Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript zur Vorlesung

Knowledge Discovery in Databases

im Somersemester 2013

Kapitel 6: Regression

Vorlesung: Dr. Arthur Zimek
Übungen: Erich Schubert

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)

Regression

Klassifikation: Jedes Objekt o hat eine Klasse $C_i \in \{C_1, ..., C_k\}$

Klassifikator: $O \rightarrow C$

C ist diskret !!

Regression: Jedem Objekt o ist eine Zielvariable $Y \in \Re$ zugeordnet.

Regression: $O \rightarrow \Re$

Aufgabe der Regression ist die Vorhersage eines kontinuierlichen Wertes.

Zum Beispiel :

Vorhersage des erwarteten Absatzes eines Produkts

oder

empfohlene Menge an Düngemittel für einen bestimmten Bodentyp .

Arten der Regression

lineare Regression

gesuchte Vorhersage Variable Y verhält sich linear.

multiple Regression

lineare Regression, bei der Y von einem Vektor abhängt.

nicht-lineare Regression

allgemeiner Fall, die beschriebene Regressionsfunktion muss nicht-linear sein.

z.B. Logistic Regression, Poisson Regression

Lineare Regression

Gegeben: Objekt ist durch Zufallsvariable X beschrieben. Trainingsobjekte haben zusätzlich noch Ausprägungen der Zielvariable Y.

Annahme: $Y = \alpha + \beta \cdot X$

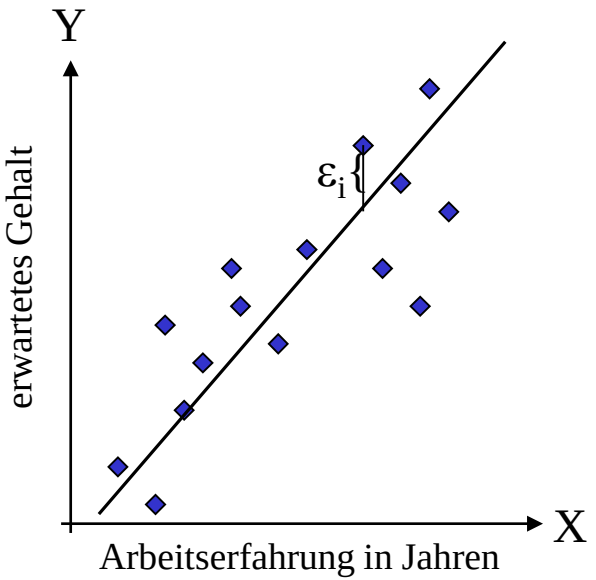
Die beobachteten Werte von Y weichen mit konstanter Varianz von der Linie ab.

Gesucht: Linie auf der Erwartungswerte liegen.

Lösung: Minimiere quadratischen Fehler. (Least Squares Methode)

D.h. bestimme α und β, so daß

$$L = \sum_{i=1}^{s} \varepsilon^2 = \sum_{i=1}^{s} \left(Y_i - \alpha - \beta \cdot X_i\right)^2 \text{ minimal wird.}$$



ε_i ist der Abstand von der angenommenen Regressionsgerade.

Lineare Regression

Lösungsansatz:
Die partiellen Ableitungen nach β und α ergeben folgende Abschätzungen.

Steigung:
$$\beta = \frac{\sum_{i=1}^{s}(x_i - E(x))(y_i - E(y))}{\sum_{i=1}^{s}(x_i - E(x))^2}$$

Y-Achsenabschnitt: $\alpha = E(Y) - \beta \cdot E(X)$

Multiple Regression

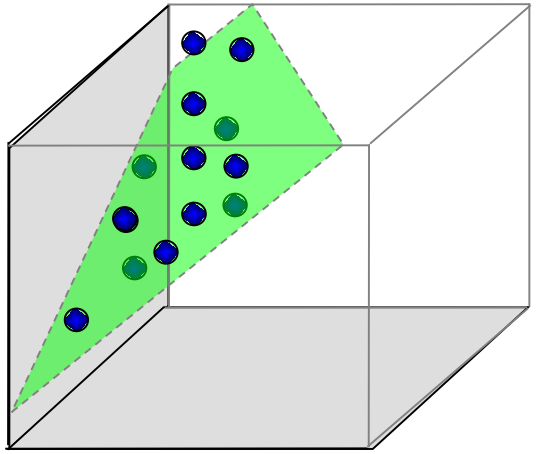
Multiple Regression:
Objekt wird durch *d*-dimensionalen Featurevektor beschrieben.

Annahme: Y sei von einer Linear-Kombination abhängig.

$$Y = \alpha + \sum_{i=1}^{d} \beta_i \cdot X_i$$

Lösung: minimiere quadratischen Fehler (Least Squares Methode)

$$L = \sum_{i=1}^{s} \varepsilon^2 = \sum_{i=1}^{s} \left(Y_i - \alpha - \sum_{k=1}^{d} (\beta_k \cdot X_{i,k}) \right)^2$$



Achtung: Es ergibt sich eine Regressionshyperebene !!

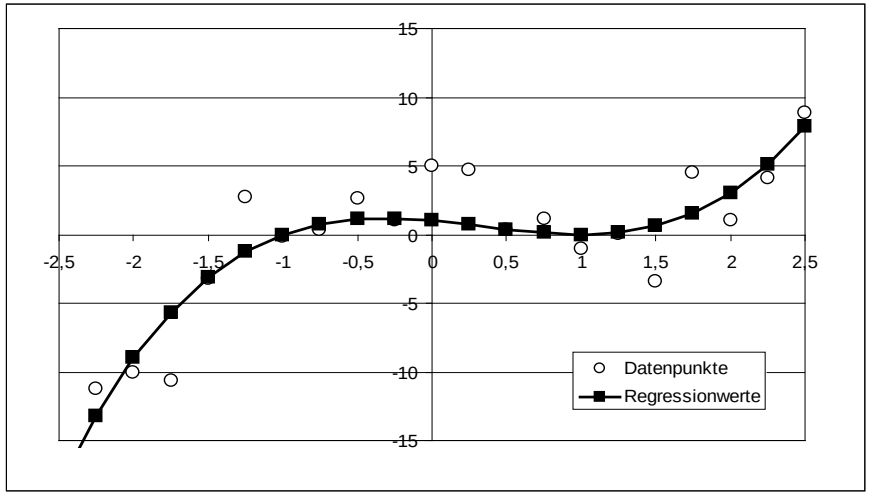
Nicht-Lineare Regression

Gegeben: Zufallsvariable X und Zielvariable Y.

Annahme: Y hängt nicht-linear von X ab.

Beispiel: polynomielle Regression
Annahme:

$$Y = \alpha + \beta_1 \cdot X + \beta_2 X^2 + \beta_3 X^3$$



Lösung: Definiere neue Variablen.

$X_1 = X$, $X_2 = X^2$, $X_3 = X^3$

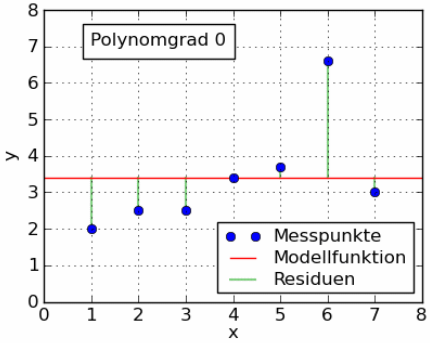
Löse lineare multiple Regression

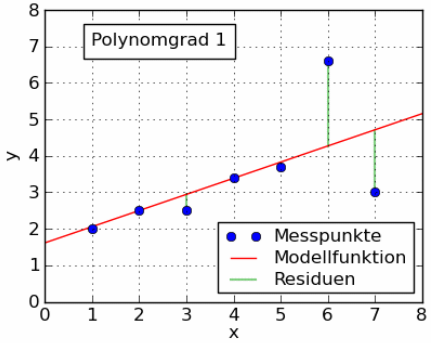
$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 X_2 + \beta_3 X_3$$

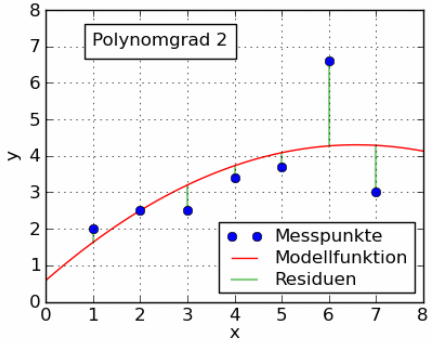
Nicht-Lineare Regression

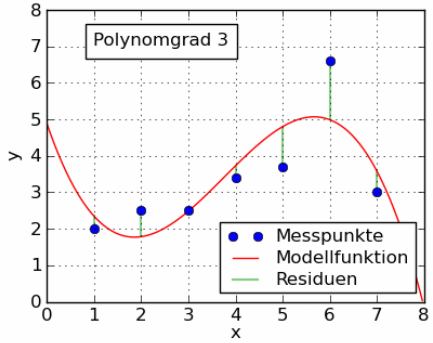
höhere Komplexität des Polynoms (=mehr Variablen, d.h., Freiheitsgrade)

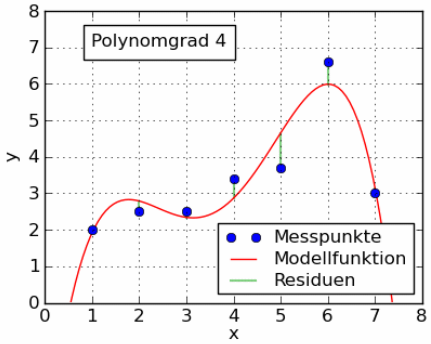
⇒ bessere Anpassung

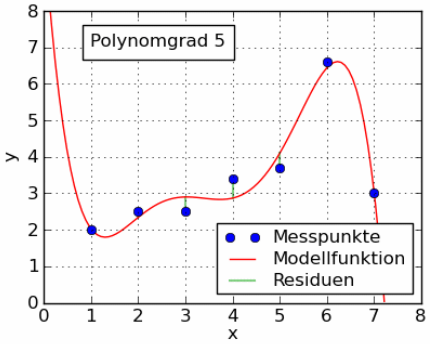


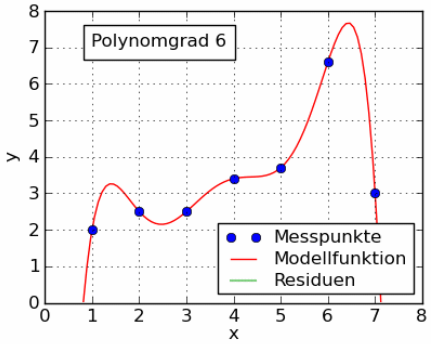












Probleme:

- Overfitting
- Outlier verzerren die Schätzung

Bild-Quelle: https://commons.wikimedia.org/wiki/File:MDKQ_anim.gif

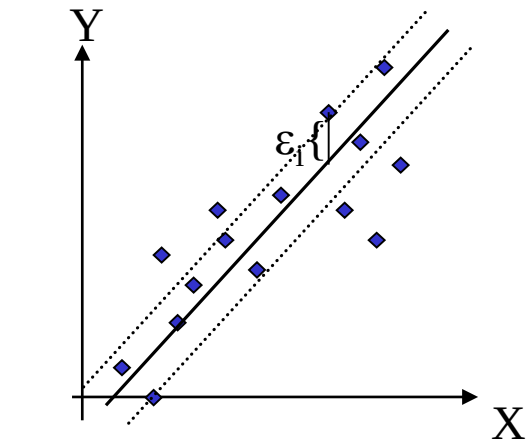
Weiterführende Regressionsverfahren

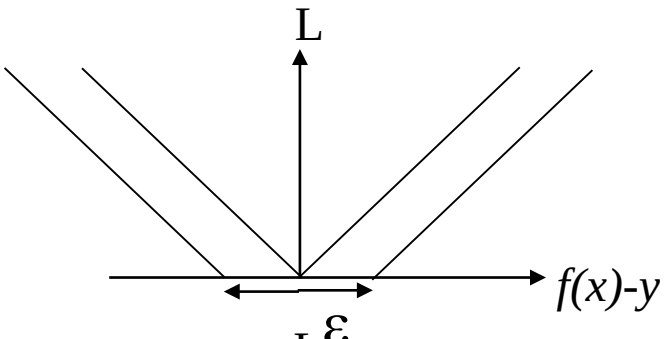
ε-insensitive Fehlerfunktion

bis jetzt müssen alle Punkte auf Regressions-gerade liegen.
Aber: Häufig ist Abstand ε von der Linie noch akzeptabel

=> definiere Rand mit Größe ε, in dem der Fehler toleriert wird.

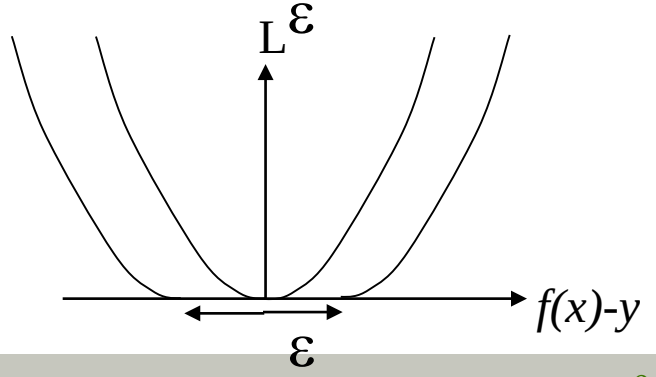
$$L^{\varepsilon}(x, y, f) = |y - f(x)|_{\varepsilon} = \max(0, |y - f(x)| - \varepsilon)$$





oder als quadratischer Fehler

$$L_2^{\varepsilon}(x, y, f) = |y - f(x)|_2^{\varepsilon}$$



Support Vector Regression

Auf Basis der ε-insensitiven Fehlerfunktion kann man jetzt ein Optimierungsproblem ähnlich zu dem der SVMs definieren

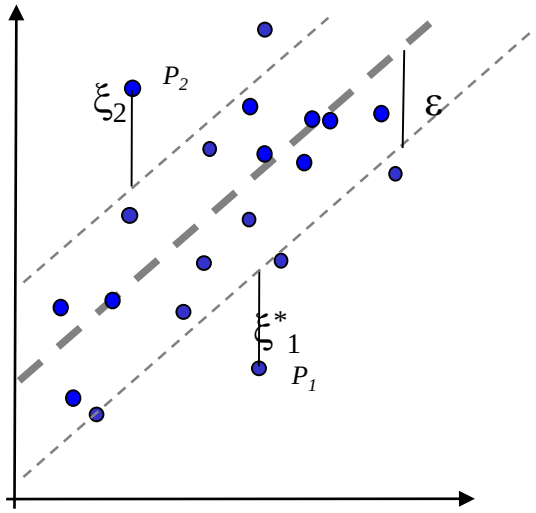
Primäres OP :

$$\text{minimiere } J(\vec{w}, b, \vec{\xi}) = \left\|\vec{w}\right\|^2 + C \cdot \sum_{i=1}^{n} \xi_i^2 + \xi_i^{*2}$$

unter Nebenbedingung für $\forall i \in [1..n]$ sei

$$y_i - \left(\left\langle \vec{w}, \vec{x}_i \right\rangle + b\right) \leq \varepsilon + \xi_i$$

$$\left(\left\langle \vec{w}, \vec{x}_i \right\rangle + b\right) - y_i \leq \varepsilon + \xi_i^* \quad \text{und} \quad \xi_i, \xi_i^* \geq 0$$



- 2 Typen von Slack-Variablen für überhalb und unterhalb des Zielwertes y
- Beachte: $\xi_i \xi^*_i = 0$, da Objekt entweder überhalb oder unterhalb der Regressionsgerade liegt.

Support Vector Regression

Überführt in eine Form mit Lagrange Multiplikatoren:

Duales OP: maximiere

$$L(\vec{\alpha}) = \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) - \varepsilon \sum_{i=1}^{n}(\alpha_i^* + \alpha_i) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^* - \alpha_i)\cdot(\alpha_j^* - \alpha_j)\cdot\left(\langle \vec{x}_i, \vec{x}_j \rangle + \frac{1}{C}\delta_{ij}\right)$$

mit Bedingung $\sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0$ und $0 \le \alpha_i, 0 \le \alpha^*_i, i=1..n$

Beachte, dass hierbei gilt: $\alpha_i \alpha^*=0$ und $\xi_i \xi^*_i=0$

Verallgemeinertes Problem mit Kernelfunktion:

Duales OP: maximiere $L(\vec{\alpha}) = \sum_{i=1}^{n} y_i\alpha_i - \varepsilon\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i \cdot \alpha_j \cdot K(\vec{x}_i, \vec{x}_j)$

mit Bedingung $\sum_{i=1}^{n}\alpha_i = 0$ und $-C \le \alpha_i \le C$, $i=1..n$

Support Vector Regression

Anmerkungen:

- Über Kernel lässt sich bequem nicht lineare Regression realisieren
- Training mit den gleichen Lösungsverfahren wie bei SVMs für die Klassifikation
- Es gibt weitere Varianten: z.B. Ridge-Regression.
  Hierbei ist $\varepsilon = 0$, d.h. es handelt sich um Least Squares Regression mit einer Einschränkung der Gewichte.

Ridge Regression :

minimiere $J(\vec{w}, b, \vec{\xi}) = \lambda \left\| \vec{w} \right\|^2 + \sum_{i=1}^{n} \xi_i^2$

unter Nebenbedingung für $\forall$i $\in$ [1..n] sei $y_i - \left(\left\langle \vec{w}, \vec{x}_i \right\rangle + b \right) = \xi_i$, i= 1,..,l

Fazit Regression

- Regression löst ein ähnliches Problem wie Klassifikation. Vorhersage: kontinuierliche Werte.
- Regressionsgeraden können häufig analytisch bestimmt werden.
- Weiterführende Verfahren mit Kernelfunktionen.
- Anmerkung: Logistische Regression Anwendung von Regression auf Klassifikation.
  (Klasse A: Y =1, Klasse B: Y = 0)