Burggraben, St. Gallen

Vorstellung

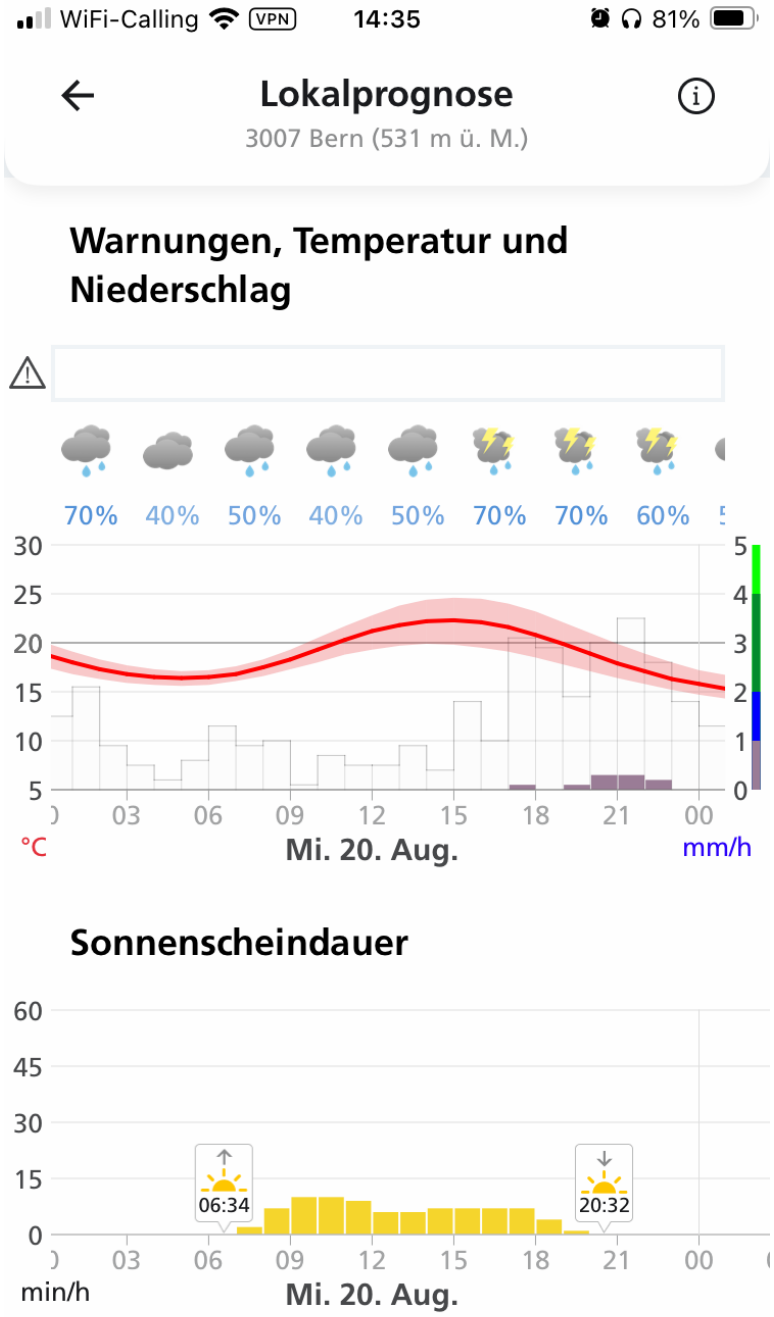
Tobias Fissler

- Aufgewachsen in Baden-Württemberg, in der Nähe von Stuttgart
- Studium der Mathematik (Nebenfächer Wirtschaftswissenschaften und Philosophie) an der Universität Heidelberg.
- PhD in Statistik an der Universität Bern
- Postdoc-Stellen am Imperial College London und an der Wirtschaftsuniversität Wien (Habilitation)
- Seit 2023 Postdoc an der ETH Zürich: Homepage
- 2023–2024: Lehrdiplom für Maturitätsschulen an der PH Bern
- Seit 2024 Mathematiklehrer am Gymnasium / FMS Lerbermatt in Köniz bei Bern

Beispiele von Vorhersagen

Welche Beispiele kommen euch in den Sinn?

Wettervorhersagen ...



...und ihre Auswirkungen

...und ihre Auswirkungen

...und ihre Auswirkungen

...und ihre Auswirkungen

Inflationsvorhersagen ...

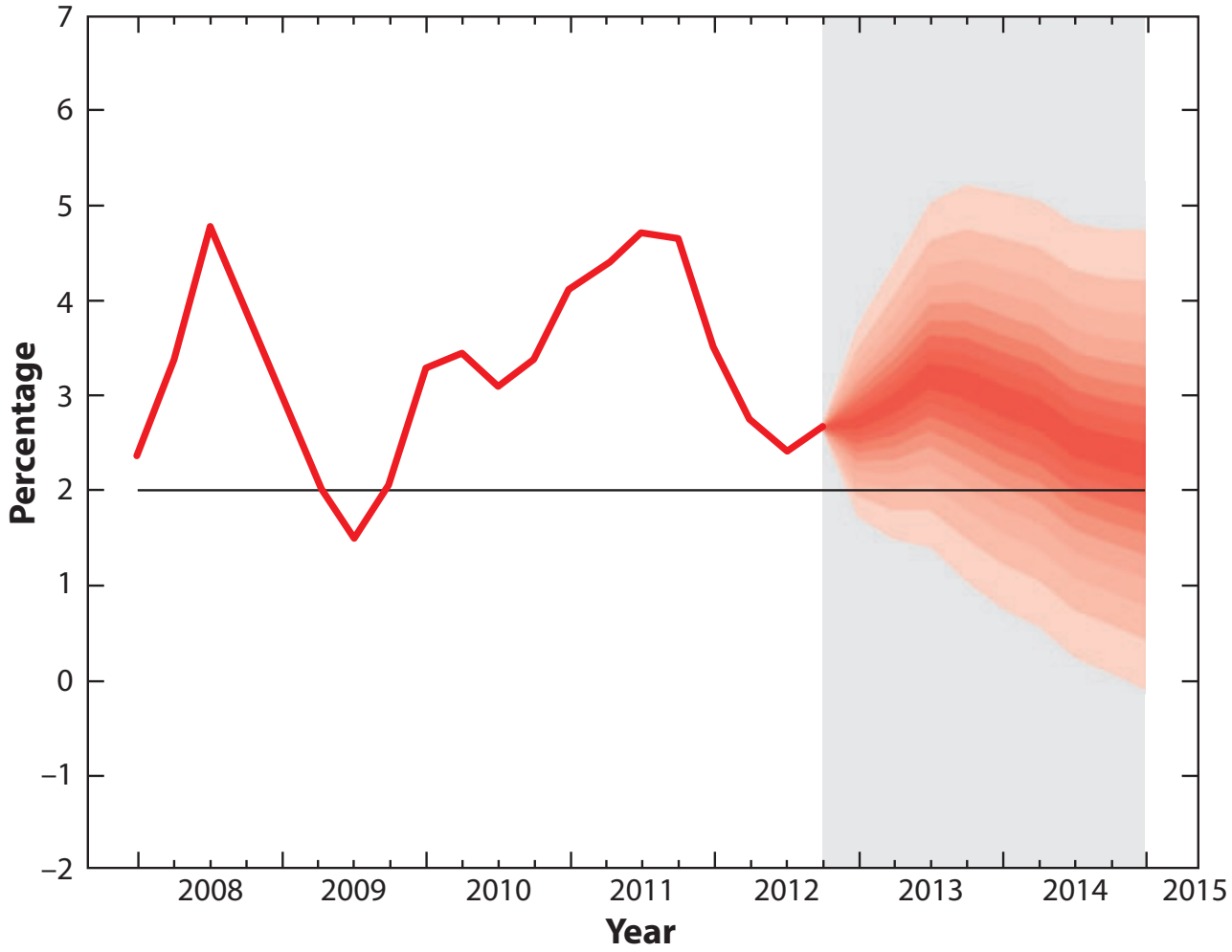
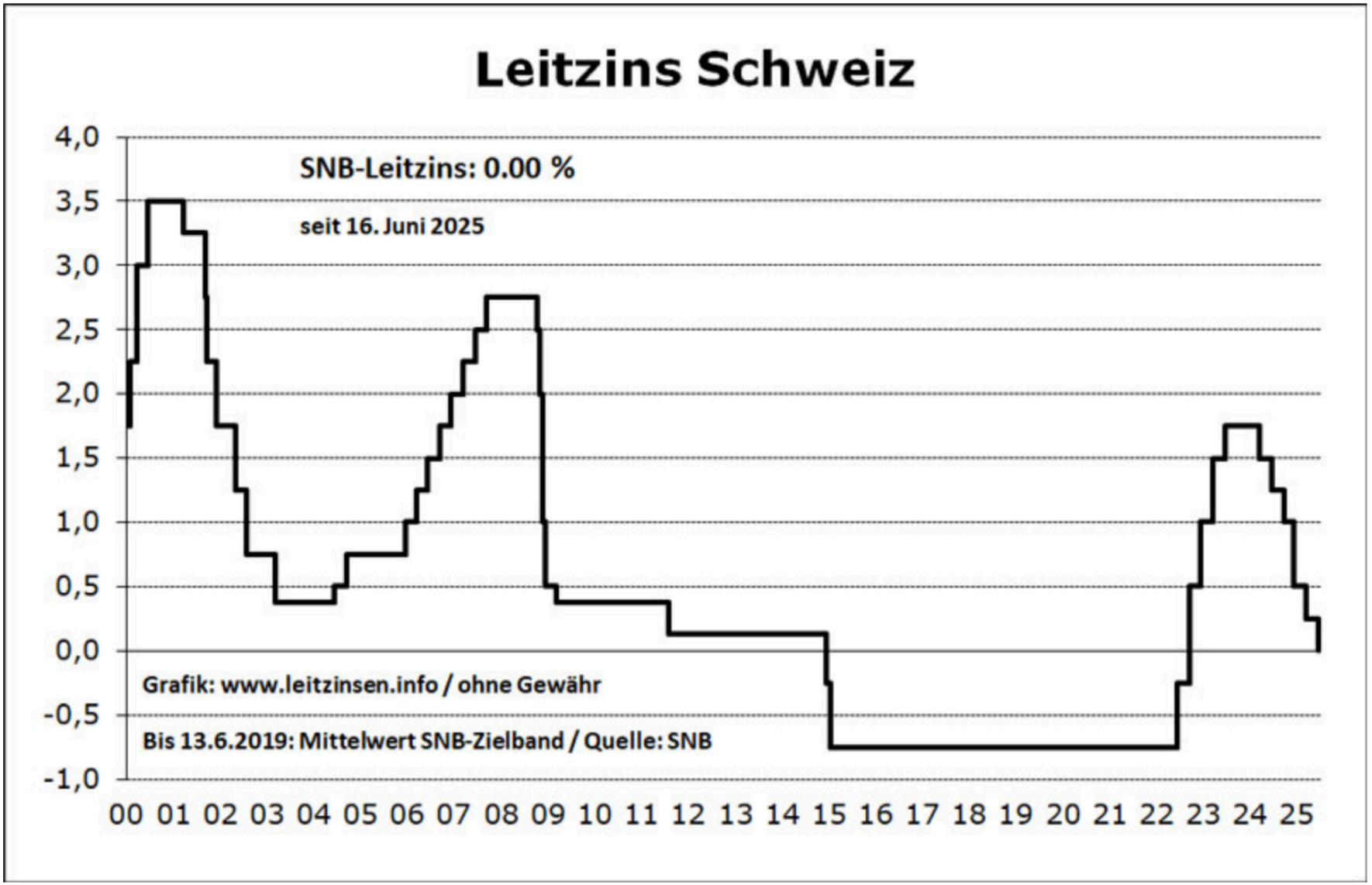


Figure 1

February 2013 Bank of England forecast of inflation in the United Kingdom as a percentage increase in the consumer price index (Bank of England 2013, with permission). The shaded bands in the fan chart show prediction intervals in increments of 10%.

...und ihre Auswirkungen



Programm

1. Welche Arten von Vorhersagen gibt es?
2. Wie werden Vorhersagen generiert?
3. **Was macht gute Vorhersagen aus?**
4. Ideen für den Unterricht

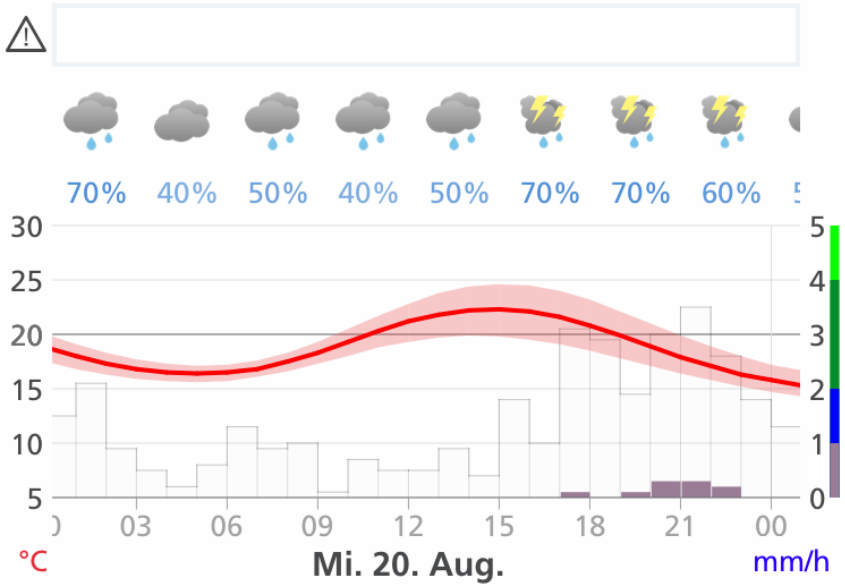
Arten von Vorhersagen

WiFi-Calling VPN 14:35 81%

Lokalprognose

3007 Bern (531 m ü. M.)

Warnungen, Temperatur und Niederschlag



Sonnenscheindauer

Arten von Vorhersagen

- **Punktvorhersagen:** Vorhersage besteht aus einer einzigen Zahl.
- **Probabilistische Vorhersagen:** Vorhersage besteht aus einer Wahrscheinlichkeitsverteilung.
  (Bei binärem Ereignis aus der Eintrittswahrscheinlichkeit)
- **Kompromiss:** Punktvorhersage zusammen mit Quantifizierung der Unsicherheit, z.B. in Form von Quantilen.

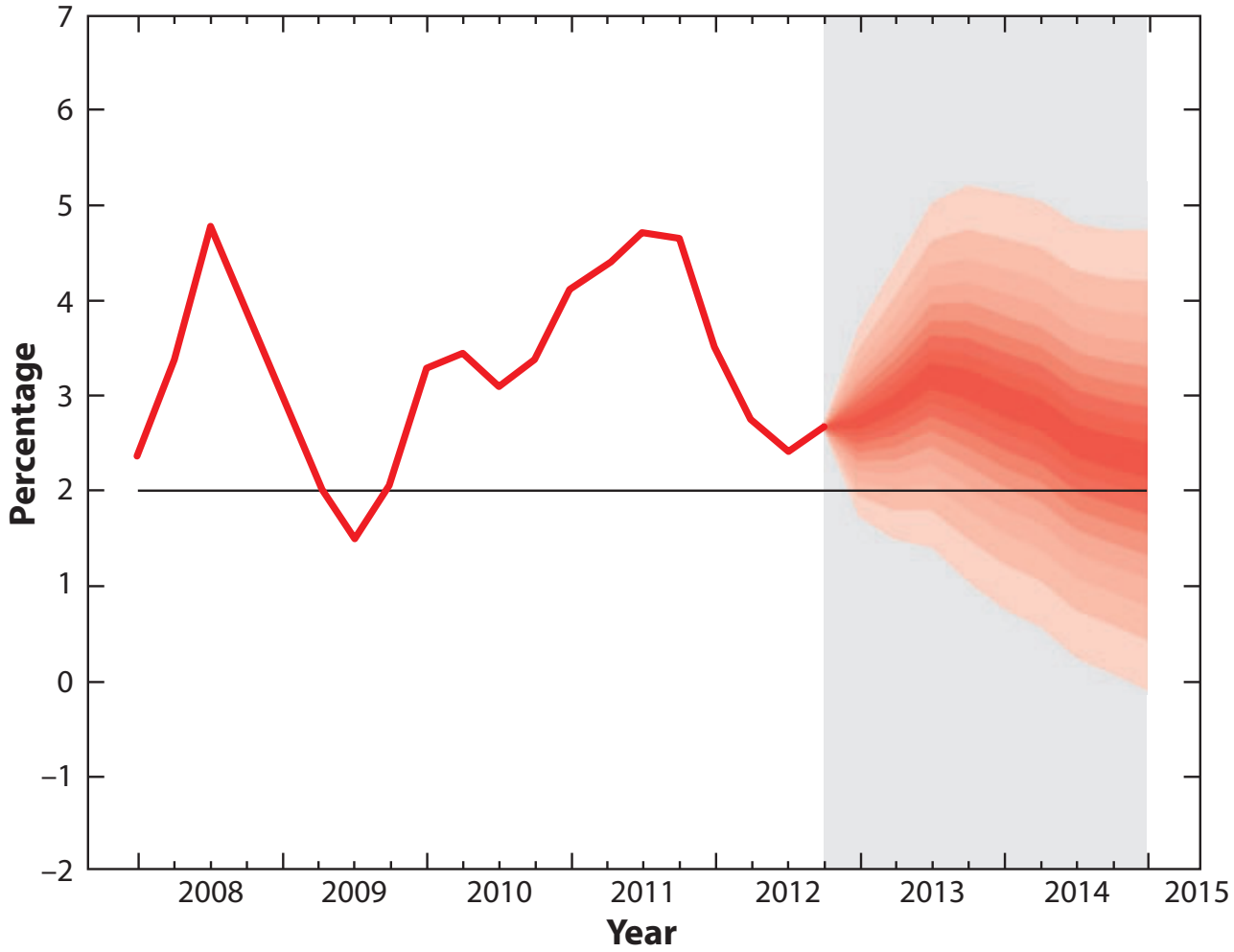


Figure 1

February 2013 Bank of England forecast of inflation in the United Kingdom as a percentage increase in the consumer price index (Bank of England 2013, with permission). The shaded bands in the fan chart show prediction intervals in increments of 10%.

Setup

- Y: Zu erklärende Variable:

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):
  - Sind aus einem teils hochdimensionalen feature space $\mathcal{X}$.

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):
  - Sind aus einem teils hochdimensionalen feature space $\mathcal{X}$.
  - Können metrisch sein, kategorial etc.

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):
  - Sind aus einem teils hochdimensionalen feature space $\mathcal{X}$.
  - Können metrisch sein, kategorial etc.
  - Können exogene Variablen sein (Querschnittsdaten) oder auch vergangene Beobachtungen von Y (Zeitreihenkontext).

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):
  - Sind aus einem teils hochdimensionalen feature space $\mathcal{X}$.
  - Können metrisch sein, kategorial etc.
  - Können exogene Variablen sein (Querschnittsdaten) oder auch vergangene Beobachtungen von Y (Zeitreihenkontext).

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):
  - Sind aus einem teils hochdimensionalen feature space $\mathcal{X}$.
  - Können metrisch sein, kategorial etc.
  - Können exogene Variablen sein (Querschnittsdaten) oder auch vergangene Beobachtungen von Y (Zeitreihenkontext).

Lernen Wir möchten die in $\boldsymbol{X}$ enthaltene Information nutzen, um Y so genau wie möglich zu beschreiben.
$\rightsquigarrow$ Welches Modell passt? Wie kann man es schätzen?

Setup

- Y: Zu erklärende Variable:
  - Typischerweise reellwertig, kann aber auch kategorial, multivariat, mengenwertig sein, ...
  - **Beispiele:** Temperatur; Niederschlag; Windgeschwindigkeit; Höhe oder Anzahl von Schadensfällen; Nachfrage nach einem Produkt; Gewinn eines Unternehmens; Wirtschaftswachstum; Inflation ...
- $\boldsymbol{X}$: Erklärende Variablen / Regressoren (Features):
  - Sind aus einem teils hochdimensionalen feature space $\mathcal{X}$.
  - Können metrisch sein, kategorial etc.
  - Können exogene Variablen sein (Querschnittsdaten) oder auch vergangene Beobachtungen von Y (Zeitreihenkontext).

Lernen Wir möchten die in $\boldsymbol{X}$ enthaltene Information nutzen, um Y so genau wie möglich zu beschreiben.
$\rightsquigarrow$ Welches Modell passt? Wie kann man es schätzen?

Vorhersagen Wir möchten die in $\boldsymbol{X}$ enthaltene Information nutzen, um ein uns unbekanntes Y so genau wie möglich vorherzusagen.
$\rightsquigarrow$ Wie kann man die Genauigkeit der Vorhersage messen und bewerten?

Was ist unser Ziel?

- Normalerweise lässt sich Y durch $\boldsymbol{X}$ nicht vollständig beschreiben: Es gibt keine deterministische Funktion g, so dass $Y = g(\boldsymbol{X})$ gilt.

Was ist unser Ziel?

- Normalerweise lässt sich Y durch $\boldsymbol{X}$ nicht vollständig beschreiben: Es gibt keine deterministische Funktion g, so dass $Y = g(\boldsymbol{X})$ gilt.

- Die verbleibende Unsicherheit von Y gegeben $\boldsymbol{X}$ lässt sich durch die bedingte Verteilung beschreiben:

$$F_{Y|\boldsymbol{x}}$$

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus
  - Quantile, Expektile

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus
  - Quantile, Expektile
  - Risikomasse: Value at Risk, Expected Shortfall

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus
  - Quantile, Expektile
  - Risikomasse: Value at Risk, Expected Shortfall
- Erzeugung einer Punktvorhersage $\widehat{T}(Y \mid \boldsymbol{X})$.

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus
  - Quantile, Expektile
  - Risikomasse: Value at Risk, Expected Shortfall
- Erzeugung einer Punktvorhersage $\widehat{T}(Y \mid \boldsymbol{X})$.
  - Informationsverlust

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus
  - Quantile, Expektile
  - Risikomasse: Value at Risk, Expected Shortfall
- Erzeugung einer Punktvorhersage $\widehat{T}(Y \mid \boldsymbol{X})$.
  - Informationsverlust
  - Implementierung einfacher

Was ist unser Ziel?

- Probabilistische Vorhersagen: Versuch, die vollständige bedingte Verteilung zu lernen und eine probabilistische Vorhersage $\widehat{F}_{Y|\boldsymbol{X}}$ zu erzeugen.
  - Sehr informativer Ansatz
  - Implementierung anspruchsvoll
  - Kommunikation kann schwierig sein – Ausnahme: binäre Ereignisse!
- Punktvorhersagen: Zusammenfassung der bedingten Verteilung durch ein Funktional der bedingten Verteilung

$$T(Y \mid \boldsymbol{X}) := T(F_{Y|\boldsymbol{X}})$$

- **Beispiele für T:**
  - Mittelwert, Median, Modus
  - Quantile, Expektile
  - Risikomasse: Value at Risk, Expected Shortfall
- Erzeugung einer Punktvorhersage $\widehat{T}(Y \mid \boldsymbol{X})$.
  - Informationsverlust
  - Implementierung einfacher
  - Kommunikation einfach

Was ist unser Ziel?

- Kompromiss: Punktvorhersage zusammen mit Quantifizierung der Unsicherheit (siehe Inflationsvorhersage).

Was ist unser Ziel?

- Kompromiss: Punktvorhersage zusammen mit Quantifizierung der Unsicherheit (siehe Inflationsvorhersage).
- **Beispiele:**
  - Mittelwert $\pm$ Standardabweichung
  - Prädiktionsinterval: Median zusammen mit kleinerem und grösserem Quantil (oder mehreren Quantilen).

Was ist unser Ziel?

- Kompromiss: Punktvorhersage zusammen mit Quantifizierung der Unsicherheit (siehe Inflationsvorhersage).
- **Beispiele:**
  - Mittelwert $\pm$ Standardabweichung
  - Prädiktionsinterval: Median zusammen mit kleinerem und grösserem Quantil (oder mehreren Quantilen).
- Weniger grobe Zusammenfassung der Verteilung. Kann beliebig verfeinert werden.

Was ist unser Ziel?

- Kompromiss: Punktvorhersage zusammen mit Quantifizierung der Unsicherheit (siehe Inflationsvorhersage).
- **Beispiele:**
  - Mittelwert $\pm$ Standardabweichung
  - Prädiktionsinterval: Median zusammen mit kleinerem und grösserem Quantil (oder mehreren Quantilen).
- Weniger grobe Zusammenfassung der Verteilung. Kann beliebig verfeinert werden.

Was ist unser Ziel?

- Kompromiss: Punktvorhersage zusammen mit Quantifizierung der Unsicherheit (siehe Inflationsvorhersage).
- **Beispiele:**
  - Mittelwert $\pm$ Standardabweichung
  - Prädiktionsinterval: Median zusammen mit kleinerem und grösserem Quantil (oder mehreren Quantilen).
- Weniger grobe Zusammenfassung der Verteilung. Kann beliebig verfeinert werden.

Weitere Notation:

- $\mathcal{A}$: Raum von Vorhersagen
- $\mathcal{F}$: Raum von möglichen bedingten Verteilungen $F_{Y|\boldsymbol{X}}$.

Wie werden Vorhersagen generiert?

Beispiel 1: Lineare Regression

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.

Beispiel 1: Lineare Regression

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung des bedingten Mittelwerts (Punktvorhersage):

$$\mathbb{E}\big[Y \big| \boldsymbol{X}\big] = \boldsymbol{X}\beta^*$$

Beispiel 1: Lineare Regression

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung des bedingten Mittelwerts (Punktvorhersage):

$$\mathbb{E}\big[Y \big| \boldsymbol{X}\big] = \boldsymbol{X}\beta^*$$

- **Schätzung:** Minimierung der durchschnittlichen quadrierten Residuen:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^{k+1}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \big(Y_i - \boldsymbol{X}_i\beta\big)^2$$

Beispiel 1: Lineare Regression

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung des bedingten Mittelwerts (Punktvorhersage):

$$\mathbb{E}\big[Y \big| \boldsymbol{X}\big] = \boldsymbol{X}\beta^*$$

- **Schätzung:** Minimierung der durchschnittlichen quadrierten Residuen:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^{k+1}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \big(Y_i - \boldsymbol{X}_i\beta\big)^2$$

- Funktioniert auch nicht-parametrisch (z.B. via isotone Regression).

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung der bedingten Dichte (probabilistische Vorhersage):

$$f_{Y|\boldsymbol{X}} = f_{\theta*}$$

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung der bedingten Dichte (probabilistische Vorhersage):

$$f_{Y|\boldsymbol{X}} = f_{\theta^*}$$

- **Schätzung:** Maximierung der durchschnittlichen log-Likelihood:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(f_\theta(Y_i \mid \boldsymbol{X}_i)\right)$$

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung der bedingten Dichte (probabilistische Vorhersage):

$$f_{Y|\boldsymbol{X}} = f_{\theta^*}$$

- **Schätzung:** Maximierung der durchschnittlichen log-Likelihood:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(f_\theta(Y_i \mid \boldsymbol{X}_i)\right)$$

- Funktioniert auch nicht-parametrisch.

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung der bedingten Dichte (probabilistische Vorhersage):

$$f_{Y|\boldsymbol{X}} = f_{\theta^*}$$

- **Schätzung:** Maximierung der durchschnittlichen log-Likelihood:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(f_\theta(Y_i \mid \boldsymbol{X}_i)\right)$$

- Funktioniert auch nicht-parametrisch.

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung der bedingten Dichte (probabilistische Vorhersage):

$$f_{Y|\boldsymbol{X}} = f_{\theta^*}$$

- **Schätzung:** Maximierung der durchschnittlichen log-Likelihood:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(f_\theta(Y_i \mid \boldsymbol{X}_i)\right)$$

- Funktioniert auch nicht-parametrisch.

Beide Lernverfahren beruhen auf der Minimierung / Maximierung von Verlustfunktionen (bis auf Vorzeichenkonvention).

Beispiel 2: Maximum Log-Likelihood Methode

- **Daten:** $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.
- **Ziel:** Schätzung der bedingten Dichte (probabilistische Vorhersage):

$$f_{Y|\boldsymbol{X}} = f_{\theta^*}$$

- **Schätzung:** Maximierung der durchschnittlichen log-Likelihood:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(f_\theta(Y_i \mid \boldsymbol{X}_i)\right)$$

- Funktioniert auch nicht-parametrisch.

Beide Lernverfahren beruhen auf der Minimierung / Maximierung von Verlustfunktionen (bis auf Vorzeichenkonvention).

$\rightsquigarrow$ Wann sollte man welche Verlustfunktion benutzen?

Konsistente Verlustfunktionen und Elizitierbarkeit

Definition 1 (Konsistenz)

Eine Verlustfunktion ist eine Abbildung

$$L\colon \mathcal{A} \times \mathbb{R} \to \mathbb{R}.$$

Manchmal fordert man noch Bedingungen wie Stetigkeit, Positivität etc.

Konsistente Verlustfunktionen und Elizitierbarkeit

Definition 1 (Konsistenz)

Eine Verlustfunktion ist eine Abbildung

$$L\colon \mathcal{A} \times \mathbb{R} \to \mathbb{R}.$$

Manchmal fordert man noch Bedingungen wie Stetigkeit, Positivität etc.
L ist $\mathcal{F}$-konsistent für ein Funktional T, falls

$$\mathrm{E}_{Y\sim F}\left[L(T(F), Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(a, Y)\right] \qquad \text{für alle } a \in \mathcal{A},\ F \in \mathcal{F}.$$

L ist strikt $\mathcal{F}$-konsistent, wenn Gleichheit nur bei $a = T(F)$ auftritt.

Konsistente Verlustfunktionen und Elizitierbarkeit

Definition 1 (Konsistenz)

Eine Verlustfunktion ist eine Abbildung

$$L\colon \mathcal{A} \times \mathbb{R} \to \mathbb{R}.$$

Manchmal fordert man noch Bedingungen wie Stetigkeit, Positivität etc.
L ist $\mathcal{F}$-konsistent für ein Funktional T, falls

$$\mathrm{E}_{Y\sim F}\left[L(T(F), Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(a, Y)\right] \qquad \text{für alle } a \in \mathcal{A},\ F \in \mathcal{F}.$$

L ist strikt $\mathcal{F}$-konsistent, wenn Gleichheit nur bei $a = T(F)$ auftritt.

Definition 2 (Elizitierbarkeit)

Ein Funktional T ist elizitierbar auf $\mathcal{F}$, falls es eine strikt $\mathcal{F}$-konsistente Verlustfunktion dafür gibt.

Erste Beispiele elizitierbare Funktionale

Der Mittelwert ist auf der Klasse der quadratintegrierbaren Verteilungen elizitierbar. Der quadratische Fehler ist eine mögliche strikt konsistente Verlustfunktion:

$$L(a, y) = (a - y)^2$$

Erste Beispiele elizitierbare Funktionale

Der Mittelwert ist auf der Klasse der quadratintegrierbaren Verteilungen elizitierbar. Der quadratische Fehler ist eine mögliche strikt konsistente Verlustfunktion:

$$L(a, y) = (a - y)^2$$

Der Median ist auf der Klasse der strikt wachsenden integrierbaren Verteilungen elizitierbar. Der absolute Fehler ist eine mögliche strikt konsistente Verlustfunktion:

$$L(a, y) = |a - y|$$

Erste Beispiele elizitierbare Funktionale

Der Mittelwert ist auf der Klasse der quadratintegrierbaren Verteilungen elizitierbar. Der quadratische Fehler ist eine mögliche strikt konsistente Verlustfunktion:

$$L(a, y) = (a - y)^2$$

Der Median ist auf der Klasse der strikt wachsenden integrierbaren Verteilungen elizitierbar. Der absolute Fehler ist eine mögliche strikt konsistente Verlustfunktion:

$$L(a, y) = |a - y|$$

Weitere Beispiele gibt es später.

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Wir betrachten das statistische Risiko eines Modells m:

$$\begin{aligned} R(m) &= \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= \mathrm{E}\Big[\mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right]\Big] \end{aligned}$$

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Wir betrachten das statistische Risiko eines Modells m:

$$\begin{aligned} R(m) &= \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= \mathrm{E}\Big[\mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right]\Big] \end{aligned}$$

- Bayes rule ist gegeben durch

$$m^* \in \operatorname*{arg\,min}_{m \in \mathcal{M}} R(m),$$

wobei $\mathcal{M}$ die Modellklasse ist.

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Wir betrachten das statistische Risiko eines Modells m:

$$\begin{aligned} R(m) &= \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= \mathrm{E}\Big[\mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right) \mid \boldsymbol{X}\Big] \end{aligned}$$

- Bayes rule ist gegeben durch

$$m^* \in \operatorname*{arg\,min}_{m \in \mathcal{M}} R(m),$$

  wobei $\mathcal{M}$ die Modellklasse ist.
- Falls die wahre Regressionsfunktion $\boldsymbol{x} \mapsto T(Y \mid \boldsymbol{X} = \boldsymbol{x})$ in $\mathcal{M}$ ist und falls L $\mathcal{F}$-konsistent für T ist, erhalten wir

$$\mathrm{E}\left[L\big(T(Y \mid \boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right] \leqslant \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right].$$

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Wir betrachten das statistische Risiko eines Modells m:

$$\begin{aligned} R(m) &= \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= \mathrm{E}\Big[\mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right) \mid \boldsymbol{X}\Big] \end{aligned}$$

- Bayes rule ist gegeben durch

$$m^* \in \operatorname*{arg\,min}_{m \in \mathcal{M}} R(m),$$

  wobei $\mathcal{M}$ die Modellklasse ist.
- Falls die wahre Regressionsfunktion $\boldsymbol{x} \mapsto T(Y \mid \boldsymbol{X} = \boldsymbol{x})$ in $\mathcal{M}$ ist und falls L $\mathcal{F}$-konsistent für T ist, erhalten wir

$$\mathrm{E}\left[L\big(T(Y \mid \boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right] \leqslant \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right].$$

- Daher ist $T(Y \mid \boldsymbol{X} = \cdot)$ eine Bayes rule.

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Wir betrachten das statistische Risiko eines Modells m:

$$\begin{aligned} R(m) &= \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= \mathrm{E}\Big[\mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right) \mid \boldsymbol{X}\Big] \end{aligned}$$

- Bayes rule ist gegeben durch

$$m^* \in \operatorname*{arg\,min}_{m \in \mathcal{M}} R(m),$$

  wobei $\mathcal{M}$ die Modellklasse ist.
- Falls die wahre Regressionsfunktion $\boldsymbol{x} \mapsto T(Y \mid \boldsymbol{X} = \boldsymbol{x})$ in $\mathcal{M}$ ist und falls L $\mathcal{F}$-konsistent für T ist, erhalten wir

$$\mathrm{E}\left[L\big(T(Y \mid \boldsymbol{X}), Y\big) \mid \boldsymbol{X}\right] \leqslant \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right) \mid \boldsymbol{X}\big].$$

- Daher ist $T(Y \mid \boldsymbol{X} = \cdot)$ eine Bayes rule.
- Man kann zeigen, dass die (strikte) Konstistenz von L auch notwendig ist.

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Es sei $D_{\text{train}} = \{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n\}$ ein Trainings-Sample. Wir definieren das empirische Risiko von m als

$$\begin{aligned} \overline{R}(m; D_{\text{train}}) &= \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D_{\text{train}}} L(m(\boldsymbol{x}_i), y_i) \\ &\approx \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= R(m). \end{aligned}$$

Lernen durch Minimierung von Verlustfunktionen (M-estimation)

- Es sei $D_{\text{train}} = \{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n\}$ ein Trainings-Sample. Wir definieren das empirische Risiko von m als

$$\begin{aligned}\overline{R}(m; D_{\text{train}}) &= \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D_{\text{train}}} L(m(\boldsymbol{x}_i), y_i) \\ &\approx \mathrm{E}\left[L\big(m(\boldsymbol{X}), Y\big)\right] \\ &= R(m).\end{aligned}$$

- M-estimator $\widehat{m}$ minimiert das empirische Risiko

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \overline{R}(m; D_{\text{train}})$$

Gefahr von Overfitting

- Der Schätzer $\hat{m}$ hängt vom Traings-Sample D_{train} ab:

Gefahr von Overfitting

- Der Schätzer $\hat{m}$ hängt vom Traings-Sample D_{train} ab:
  - Anfällig für Schätzfehler

Gefahr von Overfitting

- Der Schätzer $\hat{m}$ hängt vom Traings-Sample D_{train} ab:
  - Anfällig für Schätzfehler
  - Unterschiedliche Trainigs-Samples führen zu unterschiedlichen Schätzungen.

Gefahr von Overfitting

- Der Schätzer $\widehat{m}$ hängt vom Traings-Sample D_{train} ab:
  - Anfällig für Schätzfehler
  - Unterschiedliche Trainigs-Samples führen zu unterschiedlichen Schätzungen.
  - Gefahr, dass $\widehat{m}$ das Rauschen statt des Signals eines Samples lernt.

Gefahr von Overfitting

- Der Schätzer $\hat{m}$ hängt vom Traings-Sample D_{train} ab:
  - Anfällig für Schätzfehler
  - Unterschiedliche Trainigs-Samples führen zu unterschiedlichen Schätzungen.
  - Gefahr, dass $\hat{m}$ das Rauschen statt des Signals eines Samples lernt.
  - In-sample Performance $\overline{R}(\hat{m}; D_{\text{train}})$ kann ein schlechter Schätzer für das tatsächliche Risiko $R(\hat{m})$ sein.

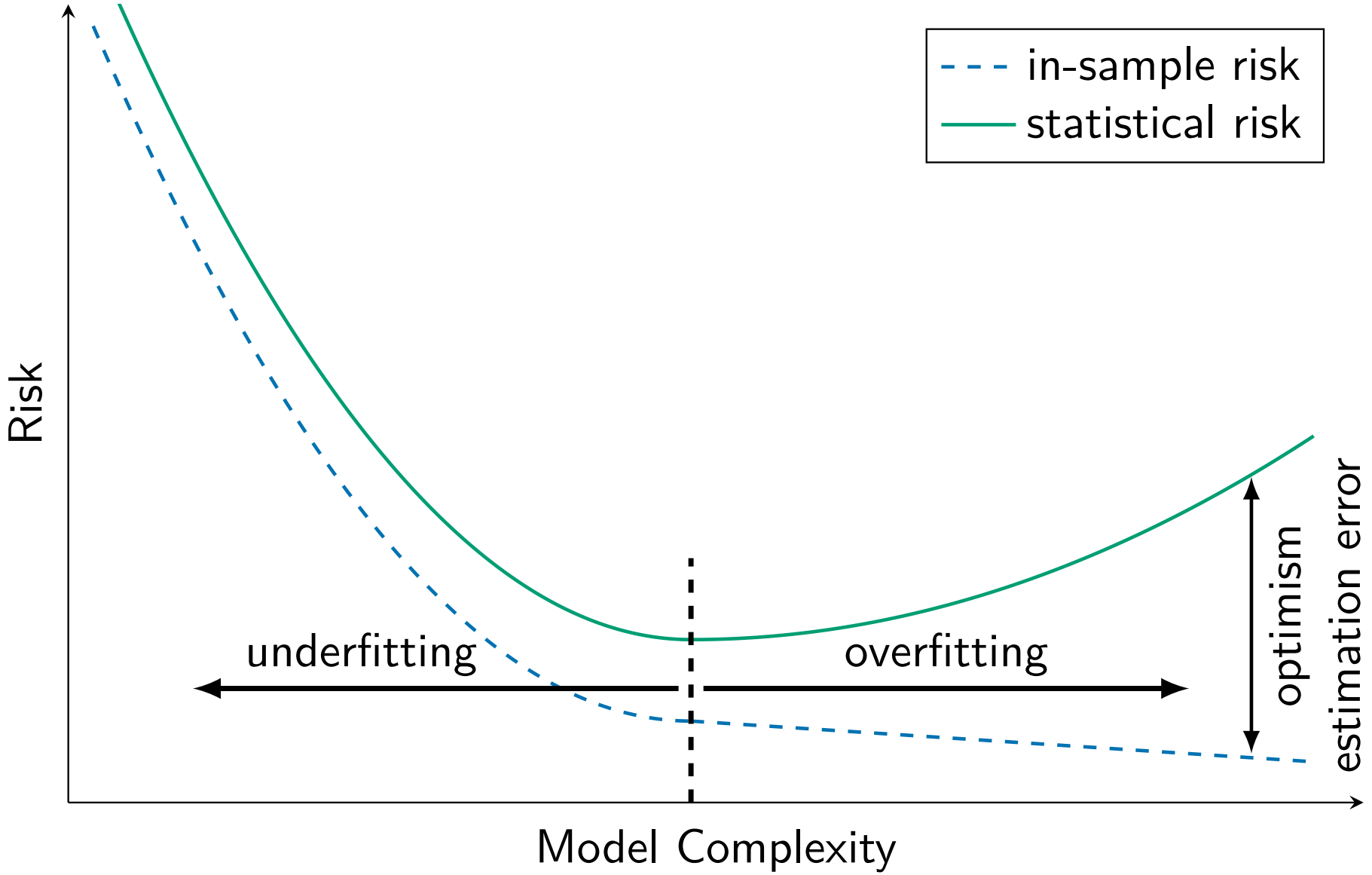
Gefahr von Overfitting

- Der Schätzer $\widehat{m}$ hängt vom Traings-Sample D_{train} ab:
  - Anfällig für Schätzfehler
  - Unterschiedliche Trainigs-Samples führen zu unterschiedlichen Schätzungen.
  - Gefahr, dass $\widehat{m}$ das Rauschen statt des Signals eines Samples lernt.
  - In-sample Performance $\overline{R}(\widehat{m}; D_{\text{train}})$ kann ein schlechter Schätzer für das tatsächliche Risiko $R(\widehat{m})$ sein.
  - $\rightsquigarrow$ Gefahr von Overfitting.

Gefahr von Overfitting

- Der Schätzer $\hat{m}$ hängt vom Traings-Sample D_{train} ab:
  - Anfällig für Schätzfehler
  - Unterschiedliche Trainigs-Samples führen zu unterschiedlichen Schätzungen.
  - Gefahr, dass $\hat{m}$ das Rauschen statt des Signals eines Samples lernt.
  - In-sample Performance $\overline{R}(\hat{m}; D_{\text{train}})$ kann ein schlechter Schätzer für das tatsächliche Risiko $R(\hat{m})$ sein.
  - $\rightsquigarrow$ Gefahr von Overfitting.
  - Das Problem wird akuter ...
    - ★ ...je komplexer das Modell ist;
    - ★ ...je kleiner (weniger repräsentativ) das Trainings-Sample ist.

Gefahr von Overfitting



Strategien gegen Overfitting

Strategien gegen Overfitting

1. **Fitting:** Einführung eines Strafterms Ω, der die Modellkomplexität misst:

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \overline{R}(m; D_{\text{train}}) + \lambda \Omega(m).$$

   Beispiele für Ω:

   - Anzahl Parameter $\rightsquigarrow$ AIC und BIC
   - Norm der Parameter $\rightsquigarrow$ Ridge- und Lasso-Regression
   - Anzahl Optimierungsschritte bei der Schätzung eines neuronalen Netzwerks

Strategien gegen Overfitting

1. **Fitting:** Einführung eines Strafterms Ω, der die Modellkomplexität misst:

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \overline{R}(m; D_{\text{train}}) + \lambda \Omega(m).$$

   Beispiele für Ω:
   - Anzahl Parameter $\rightsquigarrow$ AIC und BIC
   - Norm der Parameter $\rightsquigarrow$ Ridge- und Lasso-Regression
   - Anzahl Optimierungsschritte bei der Schätzung eines neuronalen Netzwerks

2. **Validierung:** Schätzung des out-of-sample Risikos auf einem (idealerweise) unabhängigen und identisch verteilten Validierungs-Sample $D_{\text{valid}} = \{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, l\}$ via

$$\overline{R}(\hat{m}; D_{\text{valid}})$$

   - Bessere Nährung des statistischen Risikos.
   - Kann mittels Kreuz-Validierung effizienter gemacht werden.

Was macht gute Vorhersagen aus?

Was scheint euch wichtig?

Welche Begriffe kommen euch in den Sinn?

Modellunabhängiger Prognosevergleich

- Wir haben verschiedene Methoden zur Erstellung von Vorhersagen, aber wir sind agnostisch hinsichtlich der Art und Weise, wie sie erstellt wurden.

Modellunabhängiger Prognosevergleich

- Wir haben verschiedene Methoden zur Erstellung von Vorhersagen, aber wir sind agnostisch hinsichtlich der Art und Weise, wie sie erstellt wurden.
- Beispiel zweier Prognosen: Prognose-Beobachtungsfolge

$$\left(A_i^{(1)}, A_i^{(2)}, Y_i\right) \quad i = 1, \ldots, N$$

Modellunabhängiger Prognosevergleich

- Wir haben verschiedene Methoden zur Erstellung von Vorhersagen, aber wir sind agnostisch hinsichtlich der Art und Weise, wie sie erstellt wurden.
- Beispiel zweier Prognosen: Prognose-Beobachtungsfolge

$$\left(A_i^{(1)}, A_i^{(2)}, Y_i\right) \quad i = 1, \ldots, N$$

- Rangfolge hinsichtlich der empirischen Verlustdifferenz:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(1)}, Y_i\right) \overset{?}{\gtrless} \frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(2)}, Y_i\right)$$

Modellunabhängiger Prognosevergleich

- Wir haben verschiedene Methoden zur Erstellung von Vorhersagen, aber wir sind agnostisch hinsichtlich der Art und Weise, wie sie erstellt wurden.
- Beispiel zweier Prognosen: Prognose-Beobachtungsfolge

$$\left(A_i^{(1)}, A_i^{(2)}, Y_i\right) \quad i = 1, \ldots, N$$

- Rangfolge hinsichtlich der empirischen Verlustdifferenz:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(1)}, Y_i\right) \overset{?}{\gtrless} \frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(2)}, Y_i\right)$$

  - Die Prognosemethode 1 wird als besser als 2 angesehen, wenn die linke Seite kleiner ist als die rechte.

Modellunabhängiger Prognosevergleich

- Wir haben verschiedene Methoden zur Erstellung von Vorhersagen, aber wir sind agnostisch hinsichtlich der Art und Weise, wie sie erstellt wurden.
- Beispiel zweier Prognosen: Prognose-Beobachtungsfolge

$$\left(A_i^{(1)}, A_i^{(2)}, Y_i\right) \quad i = 1, \ldots, N$$

- Rangfolge hinsichtlich der empirischen Verlustdifferenz:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(1)}, Y_i\right) \overset{?}{\gtrless} \frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(2)}, Y_i\right)$$

  - Die Prognosemethode 1 wird als besser als 2 angesehen, wenn die linke Seite kleiner ist als die rechte.
  - Tests auf gleiche Vorhersagegenauigkeit $\mathrm{E}[L(A^{(1)}, Y)] = \mathrm{E}[L(A^{(2)}, Y)]$ und bessere Vorhersagegenauigkeit $\mathrm{E}[L(A^{(1)}, Y)] \geqslant \mathrm{E}[L(A^{(2)}, Y)]$ können mittels Diebold–Mariano Tests durchgeführt werden (ähnlich zu t-Tests).

Modellunabhängiger Prognosevergleich

- Wir haben verschiedene Methoden zur Erstellung von Vorhersagen, aber wir sind agnostisch hinsichtlich der Art und Weise, wie sie erstellt wurden.
- Beispiel zweier Prognosen: Prognose-Beobachtungsfolge

$$\left(A_i^{(1)}, A_i^{(2)}, Y_i\right) \quad i = 1, \ldots, N$$

- Rangfolge hinsichtlich der empirischen Verlustdifferenz:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(1)}, Y_i\right) \overset{?}{\gtrless} \frac{1}{n}\sum_{i=1}^{n} L\left(A_i^{(2)}, Y_i\right)$$

  - Die Prognosemethode 1 wird als besser als 2 angesehen, wenn die linke Seite kleiner ist als die rechte.
  - Tests auf gleiche Vorhersagegenauigkeit $\mathrm{E}[L(A^{(1)}, Y)] = \mathrm{E}[L(A^{(2)}, Y)]$ und bessere Vorhersagegenauigkeit $\mathrm{E}[L(A^{(1)}, Y)] \geqslant \mathrm{E}[L(A^{(2)}, Y)]$ können mittels Diebold–Mariano Tests durchgeführt werden (ähnlich zu t-Tests).
- Welche Verlustfunktion sollte man nehmen?

Modellunabhängiger Prognosevergleich

Auswertung von Punktvorhersagen

- Klärung, was das Ziel der Vorhersage ist! An welchem Funktional T ist man interessiert?

Modellunabhängiger Prognosevergleich

Auswertung von Punktvorhersagen

- Klärung, was das Ziel der Vorhersage ist! An welchem Funktional T ist man interessiert?
- Strikt $\mathcal{F}$-konsistent Verlustfunktion für T wählen:

$$\mathrm{E}_{Y\sim F}\left[L(T(F), Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(a, Y)\right] \qquad \text{für alle } a \in \mathcal{A},\ F \in \mathcal{F}.$$

Modellunabhängiger Prognosevergleich

Auswertung von Punktvorhersagen

- Klärung, was das Ziel der Vorhersage ist! An welchem Funktional T ist man interessiert?
- Strikt $\mathcal{F}$-konsistent Verlustfunktion für T wählen:

$$\mathrm{E}_{Y\sim F}\left[L(T(F), Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(a, Y)\right] \qquad \text{für alle } a \in \mathcal{A},\ F \in \mathcal{F}.$$

Modellunabhängiger Prognosevergleich

Auswertung von Punktvorhersagen

- Klärung, was das Ziel der Vorhersage ist! An welchem Funktional T ist man interessiert?
- Strikt $\mathcal{F}$-konsistent Verlustfunktion für T wählen:

$$\mathrm{E}_{Y\sim F}\left[L(T(F), Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(a, Y)\right] \qquad \text{für alle } a \in \mathcal{A},\ F \in \mathcal{F}.$$

Auswertung von probabilistischen Vorhersagen

- Idee: Funktional T entspricht der Identitätsabbildung; $\mathcal{A} = \mathcal{F}$

Modellunabhängiger Prognosevergleich

Auswertung von Punktvorhersagen

- Klärung, was das Ziel der Vorhersage ist! An welchem Funktional T ist man interessiert?
- Strikt $\mathcal{F}$-konsistent Verlustfunktion für T wählen:

$$\mathrm{E}_{Y\sim F}\left[L(T(F), Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(a, Y)\right] \qquad \text{für alle } a \in \mathcal{A},\ F \in \mathcal{F}.$$

Auswertung von probabilistischen Vorhersagen

- Idee: Funktional T entspricht der Identitätsabbildung; $\mathcal{A} = \mathcal{F}$
- Strikte Konsistenz für probabilistische Vorhersagen nimmt somit folgende Form an:

$$\mathrm{E}_{Y\sim F}\left[L(F, Y)\right] \leqslant \mathrm{E}_{Y\sim F}\left[L(G, Y)\right] \qquad \text{für alle } F, G \in \mathcal{F}.$$

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

T	$L(a, y)$
Mittelwert	$(a-y)^2$
Median	$\lvert a-y\rvert$
τ-Expektil	$\lvert\mathbb{1}\{y \leqslant x\} - \tau\rvert(a-y)^2$
α-Quantil	$\lvert\mathbb{1}\{y \leqslant x\} - \alpha\rvert\,\lvert a-y\rvert$

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

T	$L(a, y)$
Mittelwert	$(a-y)^2$
Median	$\lvert a-y\rvert$
τ-Expektil	$\lvert\mathbb{1}\{y \leqslant x\} - \tau\rvert(a-y)^2$
α-Quantil	$\lvert\mathbb{1}\{y \leqslant x\} - \alpha\rvert\,\lvert a-y\rvert$
Modus	$1\{a \neq y\}$, ×

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

T	$L(a, y)$
Mittelwert	$(a-y)^2$
Median	$\lvert a-y\rvert$
τ-Expektil	$\lvert\mathbb{1}\{y \leqslant x\} - \tau\rvert(a-y)^2$
α-Quantil	$\lvert\mathbb{1}\{y \leqslant x\} - \alpha\rvert\,\lvert a-y\rvert$
Modus	$1\{a \neq y\}$, ×
Varianz	×

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

T	$L(a, y)$
Mittelwert	$(a-y)^2$
Median	$\lvert a-y\rvert$
τ-Expektil	$\lvert\mathbb{1}\{y \leqslant x\} - \tau\rvert(a-y)^2$
α-Quantil	$\lvert\mathbb{1}\{y \leqslant x\} - \alpha\rvert\,\lvert a-y\rvert$
Modus	$1\{a \neq y\}$, ×
Varianz	×
(Mittelwert, Varianz)	✓

Das Elicitation Problem

Es sei $T\colon \mathcal{F} \to \mathcal{A}$ ein Funktional.

(a) Ist T elizitierbar?

(b) Wie sieht die Klasse der (strikt) konsistenten Verlustfunktionen für T aus?

(c) Welche Verlustfunktion ist besonders geschickt?

(d) Was kann man tun, falls T nicht elizitierbar ist?

T	$L(a, y)$
Mittelwert	$(a-y)^2$
Median	$\lvert a-y\rvert$
τ-Expektil	$\lvert\mathbb{1}\{y \leqslant x\} - \tau\rvert(a-y)^2$
α-Quantil	$\lvert\mathbb{1}\{y \leqslant x\} - \alpha\rvert\,\lvert a-y\rvert$
Modus	$1\{a \neq y\}$, ×
Varianz	×
(Mittelwert, Varianz)	✓
Identität (probabilistische Prognose)	$L(F, y) = -\log(f(y))$

Was macht gute Vorhersagen aus?

Zerlegung von Verlustfunktionen

$$
\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} \quad \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} \quad \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] \quad \text{(uncertainty / entropy)}.
\end{aligned}
$$

Zerlegung von Verlustfunktionen

$$
\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] && \text{(uncertainty / entropy).}
\end{aligned}
$$

Miscalibration **Systematischer Fehler** zwischen Prognose A und dem, was man idealerweise mit der Information in $\boldsymbol{X}$ hätte erreichen können mittels Prognose $T(Y \mid \boldsymbol{X})$.

Zerlegung von Verlustfunktionen

$$
\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} \quad \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} \quad \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] \quad \text{(uncertainty / entropy)}.
\end{aligned}
$$

Miscalibration **Systematischer Fehler** zwischen Prognose A und dem, was man idealerweise mit der Information in $\boldsymbol{X}$ hätte erreichen können mittels Prognose $T(Y \mid \boldsymbol{X})$.

Discrimination Potentielle **Verminderung der Unsicherheit über Informationsgewinn**, von idealer uninformierter Prognose $T(Y)$ zu idealer und informierter Prognose $T(Y)$.

Zerlegung von Verlustfunktionen

$$
\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] && \text{(uncertainty / entropy).}
\end{aligned}
$$

Miscalibration **Systematischer Fehler** zwischen Prognose A und dem, was man idealerweise mit der Information in $\boldsymbol{X}$ hätte erreichen können mittels Prognose $T(Y \mid \boldsymbol{X})$.

Discrimination Potentielle **Verminderung der Unsicherheit über Informationsgewinn**, von idealer uninformierter Prognose $T(Y)$ zu idealer und informierter Prognose $T(Y)$.

Uncertainty **Inhärente Unsicherheit** von Y, durch L ausgedrückt. Hängt nur von der Verteilung von Y ab, nicht von der Prognose.

Zerlegung von Verlustfunktionen

$$
\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] && \text{(uncertainty / entropy).}
\end{aligned}
$$

Miscalibration **Systematischer Fehler** zwischen Prognose A und dem, was man idealerweise mit der Information in $\boldsymbol{X}$ hätte erreichen können mittels Prognose $T(Y \mid \boldsymbol{X})$.

Discrimination Potentielle **Verminderung der Unsicherheit über Informationsgewinn**, von idealer uninformierter Prognose $T(Y)$ zu idealer und informierter Prognose $T(Y)$.

Uncertainty **Inhärente Unsicherheit** von Y, durch L ausgedrückt. Hängt nur von der Verteilung von Y ab, nicht von der Prognose.

Zerlegung von Verlustfunktionen

$$
\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} && \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] && \text{(uncertainty / entropy).}
\end{aligned}
$$

Miscalibration **Systematischer Fehler** zwischen Prognose A und dem, was man idealerweise mit der Information in $\boldsymbol{X}$ hätte erreichen können mittels Prognose $T(Y \mid \boldsymbol{X})$.

Discrimination Potentielle **Verminderung der Unsicherheit über Informationsgewinn**, von idealer uninformierter Prognose $T(Y)$ zu idealer und informierter Prognose $T(Y)$.

Uncertainty **Inhärente Unsicherheit** von Y, durch L ausgedrückt. Hängt nur von der Verteilung von Y ab, nicht von der Prognose.

Minimierung einer erwarteten Verlustfunktion ist äquivalent dazu, gleichzeitig die systematischen Fehler zu minimieren und die Information zu maximieren.

Beispiel: Zerlegung des quadratischen Fehlers

$$\begin{aligned}
\mathrm{E}[L(A, Y)] = &\Big\{ \mathrm{E}[L(A, Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} \quad \geqslant 0 \quad \text{(miscalibration)} \\
&- \Big\{ \mathrm{E}[L(T(Y), Y)] - \mathrm{E}[L(T(Y \mid \boldsymbol{X}), Y)] \Big\} \quad \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{E}[L(T(Y), Y)] \quad \text{(uncertainty / entropy)}.
\end{aligned}$$

$L(A, Y) = (A - Y)^2$:

$$\begin{aligned}
\mathbb{E}\big[(A - Y)^2\big] = &\ \mathbb{E}\big[(A - \mathbb{E}[Y|\boldsymbol{X}])^2\big] \quad \geqslant 0 \quad \text{(miscalibration)} \\
&- \mathrm{Var}\big[\mathbb{E}[Y|\boldsymbol{X}]\big] \quad \geqslant 0 \quad \text{(discrimination)} \\
&+ \mathrm{Var}[Y] \quad \text{(uncertainty / entropy)}.
\end{aligned}$$

Ideen für den Unterricht

Ideen für den Unterricht

Charakterisierung von statistischen Lagemassen (Mittelwert, Median und Modus, evtl. auch von Quantilen) als Lösungen von unterschiedlichen Minimierungsproblemen.

- Definition nur für empirische Grössen in einem Sample.

Ideen für den Unterricht

Charakterisierung von statistischen Lagemassen (Mittelwert, Median und Modus, evtl. auch von Quantilen) als Lösungen von unterschiedlichen Minimierungsproblemen.

- Definition nur für empirische Grössen in einem Sample.
- Mittelwert:

$$\arg\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(a-y_i)^2$$

Ideen für den Unterricht

Charakterisierung von statistischen Lagemassen (Mittelwert, Median und Modus, evtl. auch von Quantilen) als Lösungen von unterschiedlichen Minimierungsproblemen.

- Definition nur für empirische Grössen in einem Sample.
- Mittelwert:

$$\arg\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(a-y_i)^2$$

- Median:

$$\arg\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}|a-y_i|$$

Ideen für den Unterricht

Charakterisierung von statistischen Lagemassen (Mittelwert, Median und Modus, evtl. auch von Quantilen) als Lösungen von unterschiedlichen Minimierungsproblemen.

- Definition nur für empirische Grössen in einem Sample.
- Mittelwert:

$$\arg\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(a-y_i)^2$$

- Median:

$$\arg\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}|a-y_i|$$

- Modus (bei diskreten Daten):

$$\arg\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}1\{a\neq y\} = \arg\max_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}1\{a= y\}$$

Ideen für den Unterricht

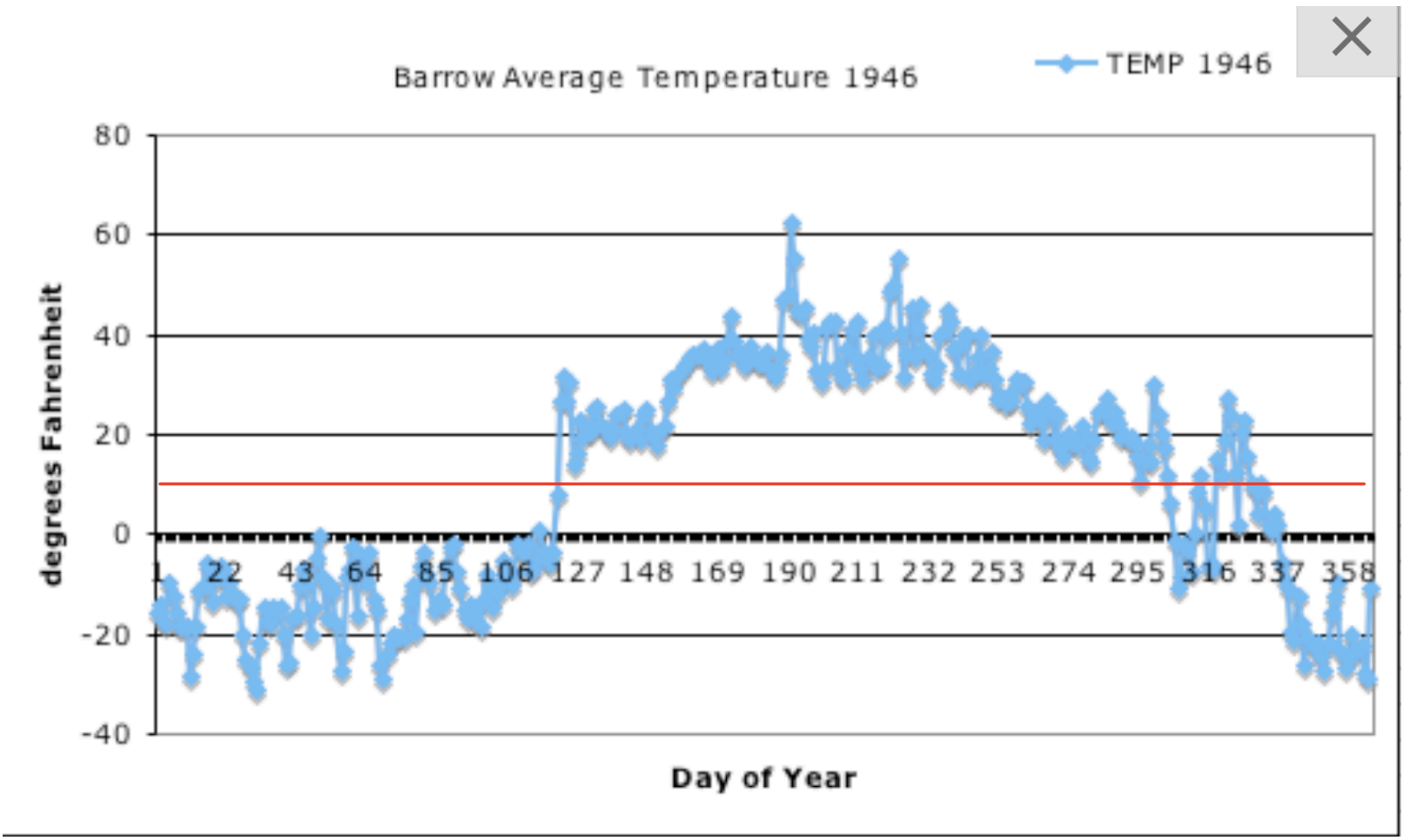
Diskussion über Modelle für Temperaturvorhersagen

- Was ist wichtig für eine gute Vorhersage?
  $\rightsquigarrow$ möglichst viel Information. Information möglichst gut verwenden.

Ideen für den Unterricht

Diskussion über Modelle für Temperaturvorhersagen

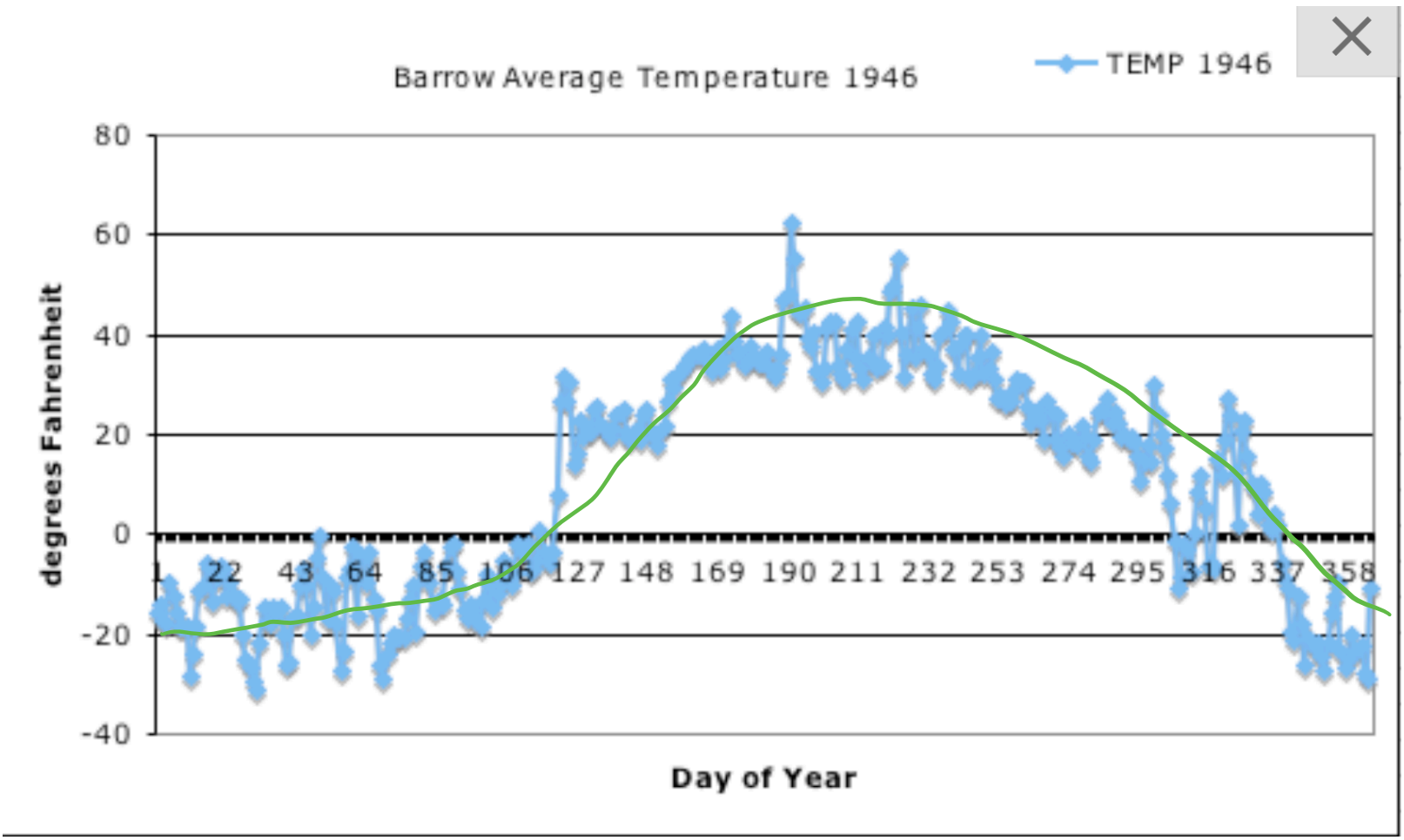
- Was ist wichtig für eine gute Vorhersage?
  $\leadsto$ möglichst viel Information. Information möglichst gut verwenden.
- Konstante Vorhersage der Jahresdurchschnittstemperatur


Barrow Average Temperature 1946
TEMP 1946
80
60
40
20
0
-20
-40
degrees Fahrenheit
1
22
43
64
85
106
127
148
169
190
211
232
253
274
295
316
337
358
Day of Year

Ideen für den Unterricht

Diskussion über Modelle für Temperaturvorhersagen

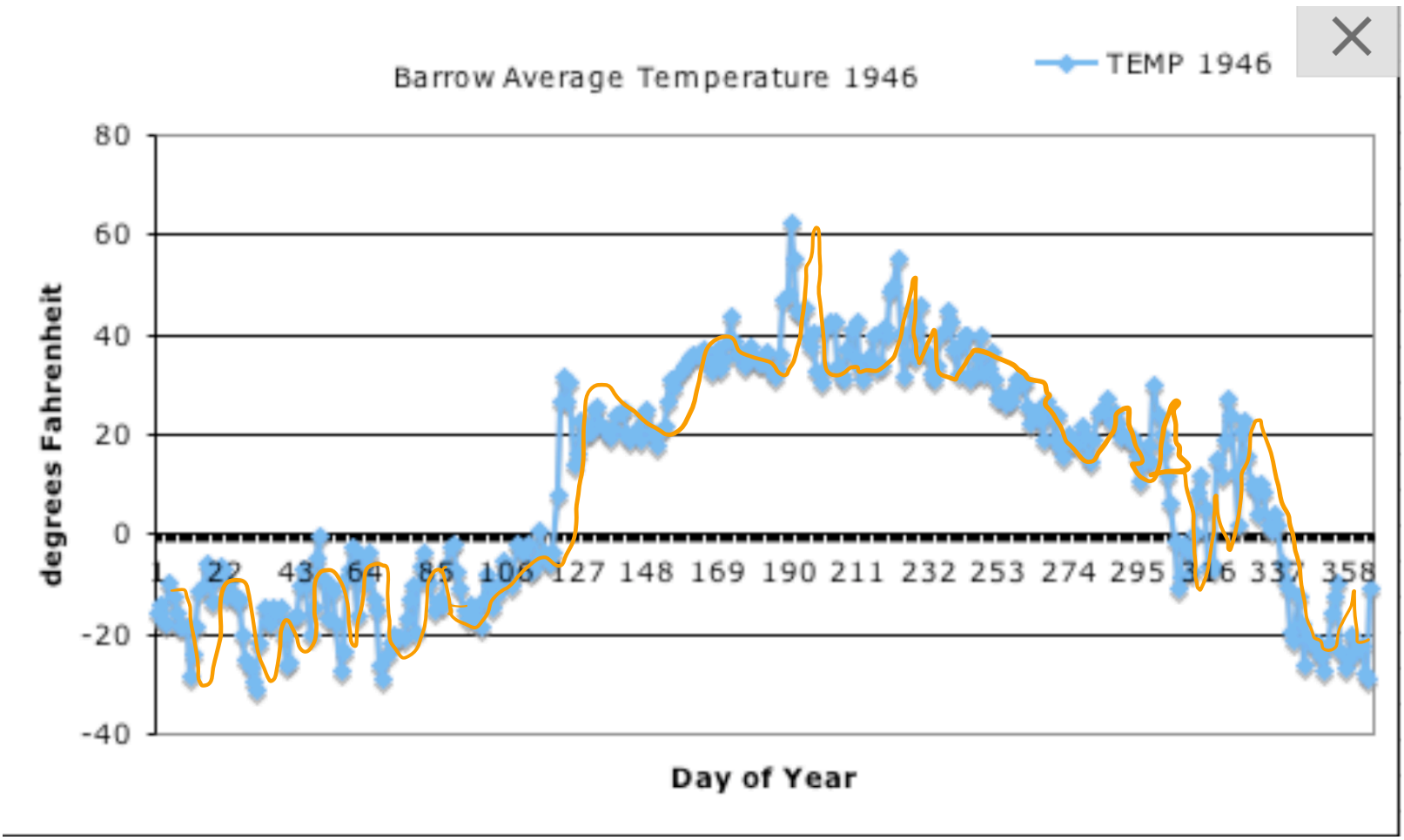
- Was ist wichtig für eine gute Vorhersage?
  $\rightsquigarrow$ möglichst viel Information. Information möglichst gut verwenden.
- Konstante Vorhersage der Jahresdurchschnittstemperatur
- Einbeziehen von Saisonalitäten


Barrow Average Temperature 1946
TEMP 1946
80
60
40
20
0
-20
-40
degrees Fahrenheit
1
22
43
64
85
106
127
148
169
190
211
232
253
274
295
316
337
358
Day of Year

Ideen für den Unterricht

Diskussion über Modelle für Temperaturvorhersagen

- Was ist wichtig für eine gute Vorhersage?
  $\leadsto$ möglichst viel Information. Information möglichst gut verwenden.
- Konstante Vorhersage der Jahresdurchschnittstemperatur
- Einbeziehen von Saisonalitäten
- Relevante Informationen: Immer die Temperatur von gestern nehmen?


Barrow Average Temperature 1946
TEMP 1946
80
60
40
20
0
-20
-40
degrees Fahrenheit
1
22
43
64
85
106
127
148
169
190
211
232
253
274
295
316
337
358
Day of Year

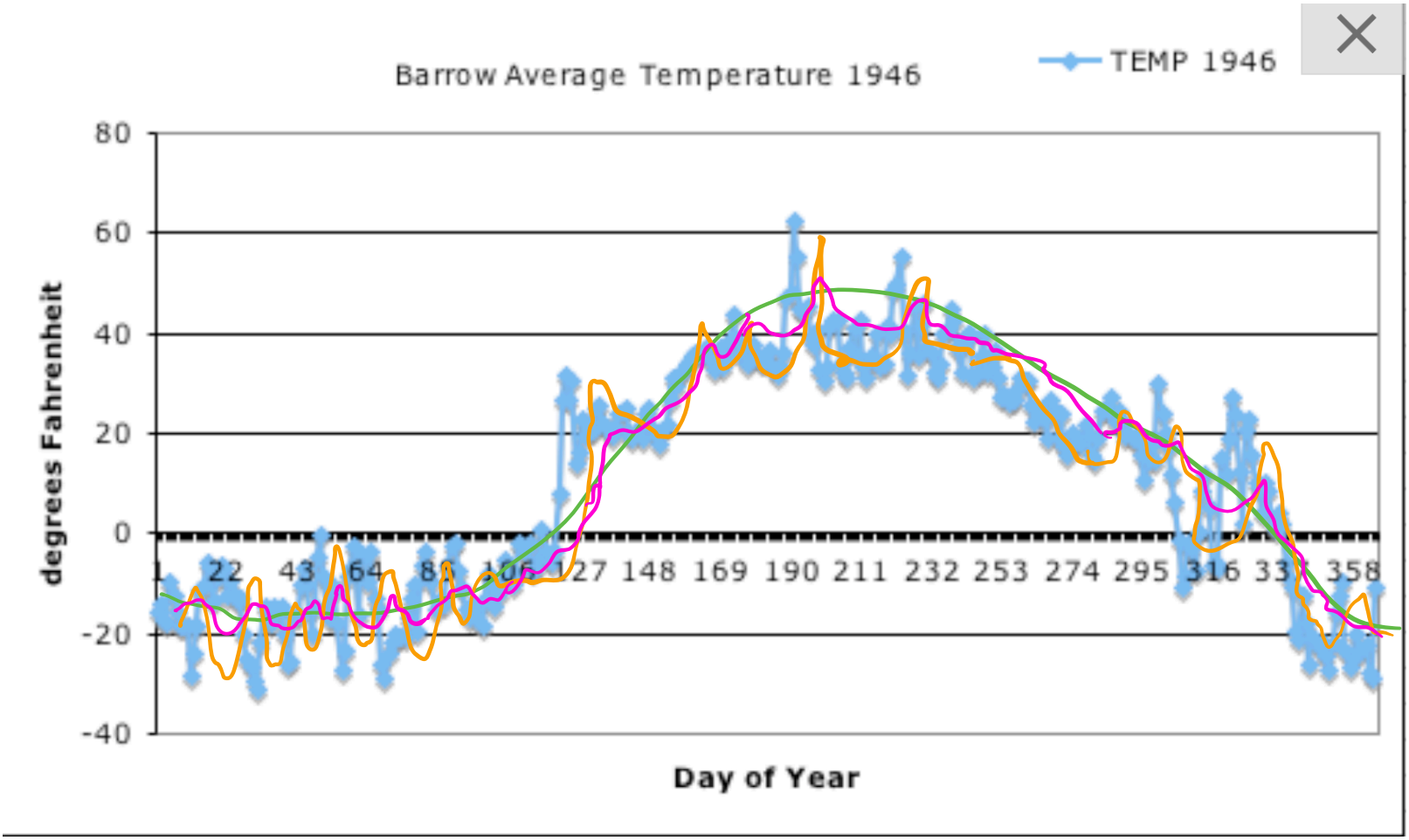
Ideen für den Unterricht

Diskussion über Modelle für Temperaturvorhersagen

- Was ist wichtig für eine gute Vorhersage?
  $\rightsquigarrow$ möglichst viel Information. Information möglichst gut verwenden.
- Konstante Vorhersage der Jahresdurchschnittstemperatur
- Einbeziehen von Saisonalitäten
- Relevante Informationen: Immer die Temperatur von gestern nehmen?
- AR(1) Modell:

$$Y_t = \theta Y_{t-1} + \varepsilon_t, \qquad 0 < \theta < 1$$

$\rightsquigarrow$ Information von gestern wird mit einbezogen, aber es gibt eine längerfristige Rückkehr zum Trend.


Barrow Average Temperature 1946
TEMP 1946
80
60
40
20
0
-20
-40
degrees Fahrenheit
1
22
43
64
85
106
127
148
169
190
211
232
253
274
295
316
337
358
Day of Year

Ideen für den Unterricht

Diskussion über Modelle für Temperaturvorhersagen

- Was ist wichtig für eine gute Vorhersage?
  $\rightsquigarrow$ möglichst viel Information. Information möglichst gut verwenden.
- Konstante Vorhersage der Jahresdurchschnittstemperatur
- Einbeziehen von Saisonalitäten
- Relevante Informationen: Immer die Temperatur von gestern nehmen?
- AR(1) Modell:

$$Y_t = \theta Y_{t-1} + \varepsilon_t, \qquad 0 < \theta < 1$$

  $\rightsquigarrow$ Information von gestern wird mit einbezogen, aber es gibt eine längerfristige Rückkehr zum Trend.
- Auswerten mittels Verlustfunktion, zum Beispiel dem quadratischen Fehler.

Zusammenfassung

- Verlustfunktionen spielen eine wichtige Rolle beim Lernen und bei der Auswertung von Prognosen.

Zusammenfassung

- Verlustfunktionen spielen eine wichtige Rolle beim Lernen und bei der Auswertung von Prognosen.
- Sie sollten immer zum relevanten Funktional passen.
  $\rightsquigarrow$ Konsistenz

Zusammenfassung

- Verlustfunktionen spielen eine wichtige Rolle beim Lernen und bei der Auswertung von Prognosen.
- Sie sollten immer zum relevanten Funktional passen.
  $\rightsquigarrow$ Konsistenz
- Strikte Konsistenz stellt sicher, dass die wahre Regressionsfunktion auch erlernt wird.

Zusammenfassung

- Verlustfunktionen spielen eine wichtige Rolle beim Lernen und bei der Auswertung von Prognosen.
- Sie sollten immer zum relevanten Funktional passen.
  $\rightsquigarrow$ Konsistenz
- Strikte Konsistenz stellt sicher, dass die wahre Regressionsfunktion auch erlernt wird.
- Strikte Konsistenz stellt anreizkompatible Vergleiche von Prognosen sicher.

Zusammenfassung

- Verlustfunktionen spielen eine wichtige Rolle beim Lernen und bei der Auswertung von Prognosen.
- Sie sollten immer zum relevanten Funktional passen.
  $\rightsquigarrow$ Konsistenz
- Strikte Konsistenz stellt sicher, dass die wahre Regressionsfunktion auch erlernt wird.
- Strikte Konsistenz stellt anreizkompatible Vergleiche von Prognosen sicher.
- Strikt konsistente Verlustfunktionen belohnen einerseits mehr Information und andererseits den richtigen Gebrauch dieser Information.

Zusammenfassung

- Verlustfunktionen spielen eine wichtige Rolle beim Lernen und bei der Auswertung von Prognosen.
- Sie sollten immer zum relevanten Funktional passen.
  $\rightsquigarrow$ Konsistenz
- Strikte Konsistenz stellt sicher, dass die wahre Regressionsfunktion auch erlernt wird.
- Strikte Konsistenz stellt anreizkompatible Vergleiche von Prognosen sicher.
- Strikt konsistente Verlustfunktionen belohnen einerseits mehr Information und andererseits den richtigen Gebrauch dieser Information.
- Diese Ideen können im Schulkontext mittels geeigneter Beispiele vermittelt werden.

Zum Weiterlesen

- **Auswertung probabilistischer Vorhersagen**:

  T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation.

  Journal of the American Statistical Association, 102:359–378, 2007

- **Gute Einführung von Elizitierbarkeit**:

  T. Gneiting. Making and evaluating point forecasts.

  Journal of the American Statistical Association, 106(494):746–762, 2011

- **Backtesting im Bereich von Quantitative Risk Management**:

  N. Nolde and J. F. Ziegel. Elicitability and backtesting: Perspectives for banking regulation.

  The Annals of Applied Statistics, 11(4):1833–1874, 2017

- **Eigene Publikationen:**
  people.math.ethz.ch/∼tfissler/publications

Vielen Dank für eure Aufmerksamkeit!

Ich freue mich auf die Diskussion!