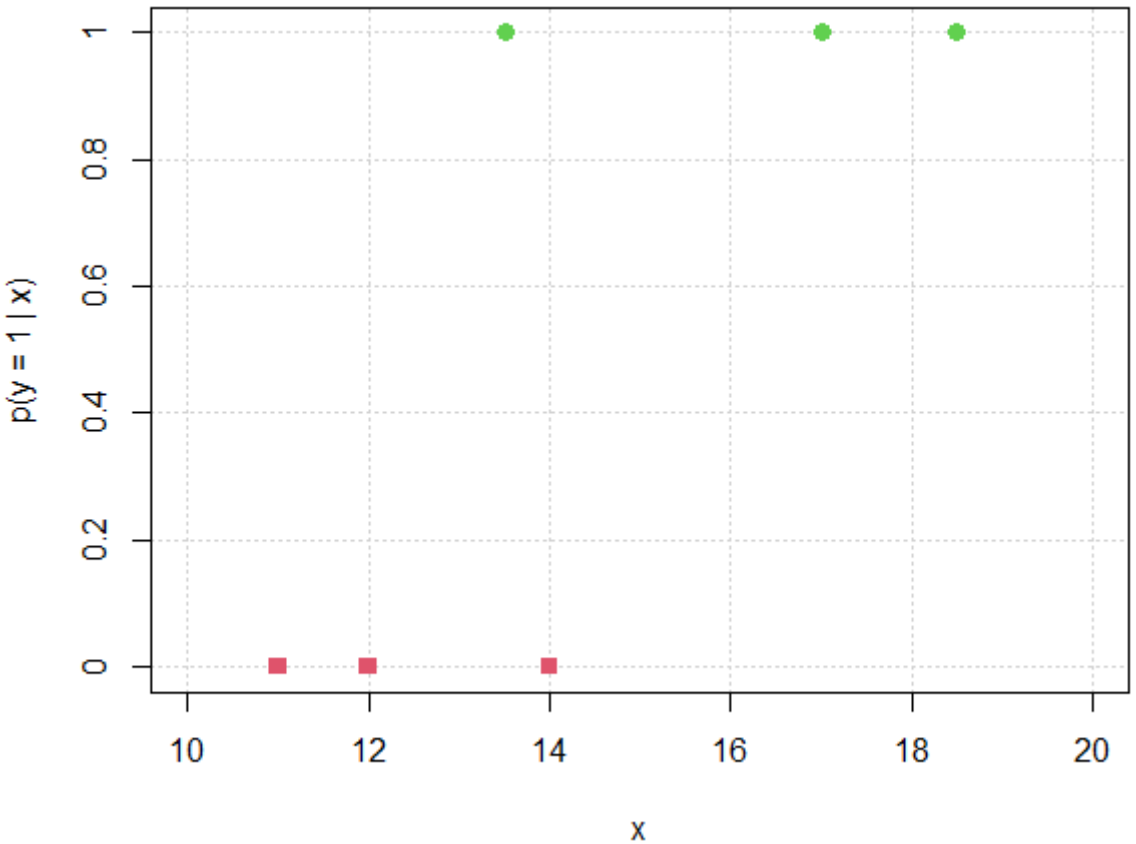
MDS – Logistische Regression

Aufgabe 1

Stellen Sie sich vor, wir haben ein einfaches Klassifikationsproblem mit nur einer Input-Variable x_i und $n = 6$ Beobachtungen:

Beobachtung	y_i	x_i
1	0	11
2	0	12
3	0	14
4	1	13.5
5	1	17
6	1	18.5

Die Daten können wie folgt in einem Streudiagramm visualisiert werden:



Sie wollen nun ein logistisches Regressionsmodell für diese 6 Datenpunkte fitten. Wie Sie aus dem Unterricht wissen, hat das Modell folgende Form:

$$p(y_i = 1|x_i) = \frac{1}{1+e^{-(w_0+w_1 \cdot x_i)}}$$

a) Für das Modell Fitting via *Gradient Descent* wählen wir die Anfangswerte für die Parameter unseres Modells. Die Anfangswerte sind wie folgt: $w_0 = -5$ und $w_1 = 0.5$. Setzen Sie die Parameterwerte in die Modellgleichung (siehe vorherige Seite) ein und rechnen Sie die vorhergesagten Wahrscheinlichkeiten für die 6 Beobachtungen.

Tipp: Verwenden Sie folgende einfache R Funktion, um für gegebene Parameterwerte und einen gegebenen x-Wert die vorhergesagten Wahrscheinlichkeiten zu rechnen.

```
logreg <- function(w0, w1, x) {
      1 / (1 + exp(- (w0 + w1 * x)))
}
```

Beobachtung	y_i	x_i	$p(y_i = 1 \mid x_i)$
1	0	11	
2	0	12	
3	0	14	
4	1	13.5	
5	1	17	
6	1	18.5	

b) Nun wollen wir den Wert der Kostenfunktion für die Parameterwerte $w_0 = -5$ und $w_1 = 0.5$ berechnen. Dazu rechnen Sie am besten zuerst die fehlenden Werte in folgender Tabelle:

Beobachtung	y_i	x_i	$-\log\left(p(y_i = 1 \mid x_i)\right)$	$-\log\left(1 - p(y_i = 1 \mid x_i)\right)$
1	0	11	-	
2	0	12	-	
3	0	14	-	
4	1	13.5		-
5	1	17		-
6	1	18.5		-

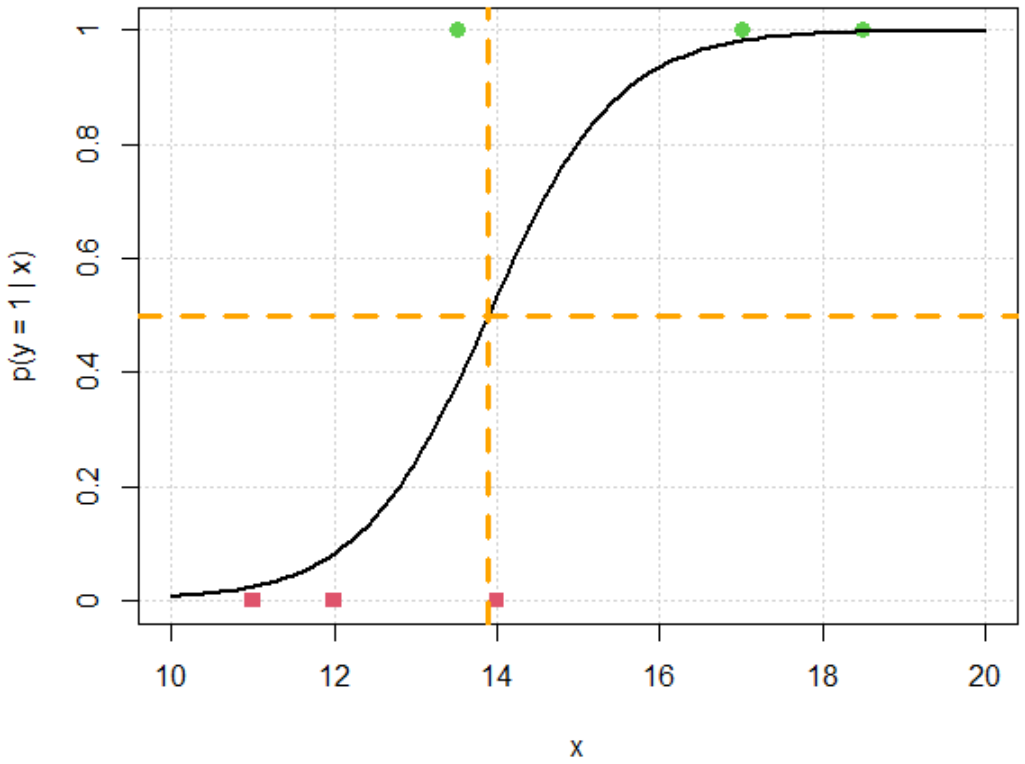
c) Addierend Sie nun die individuellen Kosten und teilen Sie das Resultat durch die Anzahl Beobachtungen. Was sind die Kosten für die Parameterwerte $w_0 = -5$ und $w_1 = 0.5$?

d) Stellen Sie sich vor, Sie haben nun ganz viele *Gradient Descent* Schritte gemacht und die aktuellen Parameterwerte sind $w_0 = -17.64$ und $w_1 = 1.27$. Sie kriegen mit dem aktuellen Modell die vorhergesagten Wahrscheinlichkeiten in folgender Tabelle:

Beob.	y_i	x_i	$p(y_i = 1\|x_i)$	$-\log\left(p(y_i = 1\|x_i)\right)$	$-\log\left(1 - p(y_i = 1\|x_i)\right)$
1	0	11	0.02	-	
2	0	12	0.08	-	
3	0	14	0.53	-	
4	1	13.5	0.38		-
5	1	17	0.98		-
6	1	18.5	1.00		-

Berechnen Sie die individuellen Kosten (fehlende Werte in Tabelle) und rechnen Sie danach die Gesamtkosten dieser Parameterwerte. Vergleichen Sie das Resultat mit den Kosten aus Aufgabe c).

e) Mit den Parameterwerten $w_0 = -17.64$ und $w_1 = 1.27$ haben wir nun also das optimale Modell gefunden. Grafisch sieht das Modell wie folgt aus:



Wie viele Fehler macht das Modell mit einem Threshold von 0.5? Wie viele davon sind False Positives? Wie viele False Negatives?