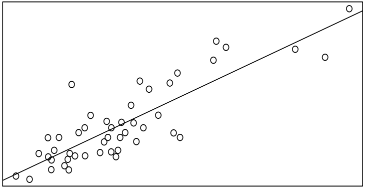
Lineare Regression

Kleinste Quadrate, Maximum Likelihood und Bayes

Jonas Nick

February 18, 2013



Lineare Regression: Least-squares, Maximum Likelihood und Bayes

Problembeschreibung

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n, \;\; y \in \mathbb{R} \wedge y \in G(Gruppen)$$

Regression

$$y = h(x) + \epsilon$$

Praktische Fragestellung, X und Y aus dem Trainingsdatensatz, x und y aus dem Testdatensatz:

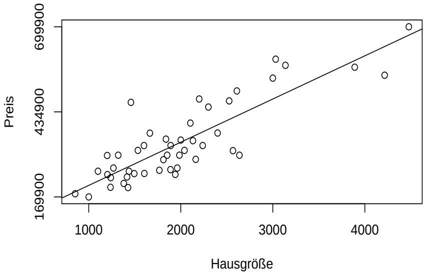
$$\operatorname*{argmax}_y p(y|x, X, Y)$$

Lineare Regression

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$h_\theta(x) = \theta^T x$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$



Nichtlineare Regression

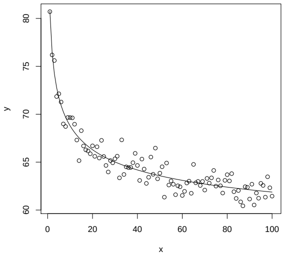
Nichtlineare Regression

Linearkombination der Basisfunktionen ϕ

$$h_\theta(x) = \theta_0 + \theta_1\phi_1(x) + \theta_2\phi_2(x) + \theta_3\phi_3(x)$$

Beispiel:

$$\phi_1(x) = x_1,$$
$$\phi_2(x) = 0.1x_2$$
$$\phi_3(x) = 30x_3^{-0.2}$$



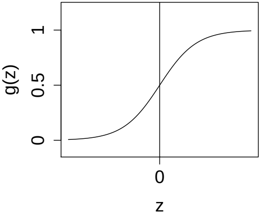
Logistische Regression

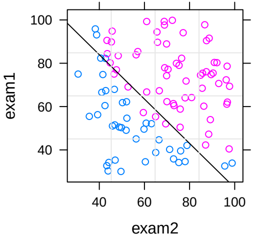
$$y \in \{0, 1\}$$

$$h_\theta(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Falls $h_\theta(x) \geq 0.5$
prognostiziere "y=1"
sonst
prognostiziere "y=0"





$$X = \begin{bmatrix} -{x^{(1)}}^T- \\ -{x^{(2)}}^T- \\ \vdots \\ -{x^{(m)}}^T- \end{bmatrix} \in \mathbb{R}^{mxn} \; Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^{m}$$

Kostenfunktion

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, J(\theta)$$

Maximum Likelihood Methode

Maximum-Likelihood-Methode

Es bezeichne $\mathcal{D} = (d^{(1)}, d^{(2)} \dots d^{(m)})$ Realisierungen von Zufallsvariablen mit zugehöriger Wahrscheinlichkeitsdichte $p(\mathcal{D}|\delta)$.

$$L(\delta) := p(\mathcal{D}|\delta)$$

Wähle zu den Beobachtungen $\mathcal{D}$ als Parameterschätzung denjenigen Parameter $\hat{\delta}$, für den die Likelihood maximal ist, d.h.

$$\hat{\delta} = \underset{\delta}{\operatorname{argmax}}\, L(\delta)$$

Beispiel:

$$p(d|\mu,\sigma) = \mathcal{N}(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(d-\mu)^2}{2\sigma^2})$$

Kleinste Quadrate Herleitung

Annahme für das Regressionsproblem:

$$L(\theta) = p(y|x, \theta) = \mathcal{N}(h_\theta(x), \sigma_1)$$

$$p(Y|X, \theta) = \prod_{i=1}^{m} \mathcal{N}(h_\theta(x^{(i)}), \sigma_1)$$

$$\ln p(Y|X, \theta) = -\frac{\sigma_1}{2} \sum_{n=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{m}{2} \ln \sigma_1 - \frac{m}{2} \ln(2\pi)$$

Kleinste Quadrate (Least Squares)

Kleinste Quadrate Kostenfunktion

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

Normalengleichung

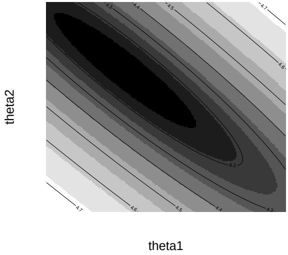
$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)} \stackrel{!}{=} 0$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T y$$

Gradientenverfahren

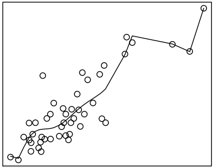
Gradientenverfahren

α: Lernrate
repeat until convergence {
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$
}



Regularisierung

Overfitting



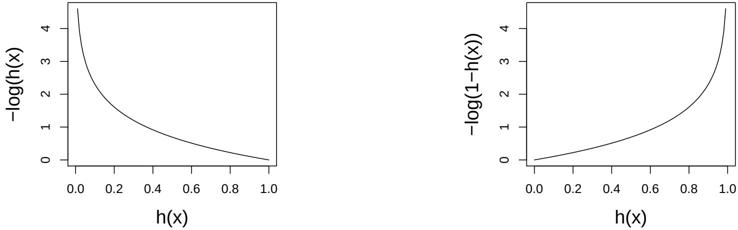
Regularisierte Kleinste Quadrate

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

Logistische Regression

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)})$$

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$



Logistische Regression Kostenfunktion

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

Bayes Inferenz

Bayes-Inferenz

Die Wahrscheinlichkeitsdichte von $\mathcal{D}$, gegeben δ, sei $p(d|\delta)$ und $L(\delta) = p(\mathcal{D}|\delta)$ die Likelihoodfunktion. Für den unbekannten Parameter wird eine a priori Dichte

$$p(\delta)$$

spezifiziert. Dann ist die a posteriori Dichte über den Satz von Bayes bestimmt durch

$$p(\delta|\mathcal{D}) = \frac{p(d|\delta)p(\delta)}{p(\mathcal{D})} = \frac{L(\delta)p(\delta)}{\int L(\delta)p(\delta)d\delta}$$

Bayes-Schätzer

Maximum a posteriori (MAP) Schätzer

Wähle denjenigen Parameterwert $\hat{\delta}_{MAP}$, für den die a posteriori Dichte maximal wird, d.h.

$$\hat{\delta}_{MAP} = \underset{\delta}{\operatorname{argmax}}\, L(\delta)p(\delta)$$

Bayes'sche Lineare Regression

Parametervektor θ

a priori Wahrscheinlichkeitsdichte: $p(\theta) = \mathcal{N}(0, \sigma_0)$
Likelihood: $L(\theta) = p(y|x, \theta) = \mathcal{N}(h_\theta(x), \sigma_1)$
a posteriori Dichte: $p(\theta|x, y) \propto p(y|x, \theta)p(\theta)$

Kostenfunktion

$$J(\theta) = \frac{\sigma_1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\sigma_0}{2} \theta^T \theta$$

Vorhersage eines Zielmerkmals

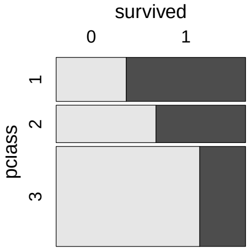
$$p(y|x, X, Y) = \int p(y|x, \theta)p(\theta|X, Y)d\theta$$

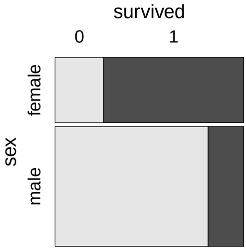
Datensatz

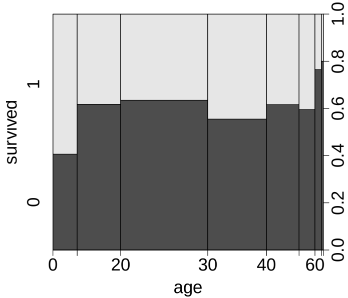
	survived	pclass	sex	age	sibsp	parch	ticket	fare	cabin	embarked
1	0	3	male	22	1	0	A/5 21171	7.2500		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.9250		S
4	1	1	female	35	1	0	113803	53.1000	C123	S
5	0	3	male	35	0	0	373450	8.0500		S
6	0	3	male	NA	0	0	330877	8.4583		Q

⋮

Datensatz


survived
0
1
pclass
1
2
3


survived
0
1
sex
female
male


survived
0
1
age
0
20
30
40
60
0.0
0.2
0.4
0.6
0.8
1.0

Modell

```
model <- glm(survived~age + sex + I(pclass==1)
                + I(pclass==2),data=titanicTrainData,
                family=binomial("logit"))
```

```
Trainingsdaten: 78.8515406162465% korrekte Vorhersagen
```

Modell

```
Call:
glm(formula = survived ~ age + sex + I(pclass == 1) + I(pclass ==
    2), family = binomial("logit"), data = titanicTrainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7303  -0.6780  -0.3953   0.6485   2.4657

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.196387   0.252649   4.735 2.19e-06 ***
age                -0.036985   0.007656  -4.831 1.36e-06 ***
sexmale            -2.522781   0.207391 -12.164  < 2e-16 ***
I(pclass == 1)TRUE  2.580625   0.281442   9.169  < 2e-16 ***
I(pclass == 2)TRUE  1.270826   0.244048   5.207 1.92e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.28  on 709  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 657.28
```

Modell

```
model <- glm(survived~I(log(age)) + sex
                + pclass +  sibsp,
                data=titanicTrainData,
                family=binomial("logit"))
```

```
Trainingsdaten: 81.3725490196078% korrekte Vorhersagen
Testdaten: 75.1% korrekte Vorhersagen
```

Modell

```
Call:
glm(formula = survived ~ I(log(age)) + sex + pclass + sibsp,
    family = binomial("logit"), data = titanicTrainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1480  -0.6031  -0.3743   0.5799   2.3857

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.1187     0.6917   8.846  < 2e-16 ***
I(log(age))  -0.9626     0.1623  -5.931 3.01e-09 ***
sexmale      -2.7015     0.2184 -12.372  < 2e-16 ***
pclass2      -1.3617     0.2786  -4.887 1.02e-06 ***
pclass3      -2.5387     0.2705  -9.384  < 2e-16 ***
sibsp        -0.5035     0.1300  -3.874 0.000107 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 623.79  on 708  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 635.79
```

Zusammenfassung

- Ein Regressionsmodell verknüpft y mit einer Funktion von x und θ.
- Lineare Regression ist die Linearkombination von gewichteten erklärenden Merkmalen.
- Kleinste Quadrate Kostenfunktion ergibt sich aus der Maximum Likelihood Parameterschätzung.
- Logistische Regression klassifiziert mittels Verbindungsfunktion.
- Bayes Regression nimmt eine a priori Verteilung der Koeffizienten an womit sich ihre a posteriori Verteilung berechnet.