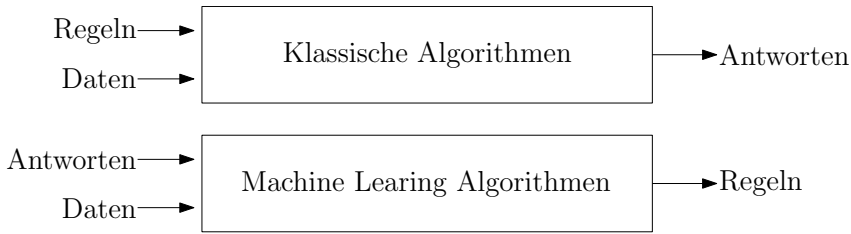
Programmierpraktikum

Pascal Wagener / Corinne Pretz

11. Juli 2022

Machine Learning



Inhaltsverzeichnis

Datensatz - Movie Ratings

Movie-Ratings

Datensatz	MovieId	UserId	Ratings (0.5-5)	Timestamp	Titel + Jahr	Genres
Ratings	✓	✓	✓	✓	×	×
Movies	✓	×	×	×	✓	✓

Ziel: Dataframe, der alle Informationen beinhaltet

1. Einlesen
2. Genres eines Filmes aufteilen (MovieId entsprechend duplizieren)
3. Zusammenführen der Dataframes (über MovieId)
4. Erscheinungsjahr vom Titel trennen und einzeln abspeichern
5. Timestamp in Daten transformieren
6. Durchschnittsbewertung und Anzahl der Bewertungen je Film hinzufügen

Jupyter - Notebook

Datensatz - Portugiesisch Kurs

Datensatz Beschreibung

- 30 Attribute (binär, numerisch, nominal)
- 3 Zielvariablen (1. Halbjahr, 2. Halbjahr, Gesamtnote)
- 649 Instanzen (Schülerdaten)
- Auszug der Attribute:

Attribut	Skala	Ausprägungen
Geschlecht	binär	männlich, weiblich
Alter	numerisch	$[15; 22]$
Vormund	nominal	Mutter, Vater, Anderer
Fahrtzeit	numerisch	< 15min, $15 - 30$min, $30 - 60$min, > 60min
Internetzugang	binär	Ja, Nein
Gesundheitszustand	numerisch	[1; 5]

Datensatz Analyse und Anpassung

1. Datensatz auf Fehldaten prüfen
2. Nominelle und binäre Attribute in Zahlenwerte umwandeln
3. (Ausreißer entfernen)
4. Korrelationen zwischen Attributen und Zielvariable anschauen (Catplots)
5. Daten entsprechend 4. anpassen

Regression

Regression

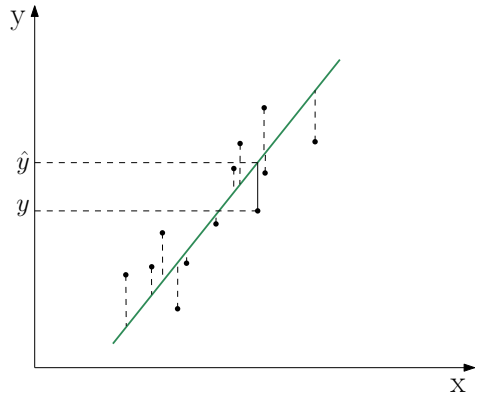
- Verfahren zur Modellierung von Zusammenhängen zwischen Daten
- Zusammenhang zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen gesucht
- Anhand dieser Zusammenhänge können Schlüsse für die Zukunft gezogen werden
- Bekannte Verfahren: Lineare Regression, Multiple lineare Regression, Polynomielle Regression
- Anwendung in der Praxis: Vorhersage von Verkaufszahlen, verrauschte Daten anpassen

Lineare Regression

Lineare Regression

Vermuteter linearer Zusammenhang zwischen abhängiger und unabhängiger Variable

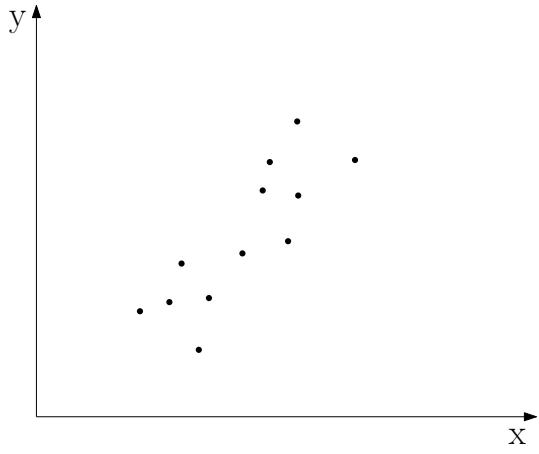
1. Spezialfall: Lineare Einfachregression
   - eine abhängige und eine unabhängige Variable
   - Funktion: $f(x) = w_0 + w_1 x$
2. Spezialfall: Multiple lineare Regression
   - mehr als eine unabhängige Variable
   - Funktion $f(x) = w_0 + \sum_{i=1}^{m} w_i x_i$



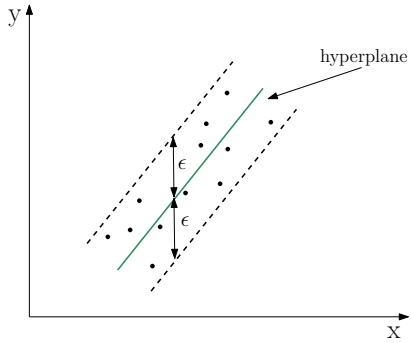
Ziel: min $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Support Vector Regression (SVR)

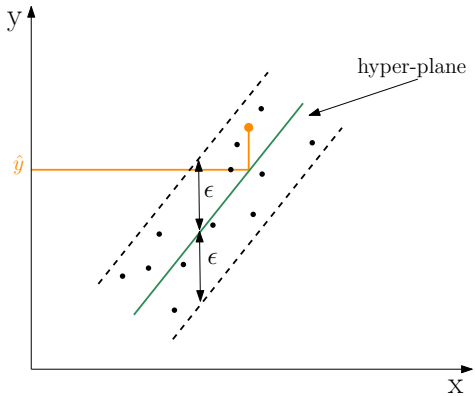
Support Vector Regression



Support Vector Regression

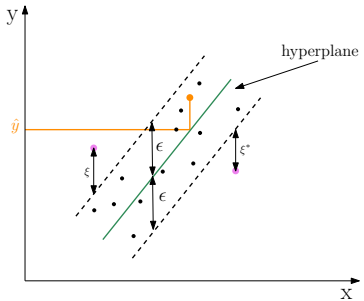


Support Vector Regression



$$\begin{aligned} &\text{min} && \frac{1}{2}||w||^2 \\ &\text{s.t.} && y_i - wx_i - b \leq \epsilon \\ & && wx_i + b - y_i \leq \epsilon \end{aligned}$$

Support Vector Regression



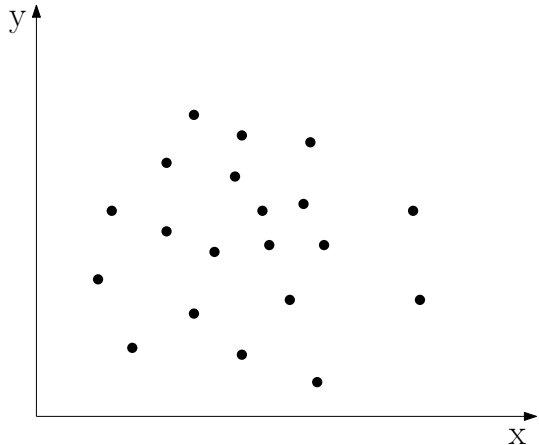
$$\text{min} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$\text{s.t.} \quad y_i - wx_i - b \leq \epsilon + \xi_i$$

$$wx_i + b - y_i \leq \epsilon + \xi_i^*$$

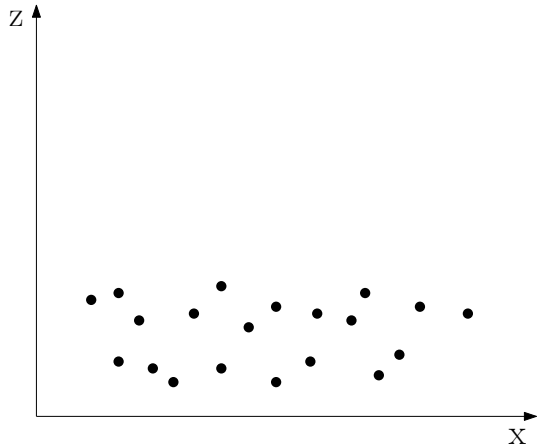
$$\xi_i, \xi_i^* \geq 0$$

SVR - Kernel Methode



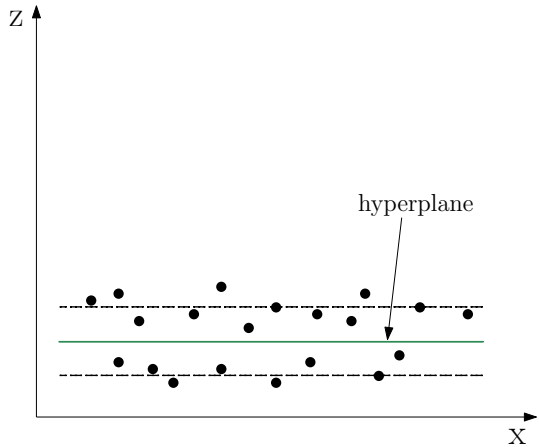
Punktewolke in einer Dimension (hier 2D) $\rightarrow$ kein linearer Zusammenhang erkennbar

SVR - Kernel Methode



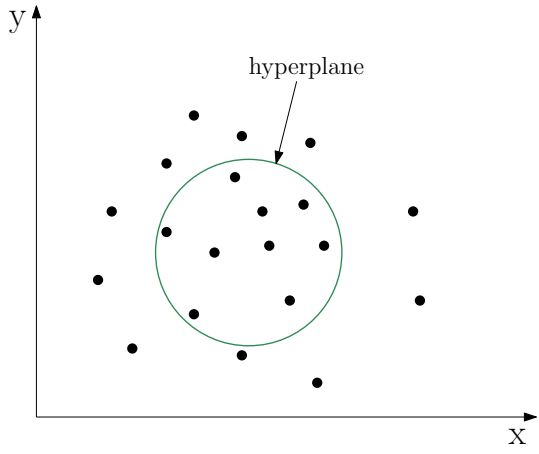
Erweiterung um eine weitere Dimension (hier 3D)

SVR - Kernel Methode



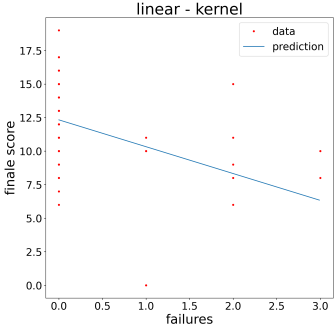
Im mehrdimensionalen ist linearer Zusammenhang erkennbar

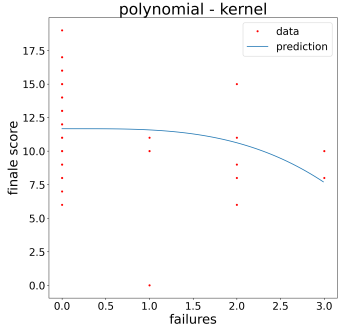
SVR - Kernel Methode

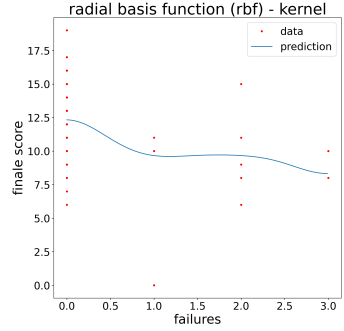


Rücktransformation in ursprüngliche Dimension

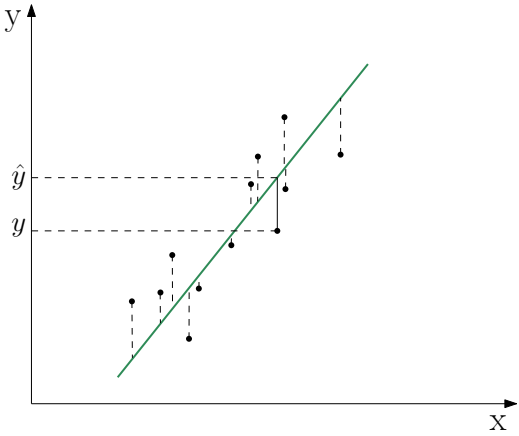
Kernel - Funktionen



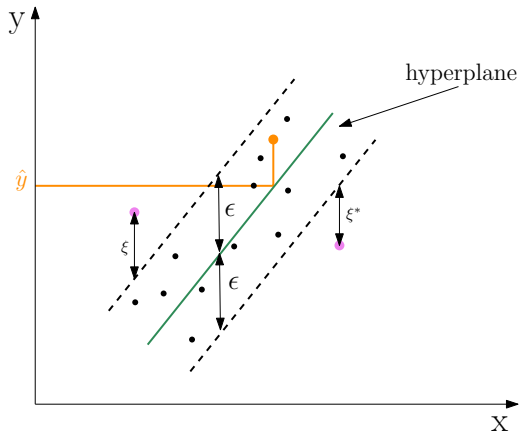




Lineare Regression vs. Support Vector Regression



Betrachtung aller Punkte



Betrachtung der Punkte außerhalb von ϵ

Statistische Kennzahlen

- Mittlerer quadratischer Fehler (MSE)

$$\frac{1}{n}\sum_{t=1}^{n}(\hat{y}-y)^2$$

- Mittlerer absoluter Fehler (MAE)

$$\frac{1}{n}\sum_{t=1}^{n}|\hat{y}-y|$$

- Bestimmtheitsmaß R^2

$$\frac{\sum_{t=1}^{n}(\hat{y}-\bar{y})^2}{\sum_{t=1}^{n}(y-\bar{y})^2}$$

Jupyter - Notebook

Literatur I

[1] Das, A.

Support vector regression (svr) based prediction with r.

https://www.youtube.com/watch?v=DBApaR2mTg0, 2022.

[2] Jahangiry, P.

Part 23-support vector machines (svm) what are they?

https://www.youtube.com/watch?v=V7mUNS3qXVY, 2022.

[3] Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K., and Tsao, C.-C.

Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang.

PloS one 12, 7 (2017), e0179763.

Literatur II

[4] Montgomery, D. C., Peck, E. A., and Vining, G. G.

Introduction to linear regression analysis.

John Wiley & Sons, 2021.

[5] scikitlearn.

sklearn.svm.svr.

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html, 2022.

[6] Sethi, A.

Support vector regression tutorial for machine learning.

https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/, 2022.