GREENBOOTCAMPS

KLASSIFIKATION & LOGISTISCHE REGRESSION

Klassifikaiton

Regression und Klassifikation sind die beiden wichtigsten Verfahren beim überwachten Lernen

Liegt eine **(meist binäre) kategoriale Zielgröße** vor, handelt es sich um Klassifikation. Neue Objekte sollen einer von endlich vielen verschiedenen Klassen zugeordnet werden. Die Kategorien der Zielgröße stellen diese Klassen der. Manche Verfahren kennzeichnen Objekte direkt mit einer Klasse, andere prognostizieren als Ausgabe eine Zahl, die als Wahrscheinlichkeit zu einer Klasse zu gehören interpretiert werden kann. Ein Verfahren, welches eine solche Wahrscheinlichkeitsabschätzung der Klassenzugehörigkeit vornimmt, ist bspw. die *logistische Regression**. Ein solches Verfahren wird weiterhin als Klassifikation eingestuft, denn die Zielgröße ist und bleibt kategorial.

*Die Bezeichnung logistische Regression könnte verwirrend sein, da es kein Regressions- sondern Klassifikationsverfahren ist.

Logistische Regression

Ziel: Finde regressionsanalytisch eine Klassifikationsregel, die die Zuordnung zu sich gegenseitig ausschließenden Kategorien ermöglicht.

Problem: Zielvariable ist nicht numerisch, sondern (im einfachsten Fall) dichotom (-> binäre logistische Regression) oder mehrkategorial (-> multinomiale logistische Regression)

(Lineares) Regressionsmodell ist ungeeignet zur Beschreibung des Zusammenhangs und zur Prognose

Lösung: Logistische Regression -> schätzt Wahrscheinlichkeiten zu den jeweiligen Kategorien zu gehören

-> auf Basis der prognostizierten Klassenzugehörigkeitswahrscheinlichkeiten kann dann die Klassifizierung erfolgen

Logistische Regression

Logistische Regression = Klassifikationsverfahren

Bezeichnung „Regression“ unzutreffend, zumindest aber unglücklich

Die logistische Regression schätzt die Wahrscheinlichkeit (numerische Werte) zur Klassenzugehörigkeit einer kategorialen Klasse, d.h. Werte der Zielvariable sind kategorial

Logistische Regression

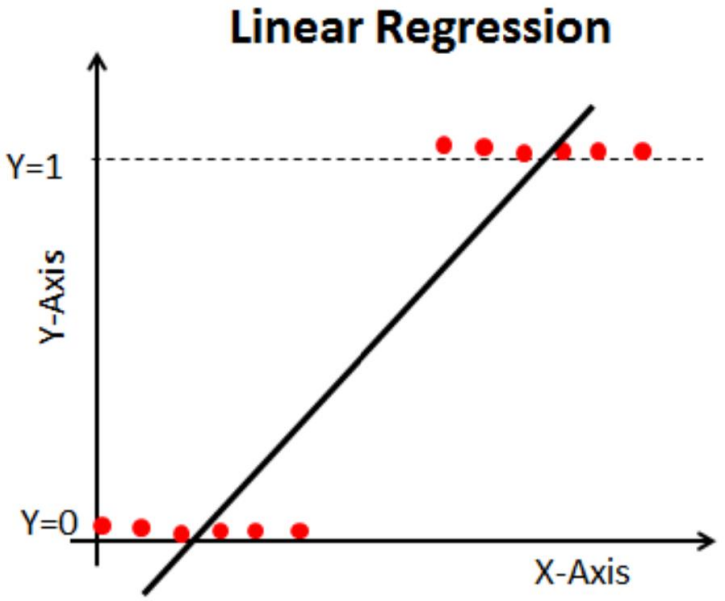
Binäre logistische Regression

Zwei Klassen -> Zielvariable Y: Angabe zur Klassenzugehörigkeit und binär codiert:

$$Y = \begin{cases} 1, wenn\ Klasse\ 1 \\ 0, wenn\ Klasse\ 2 \end{cases}$$

Lineare Regressionsgerade in diese Punkte „hineinzulegen“ ist unpassend.

Grund: Gerade prognostiziert Werte zwischen 0 und 1, kleiner 0 oder auch größer 1



Logistische Regression

Binäre logistische Regression

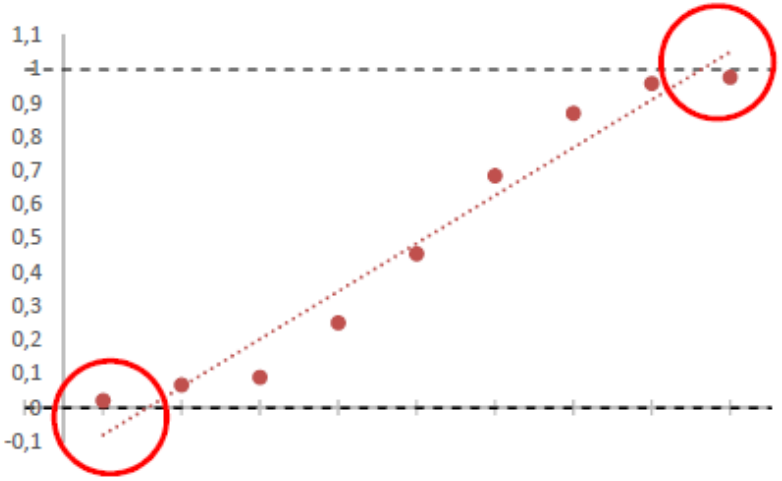
Abhängige Variable: Wahrscheinlichkeit für das Eintreten des Ereignisses Y = 1

➔ Zwischenwerte zwischen 0 und 1 sind möglich

➔ Schätzung mit MQ (minimale Quadrate) -> lineares Wahrscheinlichkeitsmodell

$$W(Y = 1|\boldsymbol{X}) = \pi(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

Nicht ausreichend, da Gerade weiterhin Werte größer 1 oder kleiner 0 liefern würde



Logistische Regression

Binäre logistische Regression

Abhängige Variable: **Odds** als Verhältnis der beiden Wahrscheinlichkeiten

$$W(Y = 1) = \pi \; und \; W(Y = 0) = 1 - \pi$$

$$\boldsymbol{Odds = \frac{\pi}{1 - \pi}}$$

➔ Werte größer 1 möglich, aber immer noch keine Werte unter 0

➔ Weitere Transformation nötig!

Logistische Regression

Binäre logistische Regression

Abhängige Variable: **Logits** (logarithmierte Odds)

$$\boldsymbol{Logit} = \ln(\boldsymbol{Odds}) = \ln\left(\frac{\boldsymbol{\pi}}{\boldsymbol{1-\pi}}\right)$$

➔ Wertebereich: von – bis + unendlich

Binäres logistisches Regressionsmodell (Logit-Modell):

$$Logit = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_m X_m$$

Logistische Regression

Binäre logistische Regression

Schätzmethode: Maximum-Likelihood

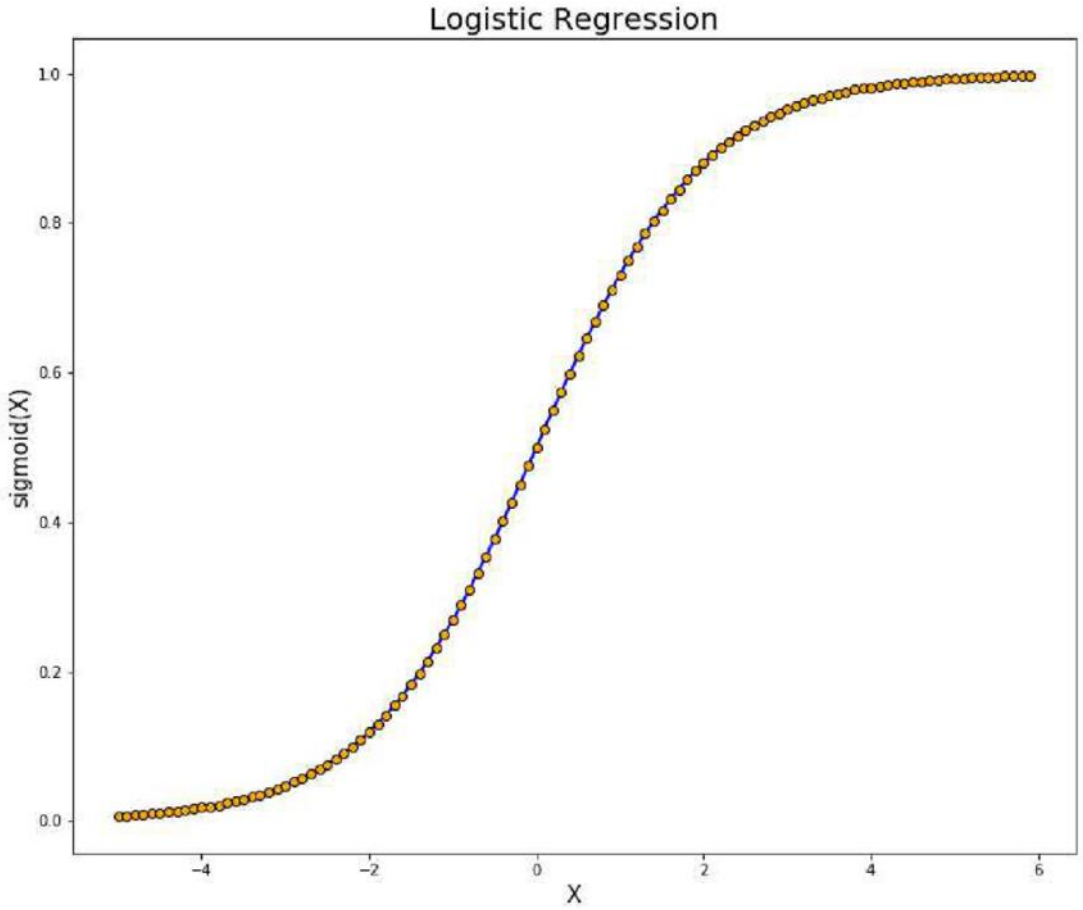
$$Logit = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_m X_m$$

$$Odds = \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots+\beta_m X_m} = e^{\beta_0} * e^{\beta_1 X_1} * e^{\beta_2 X_2} * \cdots * e^{\beta_m X_m}$$

$$\pi = W(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots+\beta_m X_m}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots+\beta_m X_m}}$$

$$= \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots+\beta_m X_m)}}$$

Logistische Regression



Logistische Regression

Binäre logistische Regression

Interpretation ist schwieriger als im linearen Regressionsmodell, da (partieller) Effekt einer unabhängigen Variablen auf abhängige Variable (W(Y=1)) davon abhängt, an welcher Stelle man sich befindet (somit vom Wert der unabhängigen Variablen selbt)

Interpretation der Koeffizienten:

- $\beta_j > 0$ mit steigender Variable Xj steigt die Wahrscheinlichkeit, dass Y=1
- $\beta_j < 0$ mit steigender Variabel Xj sinkt die Wahrscheinlichkeit, dass Y=1

Logistische Regression

Binäre logistische Regression

Zur Schätzung der Parameter der logischen Regression wird aufgrund der Nichtlinearität anstelle der MQ-Methode die Maximum-Likelihood-Methode angewendet

ML-Prinzip: Bestimme die Schätzwerte für die unbekannten Parameter so, dass die realisierten Daten maximale Plausibilität (Likelihood) erlangen

Klassifikationsgütermaßen

Korrektklassifikationsrate:

$$\frac{RP}{Total} + \frac{RN}{Total} = \frac{RP+RN}{Total}$$

Klassifikationsfehler:

$$\frac{FP}{Total} + \frac{FN}{Total} = \frac{FP+FN}{Total}$$

		Beobachtet		
		1 (Positiv)	0 (Negativ)	
Vorhersage	1 (Positiv)	Richtig Positiv (RP)	Falsch Positiv (FP)	RP + FP
	0 (Negativ)	Falsch Negativ (FN)	Richtig Negativ (RN)	FN + RN
		RP + FN	FP + RN	Total

Klassifikationsgütermaßen

(Sensitivität)

Richtig-positiv-rate = $\frac{RP}{RP+FN}$

Falsch-positiv-rate = $\frac{FP}{FP+RN}$

(Spezifität)

Richtig-negativ-rate = $\frac{RN}{RN+FP}$

Falsch-negativ-rate = $\frac{FN}{FN+RP}$

		Beobachtet		
		1 (Positiv)	0 (Negativ)	
Vorhersage	1 (Positiv)	Richtig Positiv (RP)	Falsch Positiv (FP)	RP + FP
	0 (Negativ)	Falsch Negativ (FN)	Richtig Negativ (RN)	FN + RN
		RP + FN	FP + RN	Total

Klassifikationsgütermaßen

		Beobachtet		
		1 (Positiv)	0 (Negativ)	
Vorhersage	1 (Positiv)	Richtig Positiv (RP)	Falsch Positiv (FP)	RP + FP
	0 (Negativ)	Falsch Negativ (FN)	Richtig Negativ (RN)	FN + RN
		RP + FN	FP + RN	Total

Matthews-Korrelationskoeffizient (MCC)

$$MCC = \frac{RP * RN - FP * FN}{\sqrt{(RP + FN) * (FP + RN) * (RP + FP) * (FN + RN)}}$$

|MCC| = 1 – perfekte Klassifikation, |MCC| = 0 – zufällige Zuordnung

Klassifikationsgütermaßen

Prädikative Werte

➔ Positiver prädikativer Wert (Relevanz, Präzision, Genauigkeit)

$$PPV = \frac{RP}{RP + FP}$$

➔ Negativer prädikativer Wert (Segreganz)

$$NPV = \frac{RN}{RN + FN}$$

		Beobachtet		
		1 (Positiv)	0 (Negativ)	
Vorhersage	1 (Positiv)	Richtig Positiv (RP)	Falsch Positiv (FP)	RP + FP
	0 (Negativ)	Falsch Negativ (FN)	Richtig Negativ (RN)	FN + RN
		RP + FN	FP + RN	Total

Klassifikationsgütermaßen

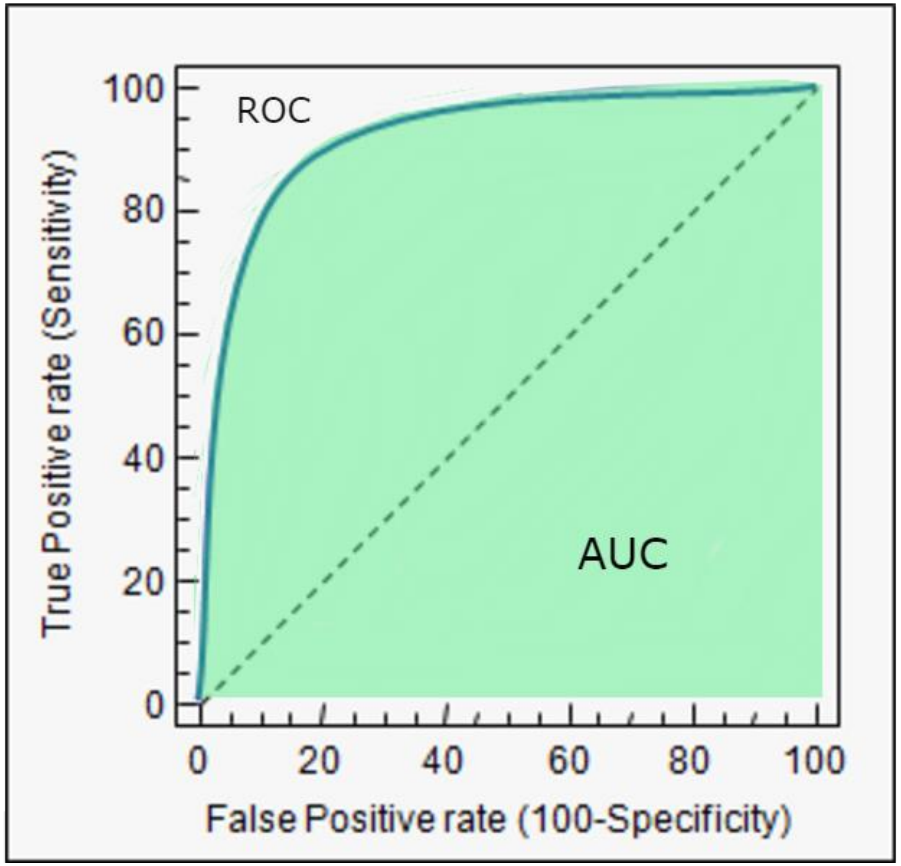
ROC-Kurve und AUC

Besondere Punkte:
(0, 0); (1, 1); (0, 1); (0.5, 0.5)

Richtig-positiv-rate = Sensitivität
Falsch-positiv-rate = 1 – Spezifität

Area **U**nder **C**urve

ROC-Kurve ist eine Treppenfunktion mit möglicherweise sehr kleinen Treppenstufen



Klassifikationsgütermaßen

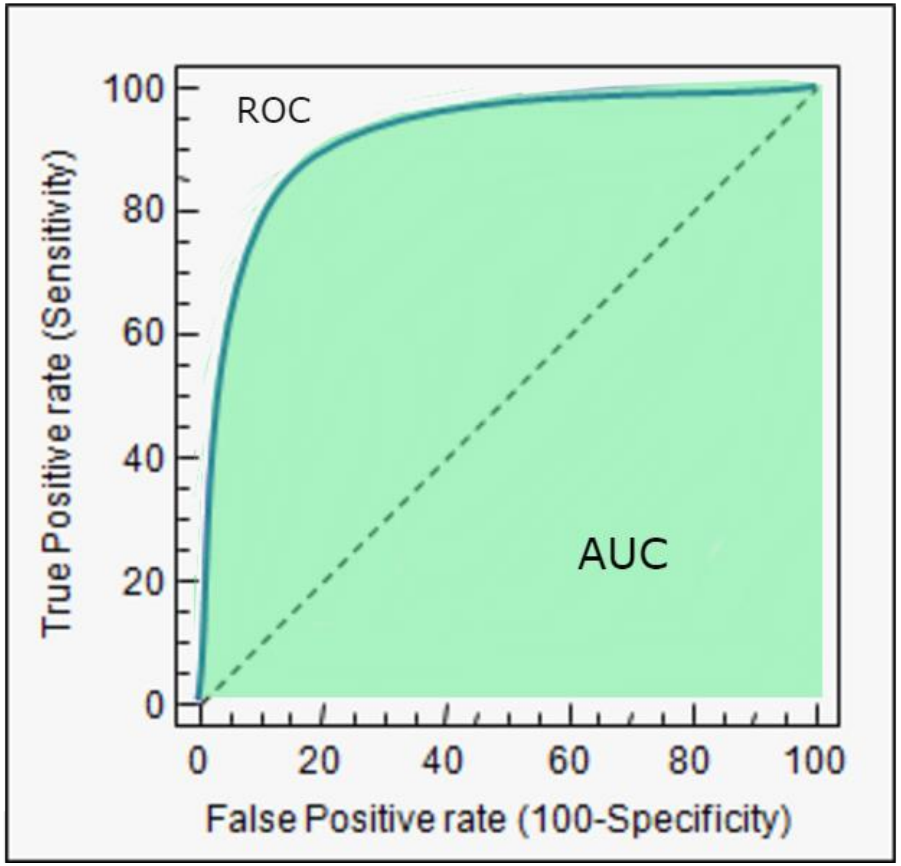
ROC-Kurve und AUC

Wo liegt der theoretisch optimale Schwellenwert?

Rechnerisch: der Schwellenwert mit dem höchsten **Youden-Index**

Youden-Index = Sensitivität + Spezifität – 1

	AUC < 0,7:	ungenügend
0,7 <=	AUC < 0,8:	akzeptabel
0,8 <=	AUC < 0,9:	exzellent
	AUC >=0,9:	außerordentlich



Klassifikationsgütermaßen

F1 Score – harmonisches Mittel kombiniert in sich Präzision und Sänsitivität

$$F1 = 2 * \frac{Präzision \ * Sensitivität}{Präzision + Sensitivität}$$

$$= \frac{RP}{RP + 0{,}5 \ * (FP + FN)}$$

Gütermaße bei Logistischer Regression

Gütermaße basieren auf dem Wert der maximierten Log-Likelihood ln L(b) (weiter **LL**). Dieser wird jedoch (da er immer negativ ist) noch verändert und man betrachtet stattdessen oft den Wert **-2LL,** der immer positiv ist.

Die Größe -2LL kann zum Vergleich verschiedener Modelle (mit gleichem Datensatz) verwendet werden, wobei die Größe an sich nichts aussagt.

-2LL ist vergleichbar mit der Summe der quadrierten Residuen (RSS) aus MQ-Methode.

Gütermaße bei Logistischer Regression

Likelihood-Ratio

$$LR = \frac{L_1}{L_2}$$

LR-Test : $LLR = -2 * ln\left(\frac{L_1}{L_2}\right) = -2 * (LL_1 - LL_2)$

Test auf Gesamtmodell: $LLR = -2 * ln\left(\frac{L_0}{L_v}\right) = -2 * (LL_0 - LL_v)$, Lv – vollständiges Modell, Lo – Nullmodell

Gütermaße bei Logistischer Regression

Pseudo-R^2

- McFadden‘s R $R_{MF}{}^2 = 1 - \left(\frac{LL_v}{LL_0}\right)$ - erreicht Extremwerte 0 und 1 bei realen Datensätzen nicht. Werte 0,2 bis 0,4 deuten auf eine gute Anpassung hin
- Cox & Snell – R $R_{CS}{}^2 = 1 - \left(\frac{L_0}{L_v}\right)^{\frac{2}{n}}$ - Kann nur Werte < 1 annhemen
- Nagelkerke‘s R $R_N{}^2 = \frac{R_{CS}{}^2}{1-L_0{}^{\frac{2}{n}}}$ - angepasstes R_{CS} , sodass auch Wert von 1 erreicht werden kann

Gütermaße bei Logistischer Regression

Devianz:

$$D = -2\ (LL - LL_s)$$

LL – Log Likelihood des zu prüfenden Modells

LLs – Log Likelihood des staturierten Modells (bestmöglichen)