Clevere Analysen in Oracle ohne kostenpflichtige Zusatzoptionen

Dr. Bernd Günther
DB Systel GmbH
Frankfurt am Main

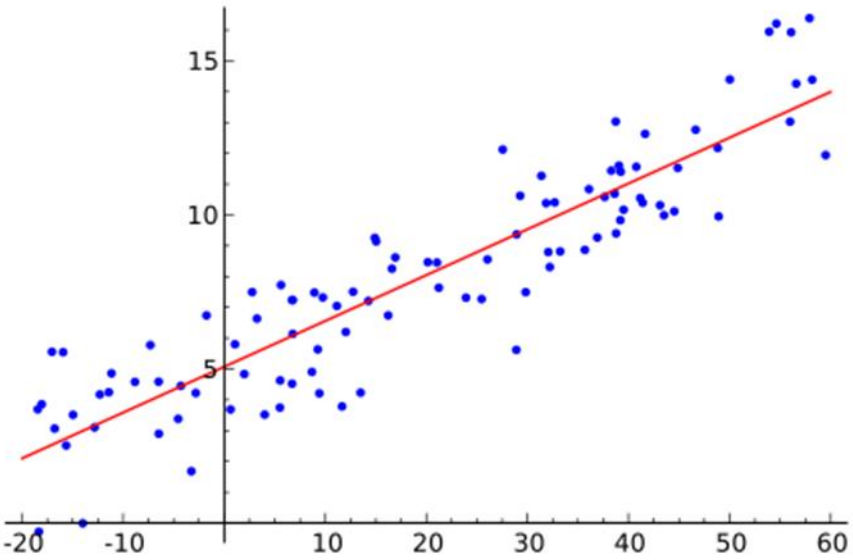
Schlüsselworte

Advanced Analytics, Lineare Regression, Numerische lineare Algebra, UTL_NLA.

Einleitung

Lineare Regressionsanalysen gehören zu den ältesten und wichtigsten Methoden der Wissensgewinnung in Datenbanken und werden u.a. von Oracle Datamining unterstützt, was jedoch eine Zusatzlizenz erfordert. Wir zeigen an Hand praktischer Beispiele, dass bereits die Standardversion der Datenbank alle erforderlichen Mittel bereitstellt, um solche Analysen einfach und effizient durchzuführen. Auch Standardschätzer und Approximationsqualität können berechnet sowie Einflussfaktoren beurteilt werden.

Der eindimensionale Fall



Im eindimensionalen Fall kann die Anpassung einer Geraden

$$y = mx + b$$

an eine vorgebene Menge von Punkten $(x^1, y^1), \cdots (x^n, y^n)$ mit Hilfe der analytischen Funktionen bzw. Aggregatfunktionen REGR_SLOPE und REGR_INTERCEPT usw. vorgenommen werden.

Der in der Praxis wird meist gefragte, zwei- oder mehrdimensionale Fall der Anpassung einer Ebene

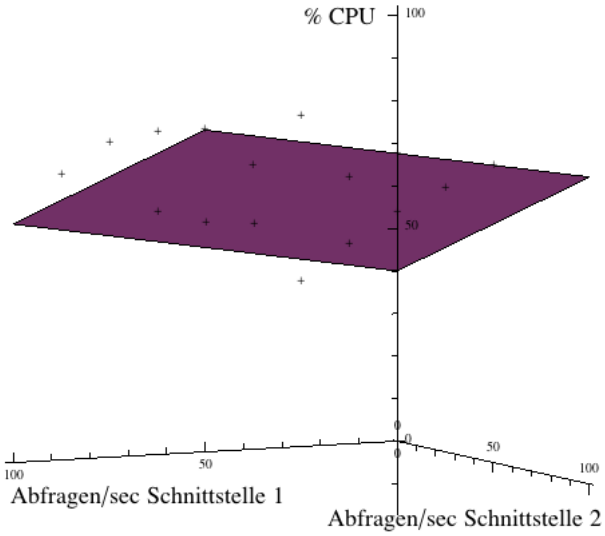
$$y = a_0 + a_1 x_1 + a_2 x_2$$

(bzw. $y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$) an eine Punktwolke ist hierdurch jedoch nicht abgedeckt.

Der mehrdimensionale Fall

Anwendungsbeispiel:

$$\%CPU \approx a + b \times Abfragen\_SS1/sec + c \times Abfragen\_SS2/sec$$



Ein möglicher Weg ist die Erstellung eines linearen Modells im Oracle-Dataminer:

bzw. der entsprechenden PL/SQL-Funktionalität

```
DBMS_DATA_MINING.CREATE_MODEL.
```

(Näheres hierzu siehe Brendan Tierney: Predictive Analytics Using Oracle Data Miner, Kapitel 10 und 17). Dies erfordert jedoch den Kauf der Advanced-Analytics-Option sowie die Bewältigung des Modelling Frameworks.

Lineare Regression in der Oracle-Standard-Version

Tatsächlich hält bereits die Standard-Version der Oracle-Datenbank für mehrdimensionale lineare Regressionen alle erforderlichen Bausteine, die nur noch geschickt zusammengesetzt werden müssen. Schauen wir nämlich in irgendein Buch über Statistik (z.B.: Ansgar Steland: Basiswissen Statistik, 3. Auflage), so erkennen wir, dass lediglich zwei Ingredienzen benötigt werden:

1. Aggregatfunktionen (SUM, AVG)
2. sowie die Lösung linearer Gleichungssysteme.

Ersteres bedarf keiner weiteren Ausführung. Die Behandlung linearer Gleichungen erfolgt in Oracle mittels der Package

UTL_NLA

(utility numerical linear algebra), die Oracles Implementierung der Open Source Packages BLAS (Basic Linear Algebra Subprograms) und LAPACK (Linear Algebra PACKage) zusammenfasst. In unserem Kontext wird davon die Funktion

```
UTL_NLA.LAPACK_GESV
```

benötigt (Auflösung linearer Gleichungssysteme, Matrixinvertierung).

Bewertung der Representativität des Datenbestandes, Einflussfaktoren und Überanpassung

Fügt man dem linearen Ausdruck $y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$ immer mehr unabängige Variablen $x_{m+1}, x_{m+2}, \cdots$ hinzu, so wird zwar die Genauigkeit der *Anpassung* immer besser, aber falls der Datenbestand nicht umfangreich genug ist, können die Koeffizienten nicht genau genug bestimmt werden und die *Prognosequalität* nimmt möglicherweise ab: Es liegt ein Fall von Überanpassung vor. Diesen Sachverhalt kann man erkennen, indem man die Wolke der Datenpunkte durch ein Ellipsoid beschreibt und prüft, ob es raumfüllend ist oder ob eine seiner Hauptachsen sehr flach ausfällt. Hierbei kommt die Funktion UTL_NLA.LAPACK_SYEV (Eigenwerte und -vektoren, Hauptachsentransformation) zum Einsatz; der Dataminer stellt diese Funktionalität nicht zur Verfügung.

Kontaktadresse:

Dr. Bernd Günther
Reporting & Analytics (I.LPA46)
DB Systel GmbH
Jürgen-Ponto-Platz 1
60329 Frankfurt am Main