Empolis Machine Learning Workshop

– März 2016 –

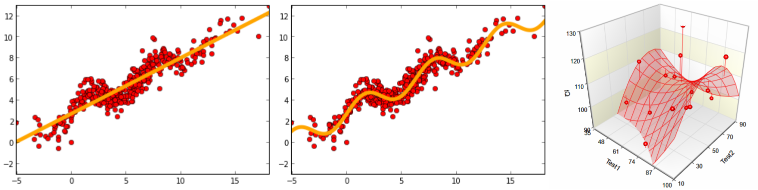
# Logistische Regression

Prof. Dr. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

21. März 2016

# Regression

- Gegeben: Trainingsamples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ mit zugehörigen Labels $y_1, \ldots, y_n$
- **Klassifikation**
  - Labels drücken Klassen-Zugehörigkeit aus
  - Lerne eine **Klassifikator**-Funktion $\mathcal{X} \to \{1, \ldots, C\}$, die den Samples Klassen zuordnet.
- **Regression**
  - Labels sind **reellwertig**!
  - Lerne eine Funktion $f : \mathcal{X} \to \mathbb{R}$, die den Samples reelle Werte zuordnet
- **Beispiele** *(incl. dem "Klassiker": Lineare Regression)*



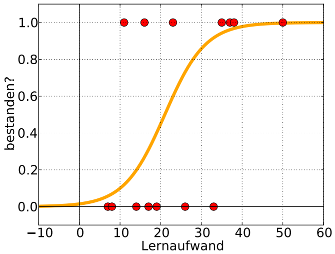
# Logistische Regression: Ansatz

- Kern-Idee der logistischen Regression: Verwende ein **Regressions-Modell zur Klassifikation**.
- Berechne für jede Klasse einen **Score** mittels Regression.
- Der Klassifikator entscheidet sich für die Klasse mit **maximalem Score**.
- Logistische Regression ist ein weit verbreiteter Ansatz in der statistischen Datenanalyse.
- Logistische Regression ist auch bekannt unter dem Namen **Maximum Entropy**.

# Logistische Regression: Ansatz

- Annahme: 2 Klassen namens 0 und 1
  *(Erfolg/Misserfolg, krank/gesund, ...)*
- Falls $> 2$ Klassen: Zerlege das Problem in viele binäre Klassifikationsprobleme, entweder **one-vs-one** ($\frac{n(n-1)}{2}$ Stück) oder **one-vs-all** ($n$ Stück).
- **Gegeben**: Ein Test-Sample $\mathbf{x}$.
- Wir ermitteln per Regression die Wahrscheinlichkeit $P(Y = 1|\mathbf{x})$.
- Wir entscheiden uns für die Klasse, für die diese Wahrscheinlichkeit maximal ist.

## Beispiel (Klausur)

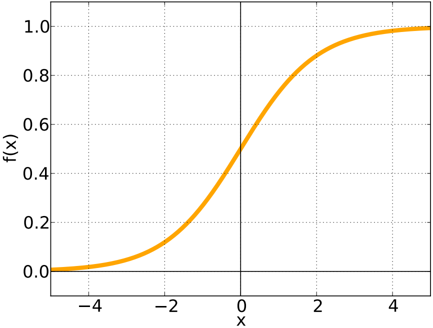


# Logistische Regression: Grundmodell

- Wir verwenden für die Regressionsfunktion als **Grundmodell** eine sogenannte **Sigmoid**-Funktion

$$P(C = 1|x) := f(x) = \frac{1}{1 + e^{-x}}$$

- Es gilt: $lim_{x \to -\infty} f(x) = 0$ und $lim_{x \to \infty} f(x) = 1$
- Es gilt: $P(C = 1|x = 0) = f(0) = 0.5$. Wir entscheiden uns also für Klasse 1, falls $x \geq 0$.



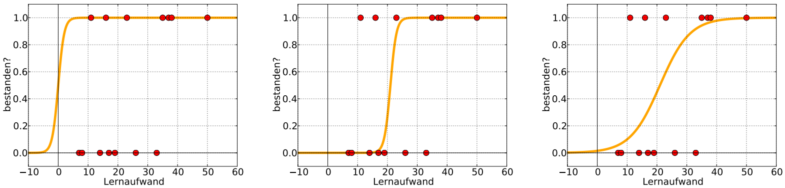
# Logistische Regression: Grundmodell

## Erweiterung

- Wir erlauben eine Verschiebung und Streckung/Stauchung/Spiegelung der Funktion!
- Wir erhalten als Modell:

$$f(x; w_0, w) = \frac{1}{1 + e^{-(w_0 + w \cdot x)}}$$

- Die Parameter $w_0, w$ werden mittels Lernen ermittelt (gleich)



# Logistische Regression: Anmerkungen

- **Warum dieses Modell?**
  - Tradition
  - Einfachheit
  - Wenige Parameter zu fitten $\rightarrow$ Gute Ergebnisse bereits bei wenigen Trainingssamples
  - Korrekt für normalverteilte Daten
- **Lineare Regression würde nicht funktionieren, denn...**
  - Die zu bestimmende Variable (die Klasse $Y$) ist binär
  - Wir können durch die Daten keine sinnvolle Gerade fitten
  - Die prognostizierten Wahrscheinlichkeiten lägen nicht zwischen 0 und 1.

# Logistische Regression im Mehrdimensionalen

- Wie funktioniert das Modell für multivariate Samples $\mathbf{x} \in \mathbb{R}^d$?
- Wir erweitern die Sigmoid-Funktion:

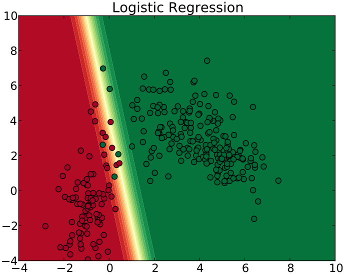
$$f(\mathbf{x}; w_0, w_1, w_2 ..., w_d) = \frac{1}{1 + e^{-(w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + ... + w_d \cdot x_d)}}$$

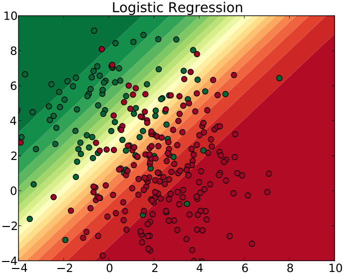
oder kurz (mit Vektor $\boldsymbol{w} := (w_1, ..., w_d)$):

$$f(\mathbf{x}; w_0, \boldsymbol{w}) = \frac{1}{1 + e^{-(w_0 + \mathbf{x} \cdot \boldsymbol{w})}}$$

- Die Entscheidungsgrenze dieses Modells liegt bei $\boldsymbol{x} \cdot \boldsymbol{w} + w_0 = 0$. Dies entspricht einer **Hyperebene** (in Normalenform)!

# Logistische Regression: Illustration




- Die Entscheidungsgrenze ist linear. Wir sprechen bei logistischer Regression deshalb auch von einem **linearen Klassifikator** *(es gibt noch einige weitere!)*.
- Der Parameter $\boldsymbol{w}$ bestimmt die Orientierung der Entscheidungsgrenze, $w_0$ verschiebt die Grenze.
- $\boldsymbol{w}$ bestimmt darüber hinaus die Glätte der Entscheidungsfunktion $f$.

# Logistische Regression: Training

## Schlüsselfrage: Training

- Gegeben ist eine Trainingsmenge von Samples $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ mit Labels $y_1, ..., y_n \in \{0, 1\}$
- **Ziel**: Bestimme $w_0$ und $\mathbf{w}$, also die Position der Entscheidungsgrenze und Glätte der Entscheidungsfunktion.

## Ansatz: Maximum-likelihood-Schätzung

- **Grundidee**: Wähle die Parameter so, dass die beobachten Daten "maximal wahrscheinlich" werden
- Für positive Samples $(y_i = 1)$ sollte $f(\mathbf{x}_i)$ möglichst groß sein:

$$f(\mathbf{x}_i) \approx 1$$

- Für negative Samples $(y_i = 1)$ sollte $f(\mathbf{x}_i)$ möglichst klein sein:

$$f(\mathbf{x}_i) \approx 0$$

# Logistische Regression: Ansatz

## Maximum-Likelihood-Schätzung

Wir formulieren eine sogenannte Likelihood-Funktion und stellen ein Maximierungsproblem auf:

$$w_0^*, \mathbf{w}^* = \arg\max_{w_0, \mathbf{w}} \underbrace{\prod_{i:y_1=1} f(\mathbf{x}_i) \cdot \prod_{i:y_i=0} (1 - f(\mathbf{x}_i))}_{\text{"Likelihood-Funktion"}\ L(w_0, \mathbf{w})}$$

Wir formen das Maximierungsproblem um:

$$\begin{aligned} w_0^*, \mathbf{w}^* &= \arg\max_{w_0, \mathbf{w}} \prod_{i:y_1=1} f(\mathbf{x}_i) \cdot \prod_{i:y_i=0} (1 - f(\mathbf{x}_i)) \\ &= \arg\max_{w_0, \mathbf{w}} \prod_{i} f(\mathbf{x}_i)^{y_i} \cdot (1 - f(\mathbf{x}_i))^{1-y_i} \qquad // \text{ log} \\ &= \arg\max_{w_0, \mathbf{w}} \sum_{i} y_i \cdot log(f(\mathbf{x}_i)) + (1 - y_i) \cdot log(1 - f(\mathbf{x}_i)) \end{aligned}$$

# Logistische Regression: Ansatz

$$w_0^*, \mathbf{w}^* = \arg\max_{w_0,\mathbf{w}} \underbrace{\prod_{i:y_1=1} f(\mathbf{x}_i) \cdot \prod_{i:y_i=0} (1 - f(\mathbf{x}_i))}_{\text{"Likelihood-Funktion"}\ L(w_0,\mathbf{w})}$$

$$= \arg\max_{w_0,\mathbf{w}} \prod_i f(\mathbf{x}_i)^{y_i} \cdot (1 - f(\mathbf{x}_i))^{1-y_i} \qquad // \log$$

$$= \arg\max_{w_0,\mathbf{w}} \sum_i y_i \cdot log(f(\mathbf{x}_i)) + (1 - y_i) \cdot log(1 - f(\mathbf{x}_i))$$

$$= \arg\max_{w_0,\mathbf{w}} \sum_i log(1 - f(\mathbf{x}_i)) + y_i \cdot log\left(\frac{f(\mathbf{x}_i)}{1 - f(\mathbf{x}_i)}\right)$$

$$= \arg\max_{w_0,\mathbf{w}} \sum_i log\left(\frac{\cancel{1} + exp(-(w_0 + \mathbf{x}_i\mathbf{w})) - \cancel{1}}{1 + exp(-(w_0 + \mathbf{x}_i\mathbf{w}))}\right) + y_i \cdot log\left(\frac{\frac{1}{\cancel{(1+exp(-(w_0+\mathbf{x}_i\mathbf{w})))}}}{\frac{exp(-(w_0+\mathbf{x}_i\mathbf{w}))}{\cancel{(1+exp(-(w_0+\mathbf{x}_i\mathbf{w})))}}}\right)$$

$$= \arg\max_{w_0,\mathbf{w}} \sum_i -log\left(\frac{1 + exp(-(w_0 + \mathbf{x}_i\mathbf{w}))}{exp(-(w_0 + \mathbf{x}_i\mathbf{w}))}\right) - y_i \cdot \cancel{log}\left(\cancel{exp}(-(w_0 + \mathbf{x}_i\mathbf{w}))\right)$$

$$= \arg\max_{w_0,\mathbf{w}} \sum_i -log\left(e^{-(w_0+\mathbf{x}_i\mathbf{w})} + 1\right) + \sum_i y_i \cdot (w_0 + \mathbf{x}_i\mathbf{w})$$

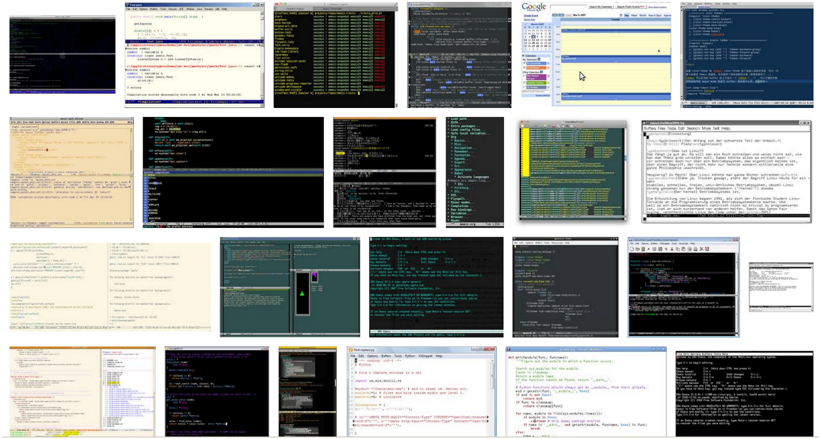
# Logistische Regression: Ansatz

$$\arg\max_{w_0,\mathbf{w}} \underbrace{\sum_i -log\left(e^{-(w_0+\mathbf{x}_i\mathbf{w})}+1\right)+\sum_i y_i\cdot(w_0+\mathbf{x}_i\mathbf{w})}_{\text{"Log-Likelihood-Funktion"}\ l(w_0,\mathbf{w})}$$

## Anmerkungen

- Diese Log-Likelihood-Funktion ist nicht analytisch minimierbar. Es gibt aber **numerische Lösungen**: Finde z.B. Nullstellen der Ableitung mit Hilfe des **Newton-Verfahrens**.
- Die Gewichte in $\mathbf{w}$ zeigen die **Bedeutung** der einzelnen Merkmale für das Klassifikationsproblem an.
- Häufig **regularisieren** wir das Problem, so dass möglichst viele Einträge in $\mathbf{w} = 0$ sind. Dies entspricht einer **Feature Selection**!
- Beispiel: $$\arg\max_{w_0,\mathbf{w}} l(w_0,\mathbf{w}) + C\cdot||\mathbf{w}||_1$$

# Logistische Regression: Code-Beispiel



# Logistische Regression: Fazit

- Logistische Regression ist ein sehr **einfaches** Modell.
- Nur lineare Entscheidungsgrenzen sind modellierbar!
- **Faustregel**: ca. 10 Trainingssamples pro Klasse pro Eingangsvariable