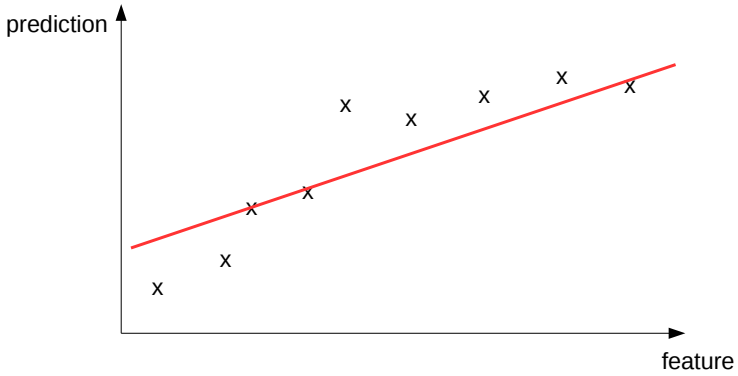
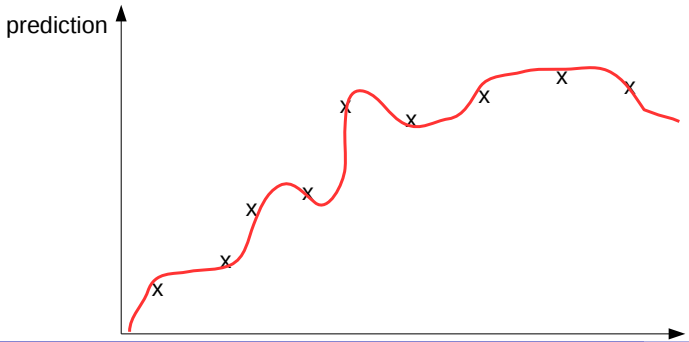
Übersicht

Feststellen von Bias oder Variance

- Bias: Unteranpassung an Trainingsdaten (underfit).
  - Modell nicht “mächtig” genug.

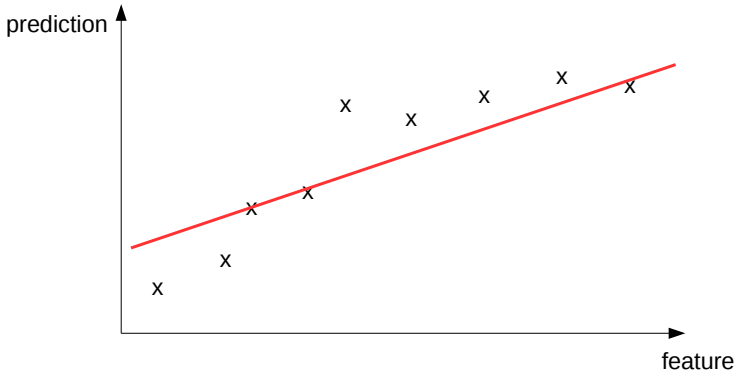


- Variance: Überanpassung an Trainingsdaten (overfit).

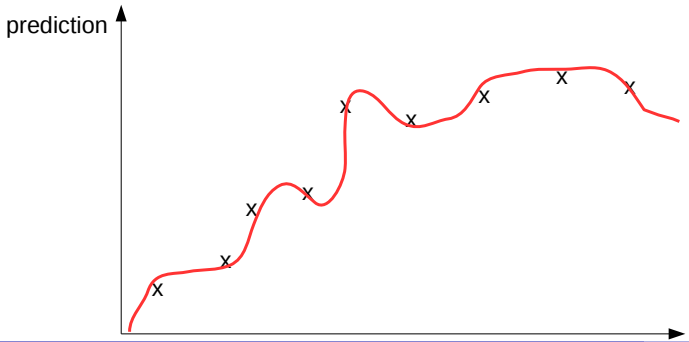


Feststellen von Bias oder Variance

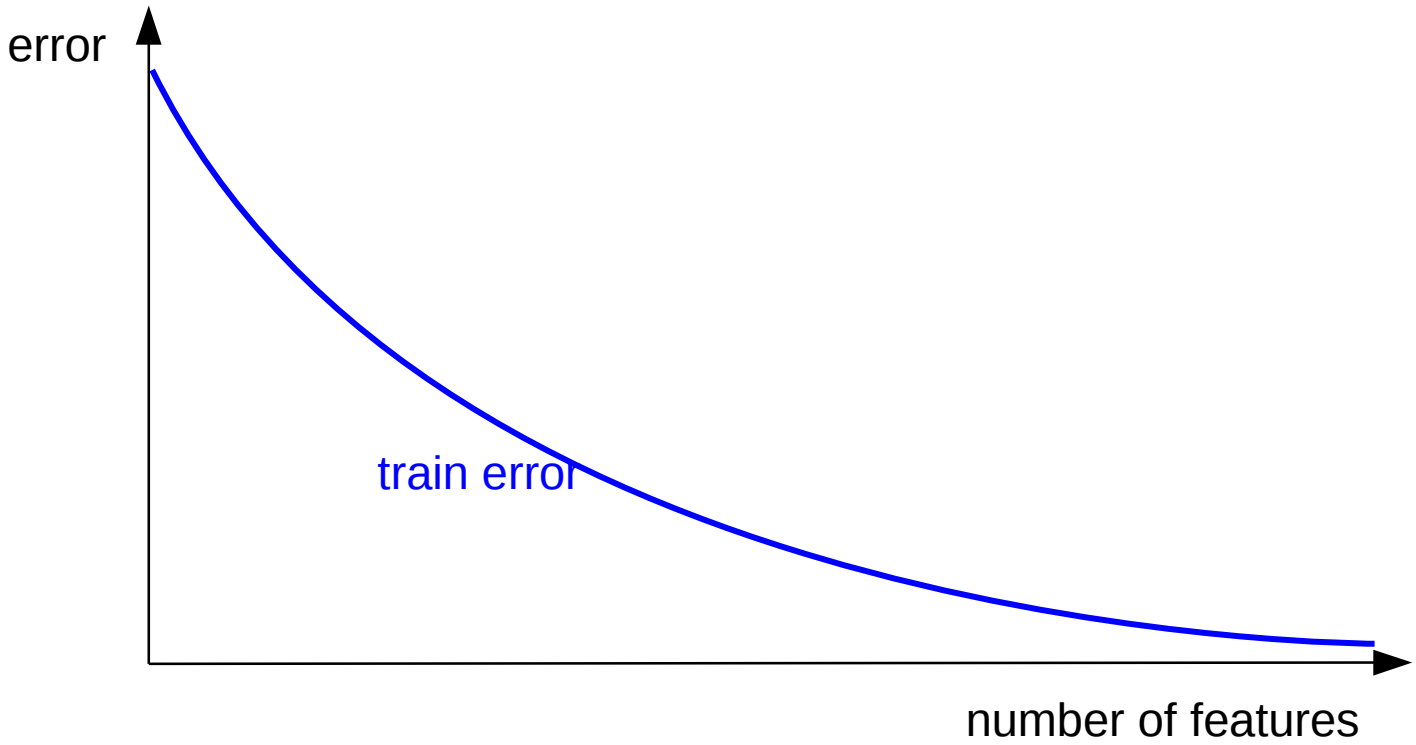
- Bias: Unteranpassung an Trainingsdaten (underfit).
  - Modell nicht "mächtig" genug.
  - Zu wenige Merkmale?
  - Zu viel Regularisierung?



- Variance: Überanpassung an Trainingsdaten (overfit).
  - Zu viele Parameter/Merkmale?
  - Zu wenig Regularisierung?
  - Zu wenige Daten?

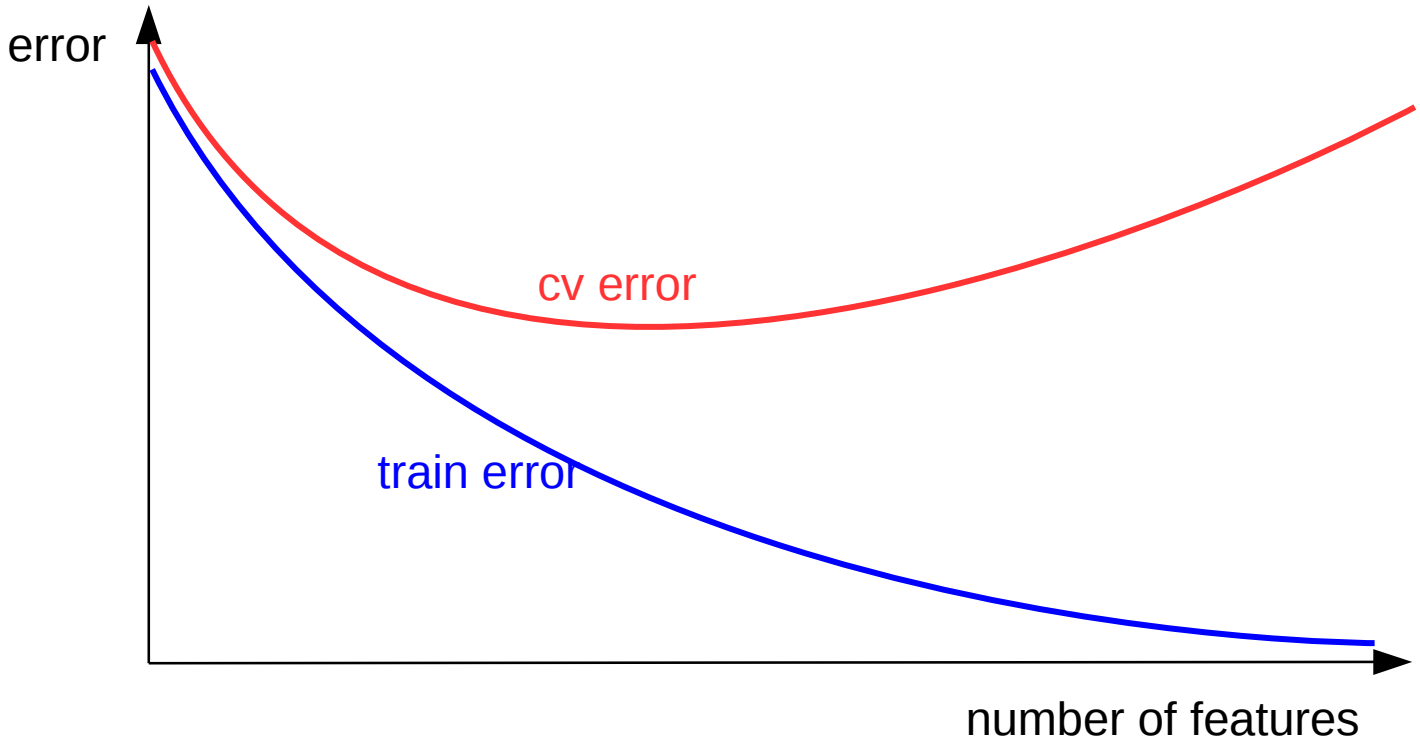


Fehlerrate bei Erhöhen der Modellkapazität



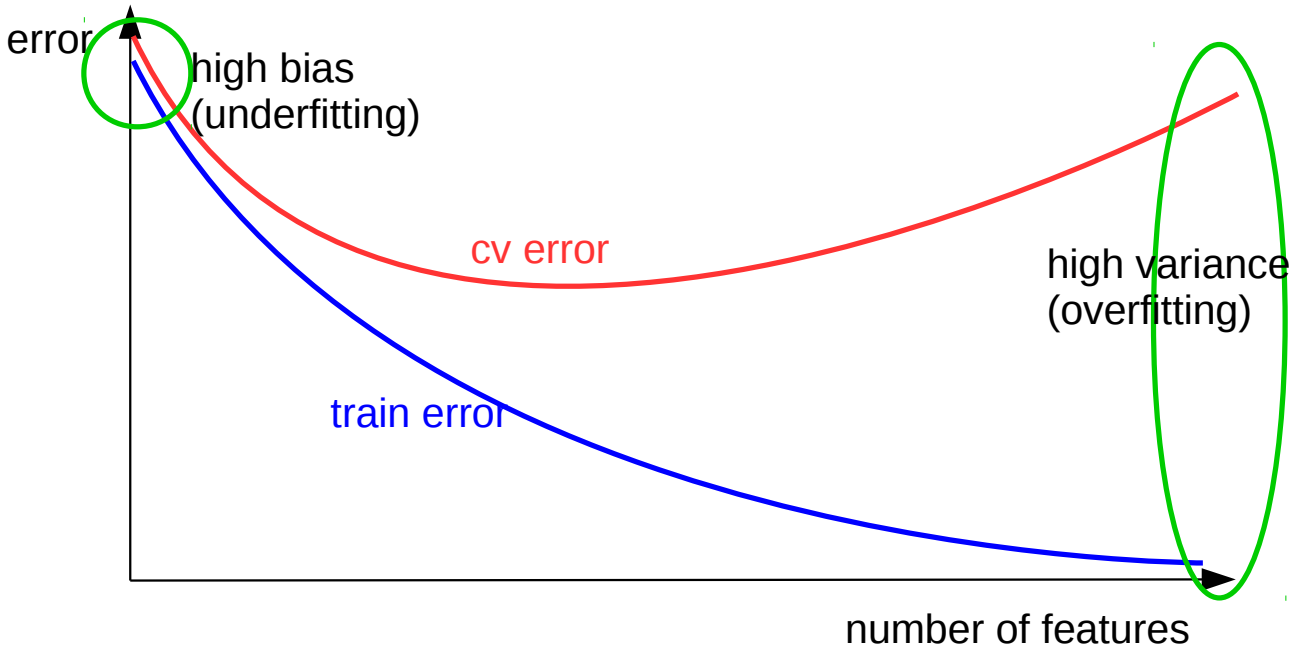
Diagnose: Underfitting oder Overfitting?

- Angenommen der Crossvalidierungsfehler is groß.
- Ist es ein Bias- (underfitting) oder Variance- (overfitting) Problem?



Diagnose: Underfitting oder Overfitting?

- Angenommen der Crossvalidierungsfehler is groß.
- Ist es ein Bias (underfitting) oder Variance (overfitting) Problem?



Diagnose: Underfitting oder Overfitting

- Bias (underfitting):
  - $J_{train}(\theta)$ hoch
  - $J_{cv}(\theta) \approx J_{train}(\theta)$
- Variance (underfitting):
  - $J_{train}(\theta)$ niedrig
  - $J_{cv}(\theta) >> J_{train}(\theta)$

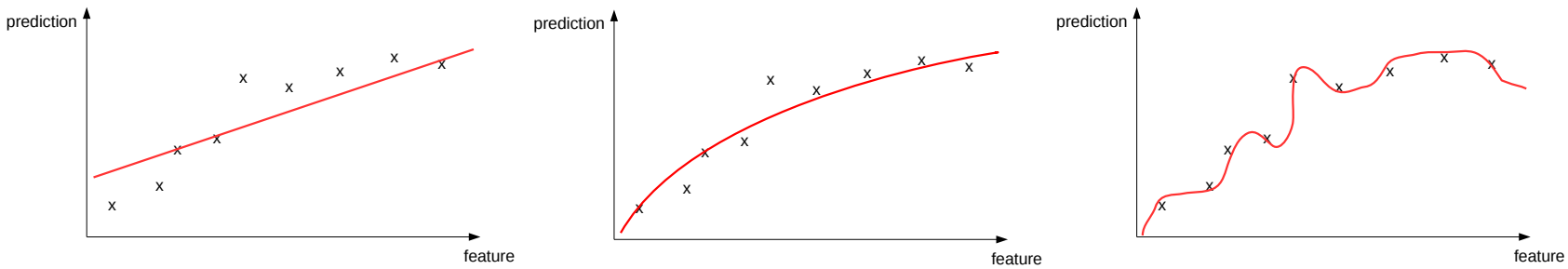
Übersicht

Regularisierung

- Term zur Fehlerfunktion hinzuaddiert (und mitoptimiert) wird, und der extreme Werte für Merkmalsgewichte bestraft.
- Zum Beispiel L2-Norm, $|\theta|_2$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_n^2$$

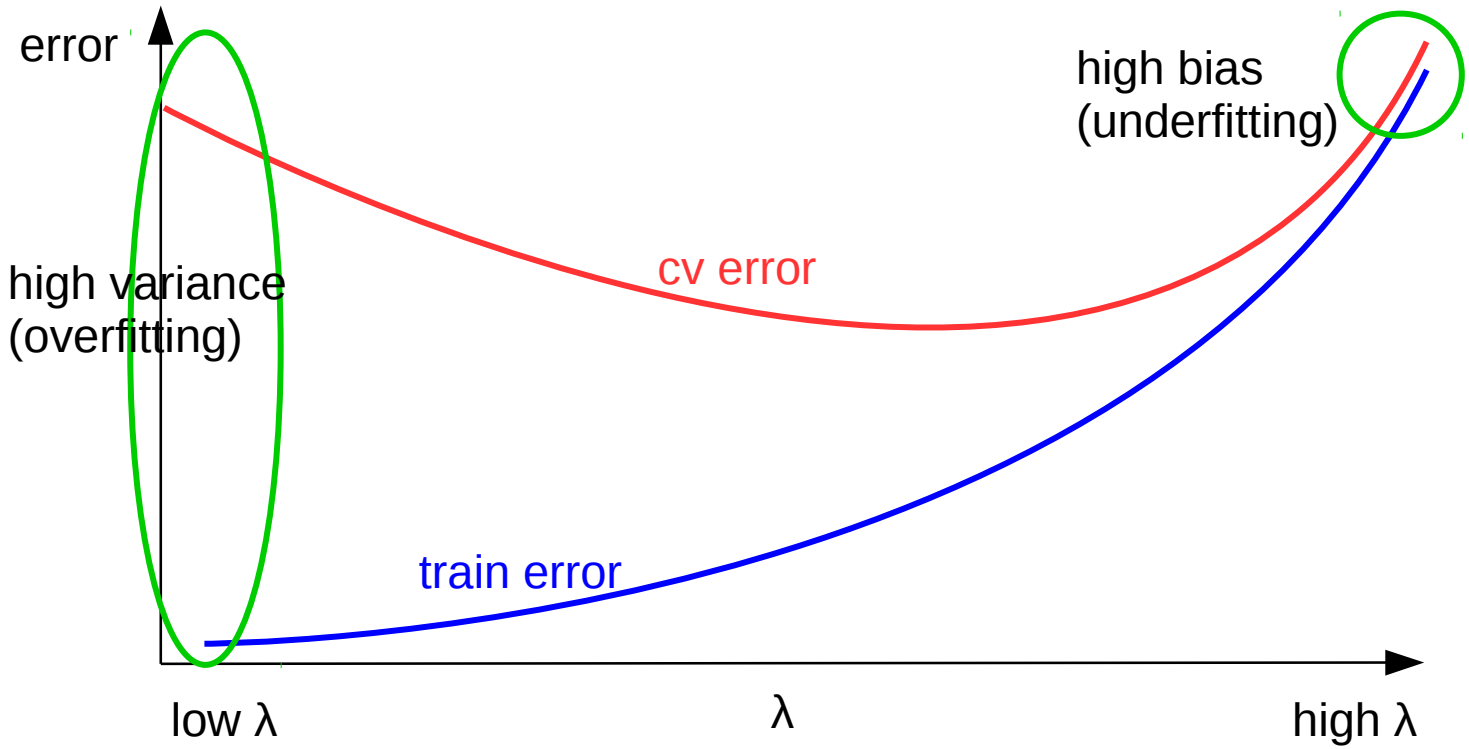
- Ausmaß der Regularisierung wird durch Hyperparameter λ gewählt.
- Kleines, "genau richtiges" and großes λ:



Auswahl des Regularisierungs-Parameters

- $J(\theta)$: zu optimierende Fehlerfunktion auf Trainingsdaten mit Regularisierungsterm.
- $J_{train}(\theta), J_{cv}(\theta), J_{test}(\theta)$: Fehlerfunktionen ohne Regularisierungsterm.
- Welches λ sollte gewählt werden? 0, 0.01, 0.02, 0.05, 0.1 ... ?
- Separates Modell für jeden Wert trainieren.
- Das Beste Modell auf Kreuzvalidierungsdaten wählen.
- Ergebnis auf den Testdaten berechnen.

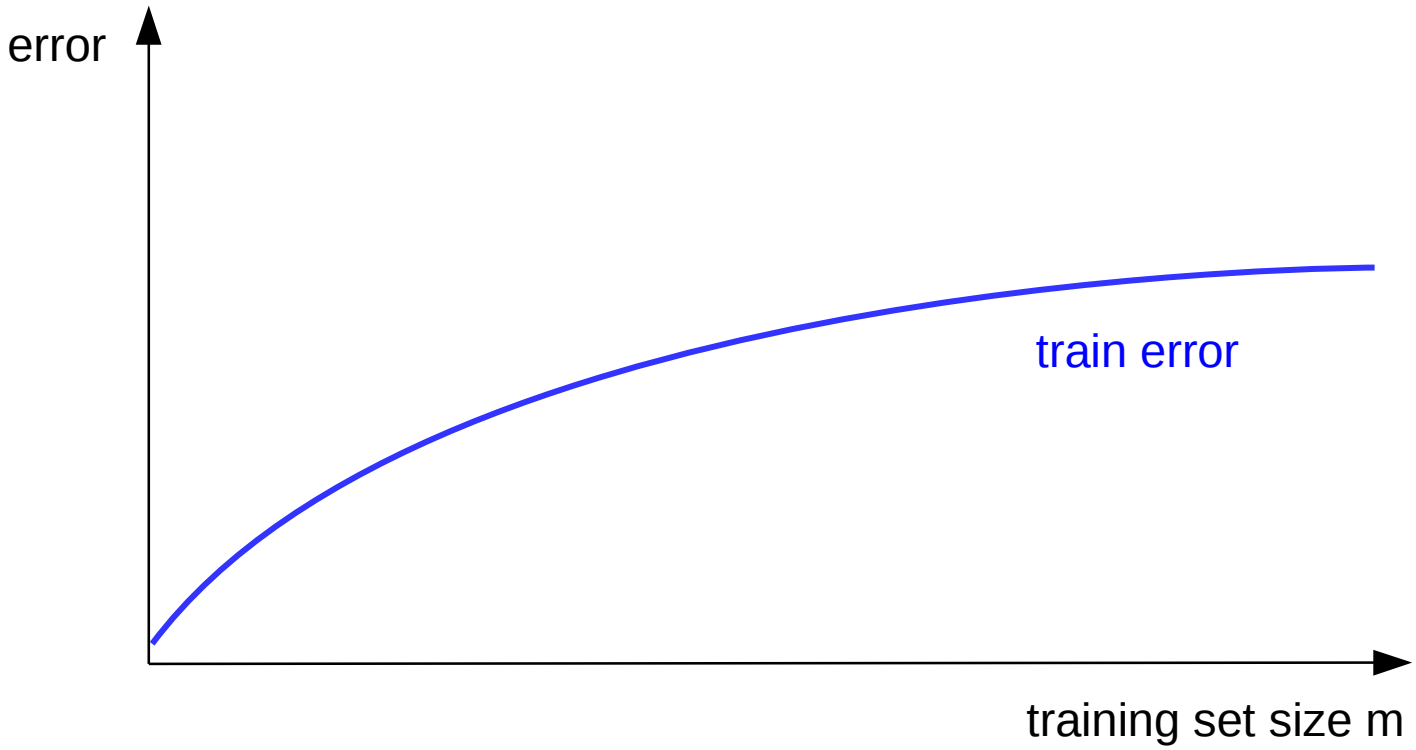
Overfitting und Underfitting in Abhängigkeit von λ



Übersicht

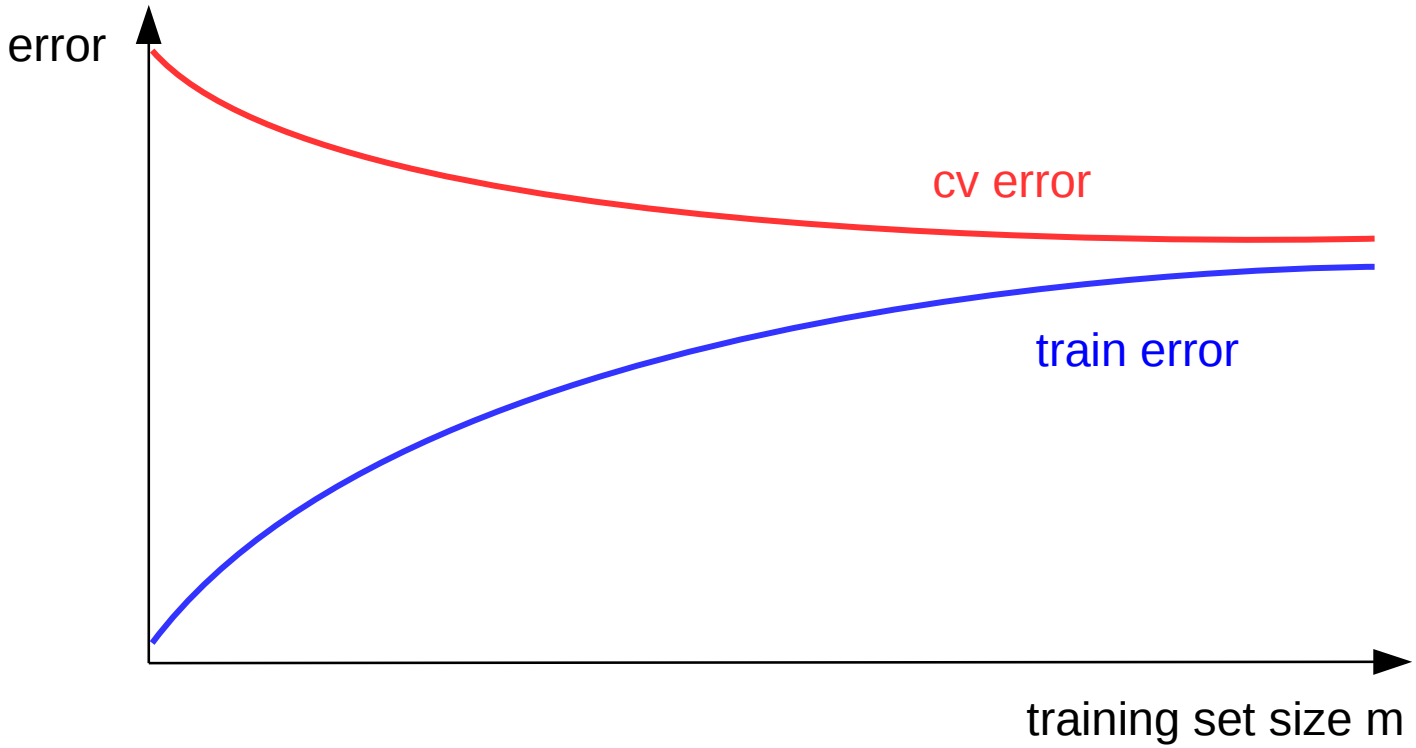
Learning Curves

- "Learning Curve": Fehlerfunktion in Abhängigkeit von der Datenmenge.
- Je mehr Daten im Training vorhanden sind, desto schwieriger ist es ein Modell zu finden, dass alle Trainingsdaten perfekt modelliert ...



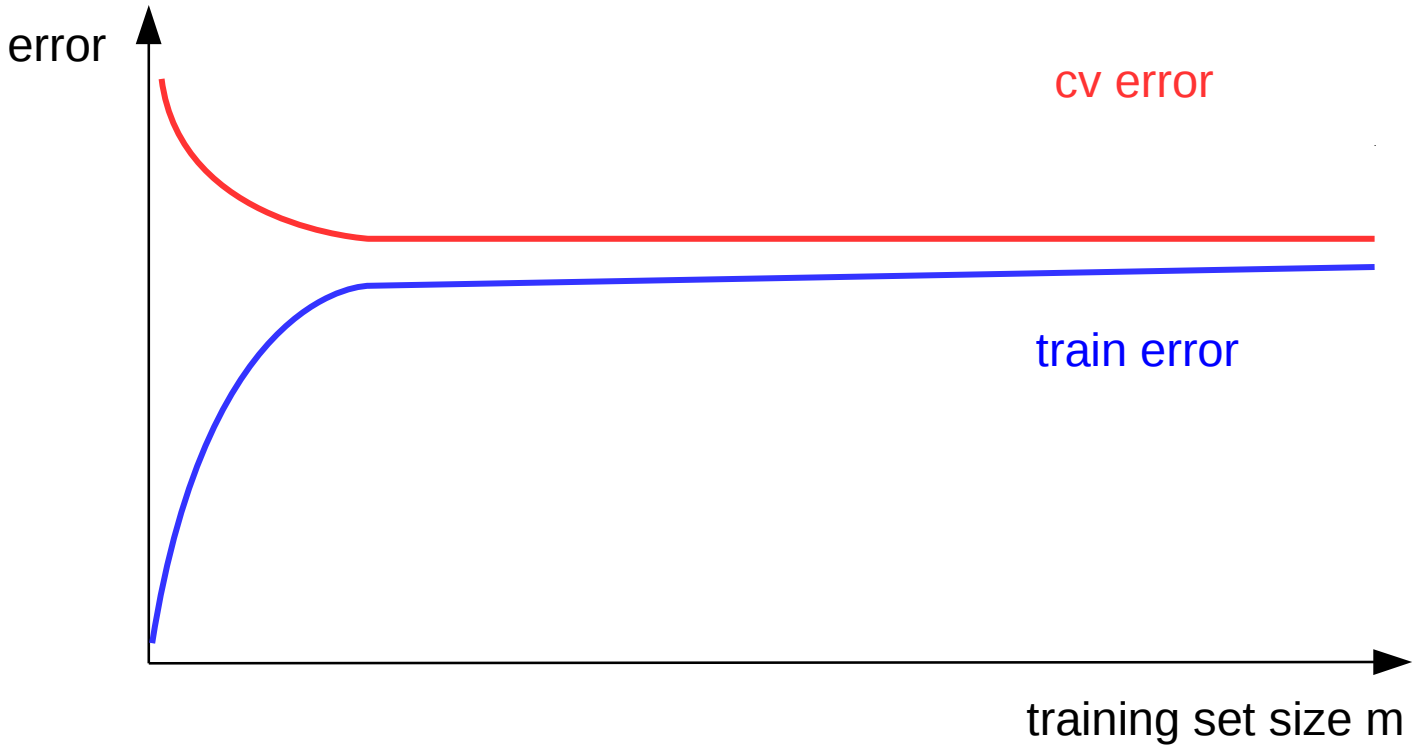
Learning Curves

- "Learning Curve": Fehlerfunktion in Abhängigkeit von der Datenmenge
- Je mehr Daten im Training vorhanden sind, desto schwieriger ist es ein Modell zu finden, dass alle Trainingsdaten perfekt modelliert ...
- ... jedoch steigt bei mehr Trainingsdaten die Qualität der Vorhersage für ungesehene Daten.



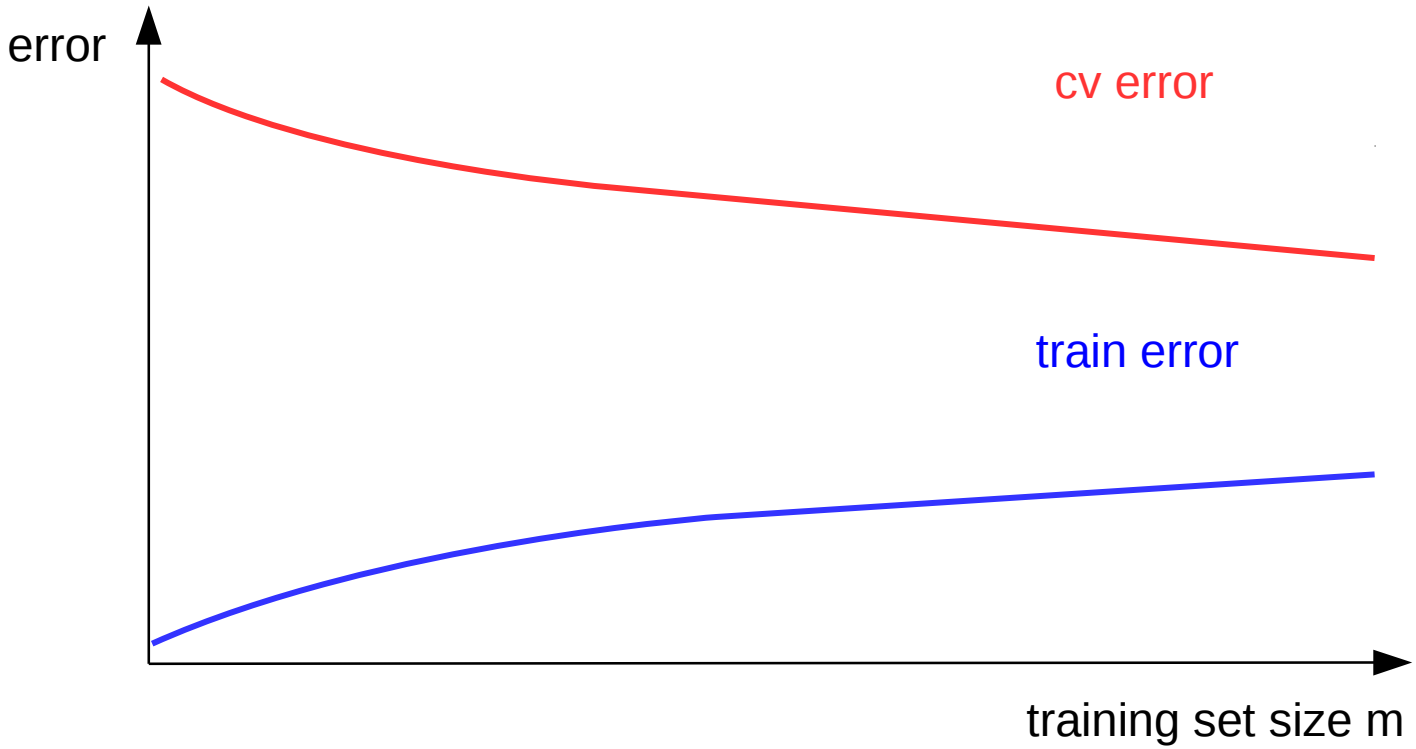
Learning Curves bei underfitting-Modellen

- Bei underfitting-Modellen ändert sich der Fehler in Abhängigkeit von zusätzlichen Daten nicht wesentlich.



Learning Curves bei overfitting-Modellen

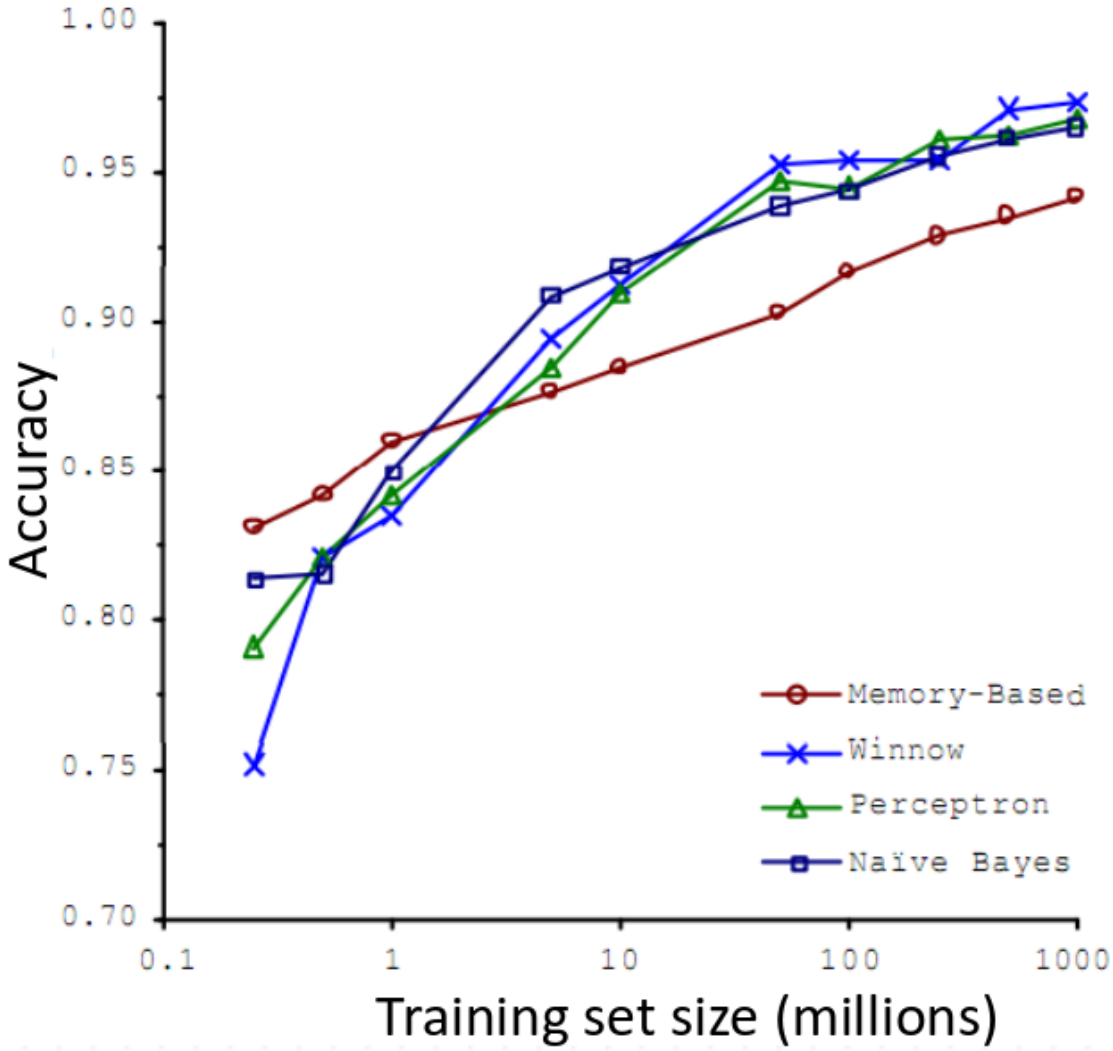
- Großer Unterschied zwischen Trainings- und Testfehler.
- Zusätzliche Trainingsdaten reduzieren den Testfehler.
- Zusätzliche Trainingsdaten erhöhen den Trainingsfehler (weniger Overfitting)



Übersicht

Sind mehr Daten immer besser?

- Daten zu gewinnen ist mit Aufwand verbunden.
- Wann lohnt sich dieser Aufwand?



[Banko & Brill, 2001]

Sind mehr Daten immer besser?

- Banko and Brill 2001: "It's not who has the best algorithm that wins. It's who has the most data."
- Annahmen:
  - Merkmale enkodieren alle wesentlichen Informationen, so dass ein Mensch die Entscheidung souverän treffen könnte.
  - Der Lernalgorithmus hat eine hohe Kapazität (hohe Varianz, overfitting).
- Unter diesen Annahmen ist es eine gute Idee, mehr Daten zu gewinnen.
- Ansonsten ist es vielversprechender, an Merkmalen und Algorithmus zu arbeiten.

Zusammenfasssung: Verbessern von Performanz

- Ausgangssituation: Klassifikator hat zu große Fehlerrate auf Kreuzvalidierungsdaten.
- Diagnostik: Learning Curves
  - Testfehler und CV-Fehler für 10%, 20%, ... 100% der Testdaten anzeigen.
  - $\Rightarrow$ Overfitting oder Underfitting.
- Nächste Schritte:
  - Problem ist Overfitting:
    - λ erhöhen
    - Weniger Merkmale
    - Mehr Trainingsdaten
  - Problem ist Underfitting:
    - λ erniedrigen
    - Merkmalskombinationen
    - Zusätzliche Merkmale

Fehleranalyse

- Beginne mit einem einfachen Algorithmus, der schnell implementiert werden kann.
- Auf Kreuzvalidierungsdaten testen und Hyperparameter optimieren.
- Learning Curves anzeigen, um zu sehen ob mehr Daten oder mehr Features helfen könnten.
- Fehleranalyse:
  - Von Hand Beispiele in den Kreuzvalidierungsdaten suchen, bei denen der Algorithmus Fehler gemacht hat.
  - Gibt es systematische Fehler?
- Falls das Problem Underfitting war, neue Features anhand der Beobachtungen konstruieren.
- Falls das Problem Overfitting war, Features anhand der Beobachtungen generalisieren (oder neue Daten gewinnen).

Fehleranalyse: Spam-Email Beispiel

- 500 Beispiele in Kreuzvalidierungsset
- 100 falsch klassifiziert
- Durchsehen und von Hand kategorisieren:
- Welche Art von Email:
  - Pharma
  - Gefälschte Produkte
  - Fishing-emails
  - andere
- Welche Features könnten helfen:
  - Länge der Email
  - Beabsichtigte 5chreibfehl3r
  - andere
- Bestimmte Merkmale müssen quantitativ evaluiert werden: Stemming, Muster in der Groß- und Kleinschreibung ...

Noch Fragen?