

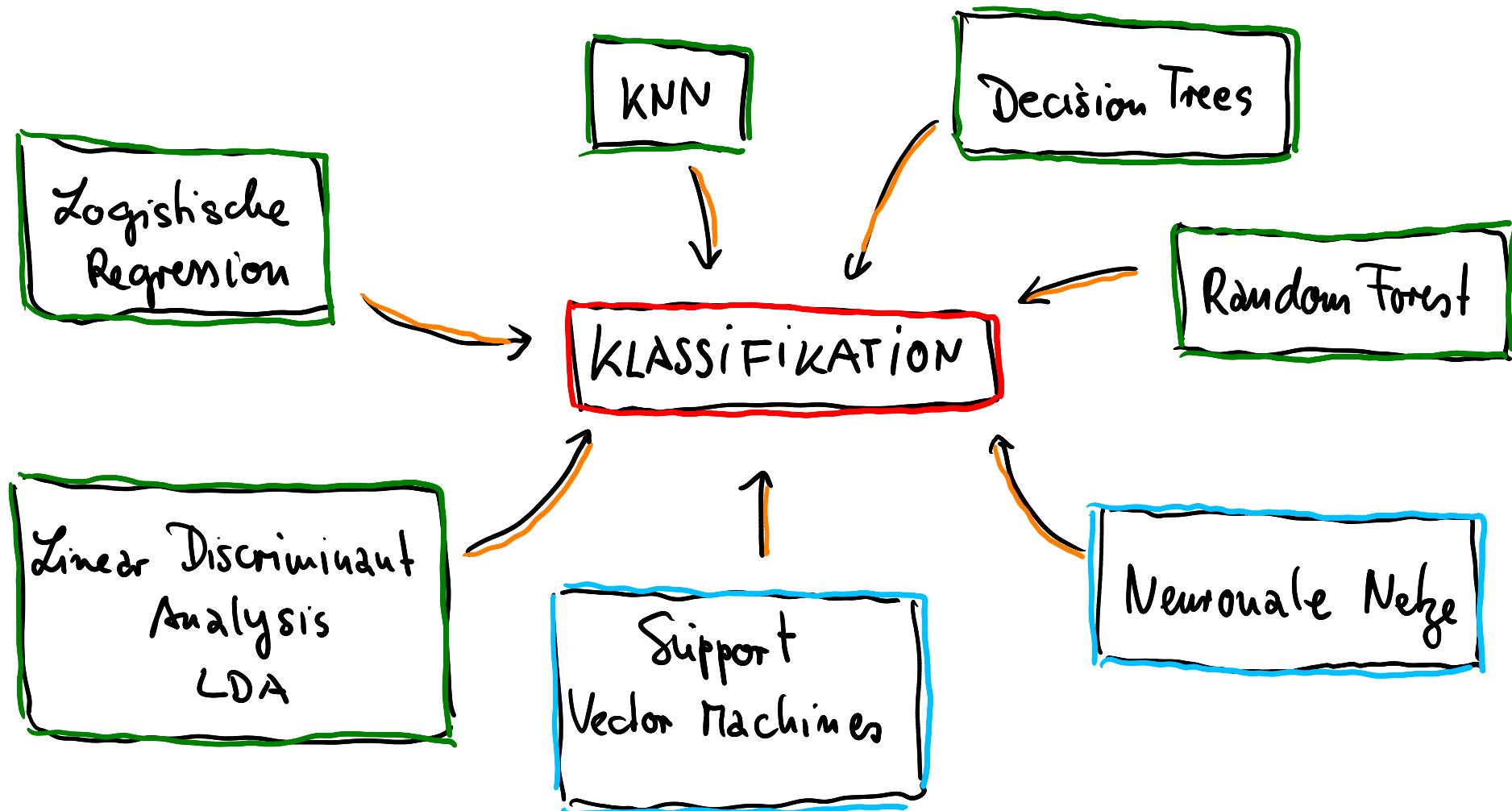
2.2) Klassifikation

Im Unterschied zu Regressionsaufgaben ist die Zielgröße Y bei der Klassifikation kategorisch.

Jede Ausprägung der Zielgröße Y definiert eine Klasse, in die die Beobachtungen eingeteilt werden sollen.

- Bsp.:
- teile Patienten aufgrund verschiedener Eigenschaften in solche mit und ohne erhöhtes Risiko für eine bestimmte Erkrankung ein
 - entscheide aufgrund von Kundendaten über die Kreditwürdigkeit
 - analysiere den Sound einer Maschine um zu entscheiden, ob ein Fehler vorliegt oder nicht
 - ...

Es gibt viele Verfahren, die sich für Klassifikation gut eignen:

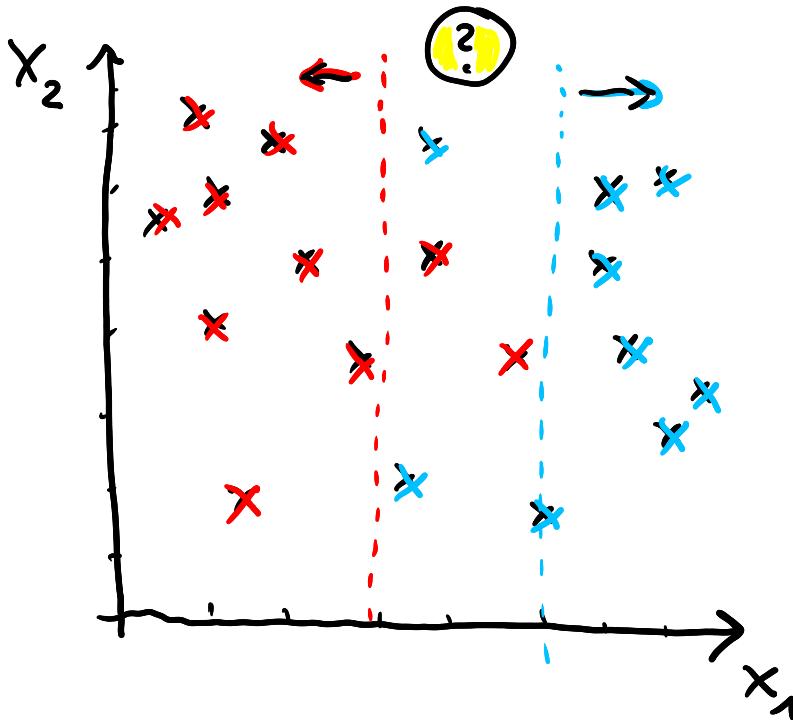


Daten: $(x_1, y_1), \dots, (x_n, y_n)$ ($i = 1, \dots, n$)

x_i ist dabei der Vektor mit den i -ten Beobachtungen der Features X_1, \dots, X_p

Bsp.: 2 Features X_1, X_2
2 Klassen \times und \circ

Es gibt offensichtlich →
Wertebereiche für X_1, X_2 , in
denen nur blaue bzw. nur rote
Punkte zu finden sind!



Zusätzliche Schwierigkeit:

- Was bei sehr vielen Klassen?
- nicht-lineare Trennlinien

ABER: mit $\text{if } \dots \times \text{ else } \circ$ ist es nicht
getan. Es gibt ja auch Bereiche
mit sowohl blauen als auch roten Punkten!

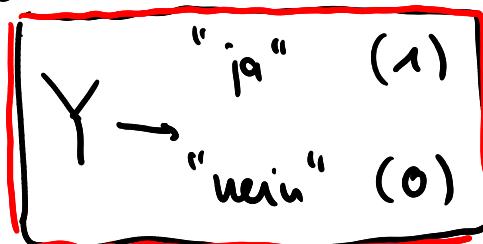
➡ Benchmark:

- Logistische Regression
- LDA
- KNN

2.2.1) Logistische Regression

119

In vielen Fällen hat die Zielgröße Y nur 2 Ausprägungen



Die Wahrscheinlichkeitsverteilung von Y ist einfach zu beschreiben (Bernoulli-Variable)

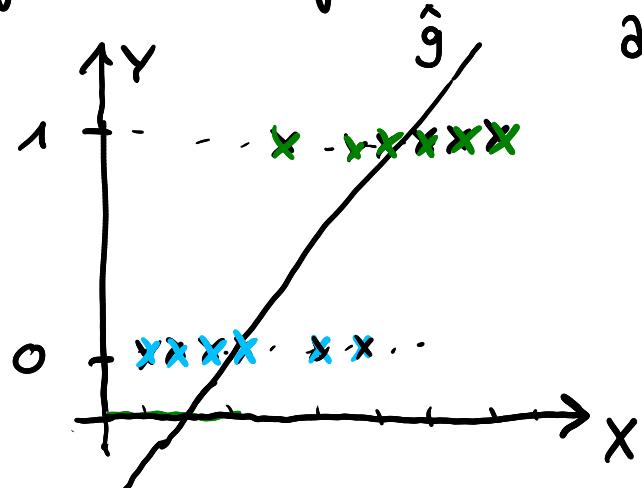
☞ kennt man $P(Y = \text{"ja"}) = p$, dann folgt sofort $P(Y = \text{"nein"}) = 1-p$

Damit ist die Zufallsgröße Y dann komplett beschrieben!

Ziel der logistischen Regression:

Modelliere die Wahrscheinlichkeit für $Y=1$ auf Basis der Beobachtungen von X .

Überlegung:



← könnten wir ja theoretisch machen, ist aber keine wirtschaftliche Idee

$\beta_0 + \beta_1 X + \varepsilon$ beschreibt keine Wahrscheinlichkeit.
→ z.B. Werte <0 , >1 möglich ↴ ↴

Beispiele:BelgrößeFeatures (erklärende Größen)

$Y \hat{=} \text{"Student zufrieden"}$	$\rightarrow ja = 1$ $\rightarrow nein = 0$	- workload - Inhalt - Engagement des Dozenten :
$Y \hat{=} \text{"Defekt liegt vor"}$	$\rightarrow ja = 1$ $\rightarrow nein = 0$	- Rep. Werte einzelner Sensoren - Laufzeit - Nutzungsmodus
$Y \hat{=} \text{"Kunde kreditwürdig"}$	$\rightarrow ja = 1$ $\rightarrow nein = 0$	- kontostand - Einkommen - Schufa

ZEL

Modelliere den Erwartungswert $E(Y)$ (Y diskret)

$$E(Y) = P(Y=0) \cdot 0 + P(Y=1) \cdot 1$$

unter Bedingung zu X :

$$P(Y=1 | X=x)$$

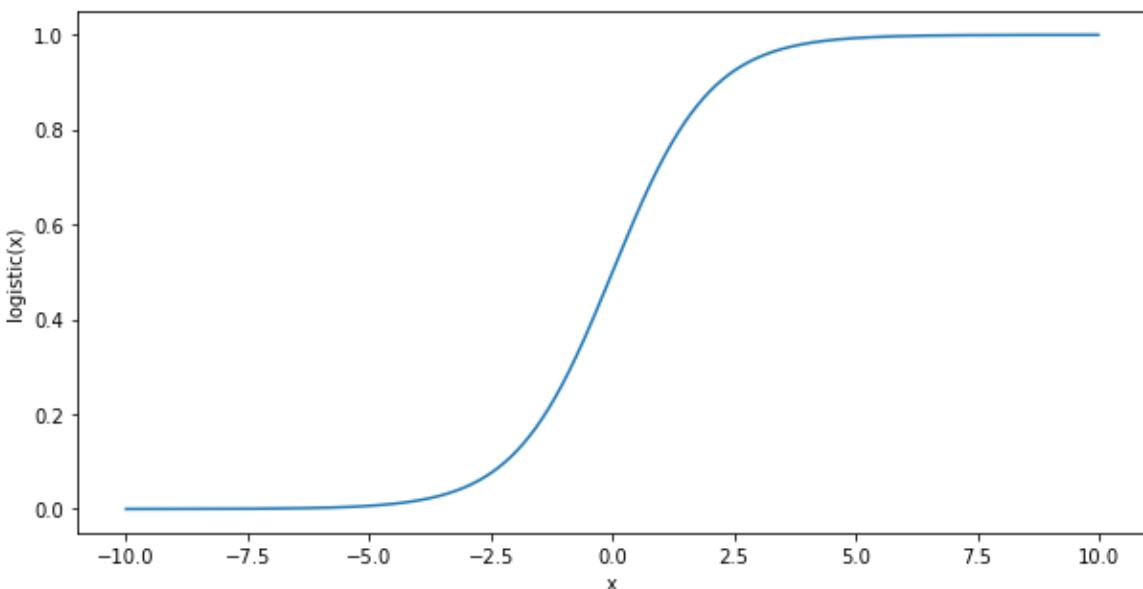


IDEE

121

Wir brauchen eine Transferfunktion (Modell) für die Sicher gestellt ist, dass die modellierten Werte als Wahrscheinlichkeit interpretiert werden können!

Funktionswerte im $[0; 1] \subset \mathbb{R}$



$$\text{logistic}(x) = \frac{e^x}{1+e^x}$$

↳ logistische Funktion

... muss noch angepasst werden



Approximiere

$$P(Y=1 | X=x)$$



optimal: Bayes-Classifier ...

aber $F(Y)$ unbekannt!

Anpassung:



Benutze die logistische Funktion um den Output von $\beta_0 + \beta_1 X$ für alle Realisierungen x von X in den Wertebereich $[0;1]$ zu transformieren

→ Def.: Logistische Funktion

Die logistische Funktion $p(x)$ ist definiert als

Ihr Wertebereich ist



$$p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

"
P(Y=y; | X=x)

$$\frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 X}$$

heißt odds.

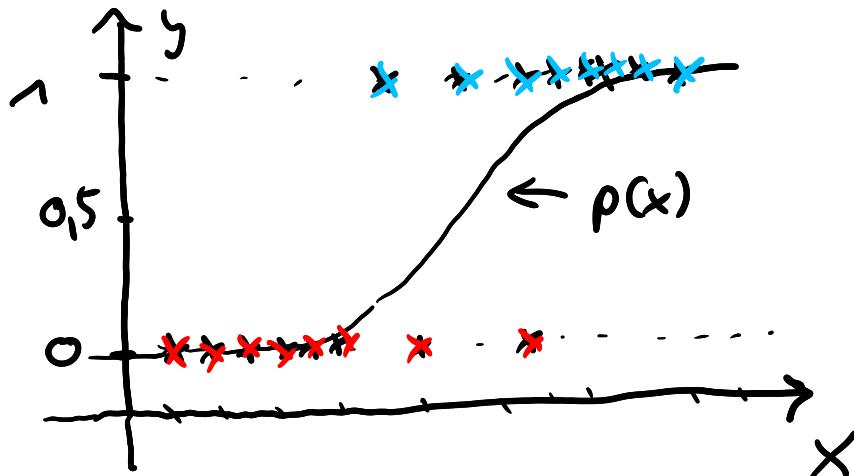
Logarithmieren liefert den logit (log-odds):

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X$$

→ Wie im Beispiel der linearen Einfachregression haben wir 2 Parameter β_0 und β_1 , die aus den Daten geschätzt werden müssen.

Diesmal ist aber die Transferfunktion komplexer

⇒ kQ (nicht-linear) ist möglich, aber sehr schwierig



Man verwendet also für die Schätzung ein Maximum Likelihood-Verfahren



Wir suchen $\hat{\beta}_0, \hat{\beta}_1$, so dass die prognostizierte Wahrscheinlichkeit $\hat{p}(x)$ "maximal gut" zum zugehörigen beobachteten Zustand von Y (0 oder 1) passt.

⇒ optimale $\hat{\beta}_0, \hat{\beta}_1$ sind so gewählt, dass die Wahrscheinlichkeit der beobachteten Zustände von Y unter den Bedingungen X maximal wird!

Def.: Likelihood-Funktion

für (x_i, y_i) ($i=1, \dots, n$) heißt

$$\ell(\beta_0, \beta_1) = \prod_{i: y_i=1} p(x_i) \cdot \prod_{j: y_j=0} (1-p(x_j))$$

Likelihood-Funktion

$\Rightarrow \ell(\beta_0, \beta_1)$ beschreibt die Wahrscheinlichkeit dafür genau die Daten (x_i, y_i) ($i=1, \dots, n$) zu beobachten, die man beobachtet hat.

\Rightarrow optimale $\hat{\beta}_0, \hat{\beta}_1$ so, dass $\ell(\beta_0, \beta_1)$ maximal wird $\Rightarrow \text{grad}(\ell(\beta_0, \beta_1)) \equiv \vec{0}$

ABER: Das Ableiten ist kompliziert (Produkt \Rightarrow Produktregel) 



$\log(\ell(\beta_0, \beta_1))$ ist maximal $\Leftrightarrow \ell(\beta_0, \beta_1)$ maximal

$$\log(\ell(\beta_0, \beta_1)) = \sum_{i: y_i=1} p(x_i) + \sum_{j: y_j=0} (1-p(x_j))$$

\leftarrow viel leichter ableiten, weil Summe !! 

LOG-LIKELIHOOD FUNKTION

Bemerkung: Die Konzepte zur Bewertung der Qualität der Parameterschätzung sind die gleichen wie bei der linearen Regression:

- Standardfehler von $\hat{\beta}_0, \hat{\beta}_1$
- Konfidenzintervalle der Parameter
- ...

Überlegungen zum Hypothesentest:



Welches Hypothesenpaar muss man aufstellen um zu beweisen, dass X geeignet ist um Y zu erklären?

betrachte

$$P(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

← Y hängt dann nicht von X ab, wenn $\beta_1 = 0$ gilt!

⇒ Teste

zum Niveau
($1-\alpha$)

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned}$$

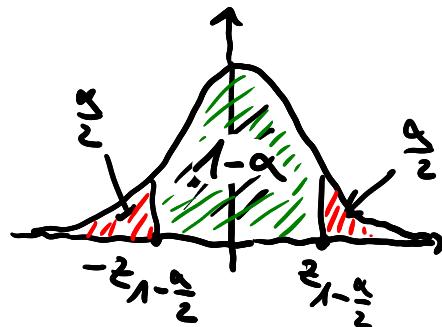
Teststatistik
(Z-Statistik)

$$\frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}} = T$$

← liegt $\hat{\beta}_1$ weit
genug von Null
weg vom H_0
abzulehnen ??

kritischer Bereich:

$$K = (-\infty; -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, +\infty)$$



wobei $z_{1-\frac{\alpha}{2}}$ die $(1 - \frac{\alpha}{2})$ -Quantile der Standardnormalverteilung $\sim N(0; 1)$ sind

INTERPRETATION Ist $\hat{\beta}_1$ weit genug von 0 weg, dann wird die Teststatistik

$T = \frac{\text{Abstand}}{\text{Standardabweichung von } \hat{\beta}_1}$ groß $\Rightarrow T$ landet im kritischen Bereich K

⇒ H_0 wird abgelehnt $\Rightarrow H_1$ gilt als bewiesen \Rightarrow Einfluss von X auf Y !

→ große Teststatistik und kleiner p-Wert sprechen für einen signifikanten Zusammenhang!

Bemerkung:

Das Logit-Modell lässt sich auch für mehrere Einflussgrößen X_1, \dots, X_p formulieren:

 Def.: Multivariates Logit-Modell

Seien X_1, \dots, X_p Prädiktoren der binären Größe Y . Dann ist das multivariate Logit-Modell definiert als

$$P(X_1, \dots, X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}} = P(Y=1 | X_1, \dots, X_p)$$

mit Modellparametern $\beta_0, \beta_1, \dots, \beta_p$

Prognose?

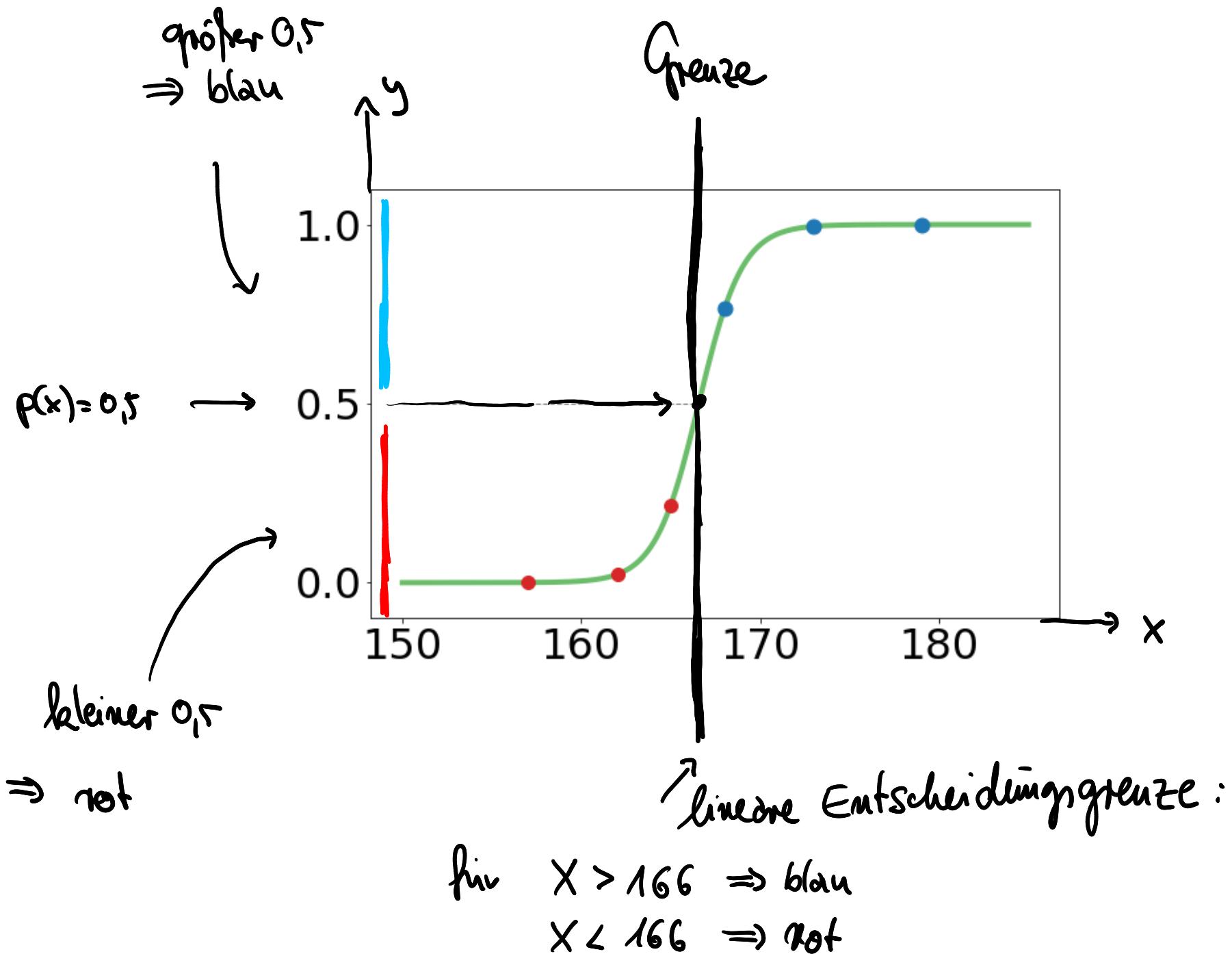
Für geschätzte Parameter $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ und eine Realisierung (x_{i1}, \dots, x_{ip}) von X_1, \dots, X_p erhält man eine Prognose durch:

$$\hat{P}(y_i = 1 \mid X_1, \dots, X_p = x_{i1}, \dots, x_{ip}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}}}$$

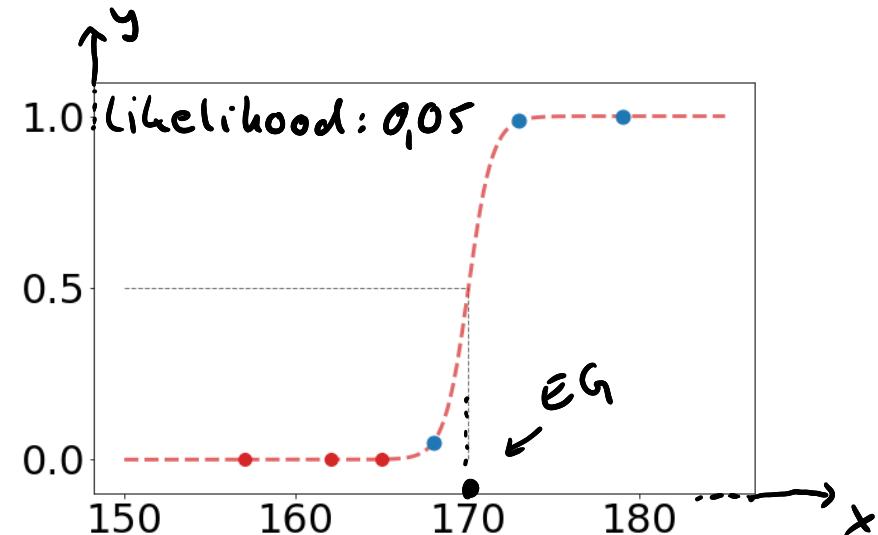
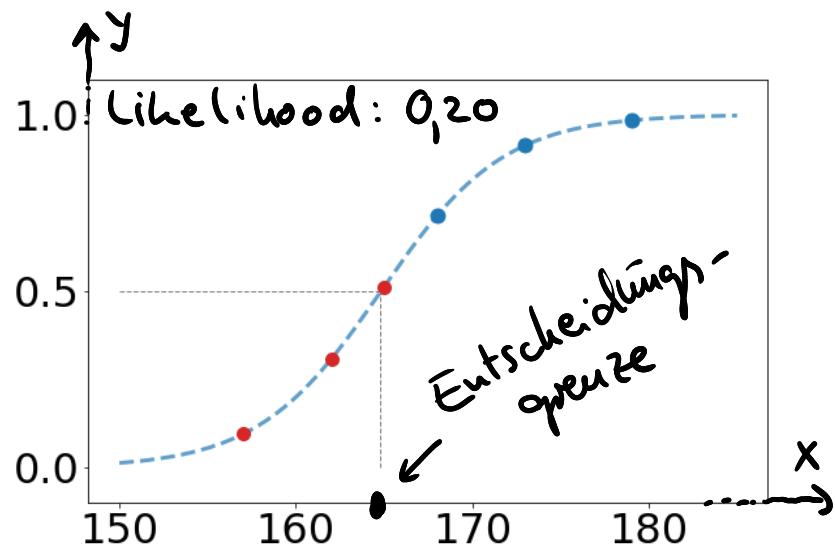
prognostizierte Wahrscheinlichkeit
für $y_i = 1$



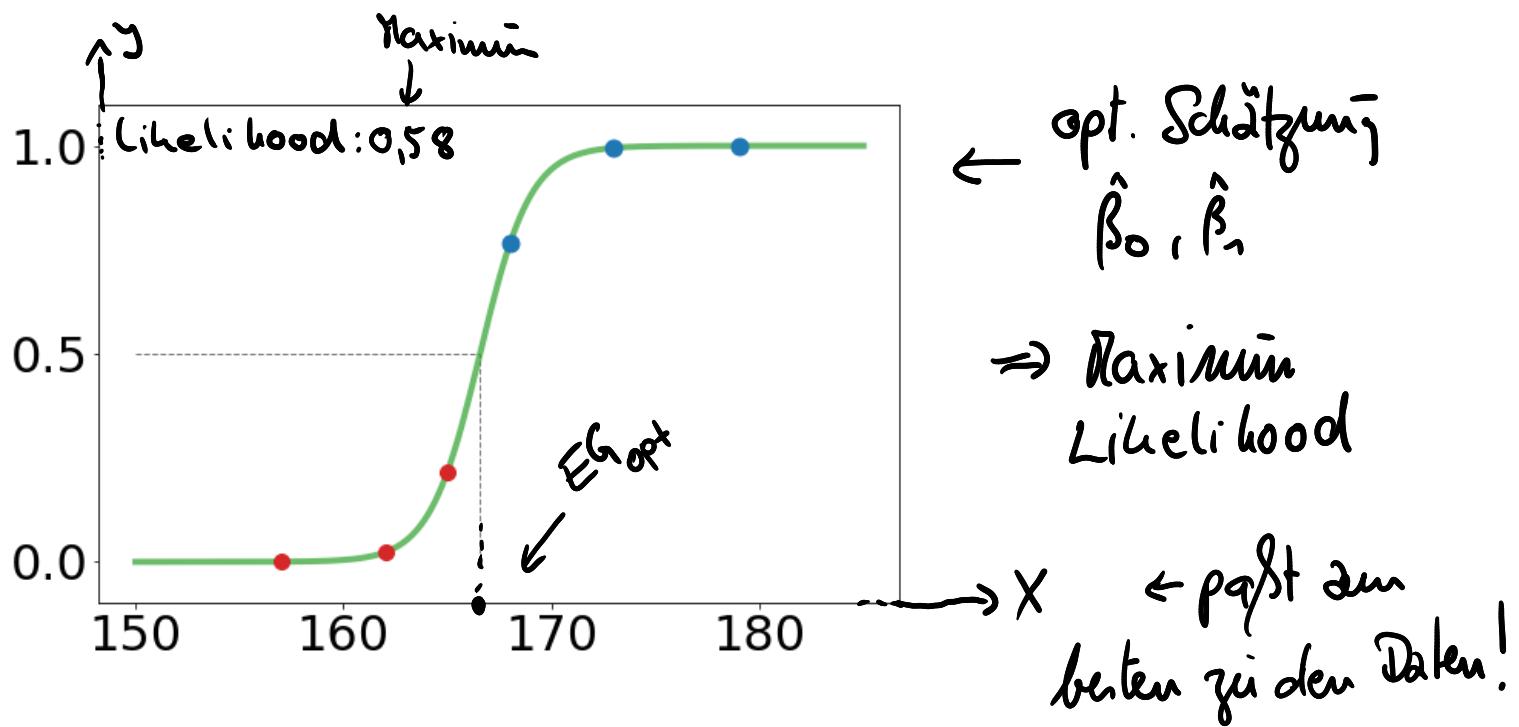
Wie sehen die "decision boundaries" aus?

Bsp.:

→ verschiedene Schätzungen für β_0, β_1



→ die optimale
Schätzung $\hat{\beta}_0, \hat{\beta}_1$
bestimmt die Struktur
der Fit-Funktion und
damit die Entschei-
dungsgrenze



Bemerkung:

- Das Maximum der Log-likelihood Funktion

$$\log(l(\beta_0, \beta_1)) = \sum_{i: y_i=1} p(x_i) + \sum_{j: y_j=0} (1-p(x_j))$$

LOG-LIKELIHOOD FUNKTION

zu bestimmen mit Hilfe der Bedingung

$$\text{grad}(\log(l(\beta_0, \beta_1))) = \vec{0}$$

ist leider analytisch nicht möglich \Rightarrow numerische Lösung (Computer)

- Die logistische Regression lässt sich auf Zielgrößen Y mit mehreren Ausprägungen erweitern. Das wird aber selten benutzt.
 → dafür gibt es geeignete Verfahren ... ☺