Übung 3: Klassifikationsmodelle I

Aufgabe 1

Betrachtet wird eine Gruppe Studierender, die an der Klausur „Statistik II“ an der Universität Hamburg teilgenommen haben. Der Trainingsdatensatz

$$\left\{(\mathbf{x}_n, t_n) \in \mathbb{R}^2 \times \{\mathcal{C}_1, \mathcal{C}_2\} : n = 1, \ldots, N\right\},$$

besteht aus den Ausprägungen der beiden Inputvariablen

$$X_1 := \text{„Klausurvorbereitungszeit“} \qquad \text{und} \qquad X_2 := \text{„Abiturdurchschnittsnote“}$$

und der Realisierung der binären Outputvariablen

$$T := \begin{cases} 1 & \text{falls, das Ergebnis der Klausur „Statistik II“ eine 1 ist (Klasse } \mathcal{C}_1) \\ 0 & \text{sonst (Klasse } \mathcal{C}_2) \end{cases}.$$

An den Trainingsdatensatz wird ein Logit-Modell mit den Basisfunktionen $\phi_1(x_1, x_2) = x_1$ und $\phi_2(x_1, x_2) = x_2$ angepasst. Dabei resultieren die Koeffizientenschätzungen $\widehat{w}_0 = -4{,}2$, $\widehat{w}_1 = 0{,}05$ und $\widehat{w}_2 = 1$.

a) Bestimmen Sie eine Schätzung für die Wahrscheinlichkeit, dass ein Studierender mit einer Vorbereitungszeit auf die Klausur von 40 Stunden und einer Abiturdurchschnittsnote von 1,7 in der Klausur „Statistik II“ eine 1 erzielt.

b) Wie viele Stunden Vorbereitungszeit auf die Klausur müsste der Studierende aus Aufgabenteil a) mindestens investieren, damit sich seine Wahrscheinlichkeit eine 1 zu erzielen auf mindestens 50% erhöht?

Aufgabe 2

Die beiden folgenden Fragen beziehen sich auf die Odds bei einem Logit-Modell.

a) Wie groß ist der Anteil der Kreditkartenbesitzer mit einem Odds von 0,37 für einen Zahlungsverzug bei ihrer Kreditkartenabrechnung, der im Durchschnitt in Verzug geraten wird?

b) Ein Kreditkartenbesitzer besitze eine Wahrscheinlichkeit von 16%, dass es bei der Kreditkartenabrechnung zu einem Zahlungsverzug kommt. Wie hoch sind die Odds, dass es zu einem Zahlungsverzug kommt.

Aufgabe 3

An den Trainingsdatensatz

$$\left\{(1,3,0), (2,4,0), (4,1,0), (3,1,1), (4,2,1) \in \mathbb{R}^2 \times \{0,1\}\right\}.$$

mit den beiden Inputs x_1 und x_2 soll ein logistisches Regressionsmodell (Logit-Modell) mit den Basisfunktionen $\phi_1(x_1, x_2) = x_1$ und $\phi_2(x_1, x_2) = x_2$ angepasst werden.

a) Geben Sie die zum Trainingsdatensatz gehörige Log-Likelihoodfunktion $l\big(\mathbf{w}; (\mathbf{x}_n, t_n)_{n=1,\ldots,5}\big)$ an.

b) Ermitteln Sie mit Hilfe des IRLS-Verfahrens (Iterative Reweighted Least Squares) Schätzungen für die Gewichte w_0, w_1 und w_2. Verwenden Sie dabei $\mathbf{w}^{(0)} = (0,0,0)^T$ als Startwert.

c) Berechnen Sie mit Hilfe des Ergebnisses aus Aufgabenteil b) eine Schätzung $\widehat{\pi}(\mathbf{x})$ für die (bedingte) Klassenzugehörigkeitswahrscheinlichkeit $\pi(\mathbf{x}) = \mathbb{P}(T = 1|\mathbf{x})$ und geben Sie den zugehörigen geschätzten Bayes-Klassifikator an.

d) Klassifizieren Sie mit Hilfe des Ergebnisses aus Aufgabenteil c) die neue Beobachtung $\mathbf{x} = (x_1, x_2) = (5,0)$. Ermitteln Sie ferner die Entscheidungsgrenze und veranschaulichen Sie diese zusammen mit den Beobachtungen im Trainingsdatensatz und der neuen Beobachtung in einer geeigneten Abbildung.

Aufgabe 4

Die folgende Tabelle enthält einen Trainingsdatensatz vom Umfang $N = 6$ bestehend aus den Werten einer Outputvariablen T und dreier Inputvariablen X_1, X_2 und X_3:

Beobachtung	X_1	X_2	X_3	T
1	0	3	0	rot
2	2	0	0	rot
3	0	1	3	rot
4	0	1	2	grün
5	−1	0	1	grün
6	1	1	1	rot

Mittels k-Nächste-Nachbarn-Klassifikation soll eine Prognose für den Wert der Outputvariable T bestimmt werden, wenn für die Werte der drei Inputvariablen $X_1 = X_2 = X_3 = 0$ gilt.

a) Berechnen Sie den Euklidischen Abstand zwischen den sechs Inputvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_6 \in \mathbb{R}^3$ und $\mathbf{x} = (0, 0, 0)^T$.

b) Welche Prognose für den Wert der Outputvariable T liefert die k-Nächste-Nachbarn-Klassifikation für $k = 1$ und $k = 3$?

c) Wenn die (optimale) Bayes-Entscheidungsgrenze sehr flexibel (d.h. hochgradig nichtlinear) ist, sollte dann der Glättungsparameter $k \in \mathbb{N}$ eher groß oder klein gewählt werden?

Aufgabe 5

Die folgende Tabelle gibt für einen Tennisspieler an, ob er in den letzten 14 Tagen Tennis gespielt hat und wie die konkreten Wetterbedingungen an diesen Tagen jeweils waren:

Tag	Ausblick (X_1)	Temperatur (X_2)	Luftfeuchtigkeit (X_3)	Wind (X_4)	Tennis (T)
1	sonnig	heiß	hoch	schwach	Nein
2	sonnig	heiß	hoch	stark	Nein
3	bewölkt	heiß	hoch	schwach	Ja
4	regnerisch	mild	hoch	schwach	Ja
5	regnerisch	kühl	normal	schwach	Ja
6	regnerisch	kühl	normal	stark	Nein
7	bewölkt	kühl	normal	stark	Ja
8	sonnig	mild	hoch	schwach	Nein
9	sonnig	kühl	normal	schwach	Ja
10	regnerisch	mild	normal	schwach	Ja
11	sonnig	mild	normal	stark	Ja
12	bewölkt	mild	hoch	stark	Ja
13	bewölkt	heiß	normal	schwach	Ja
14	regnerisch	mild	hoch	stark	Nein

In Abhängigkeit von den konkreten Wetterbedingungen soll mittels Naiver Bayes-Klassifikation ermittelt werden, ob der Tennisspieler am nächsten Tag Tennis spielt oder nicht.

a) Berechnen Sie Schätzungen für die (bedingten) Wahrscheinlichkeiten, die zur Bestimmung der Naiven Bayes-Klassifikation benötigt werden.

b) Bestimmen Sie mit Hilfe der Ergebnisse aus Teil a), ob ein Tennisspieler Tennis spielen wird, wenn der Ausblick sonnig, die Temperatur kühl, die Luftfeuchtigkeit hoch und der Wind stark ist.

c) Bestimmen Sie für die in Aufgabenteil b) genannten Wetterbedingungen eine Schätzung für die a posteriori Wahrscheinlichkeit, dass ein Tennisspieler Tennis spielen wird bzw. nicht Tennis spielen wird.

Aufgabe 6 †

Gegeben sei ein Datensatz, der zufällig in einen Trainingsdatensatz und einen Testdatensatz der gleichen Größe zerlegt wird. An die beiden Teildatensätze wird zuerst ein Logit-Modell angepasst. Dabei resultiert auf den Trainingsdaten eine Fehlerquote (Anzahl der falschen Klassifikationen geteilt durch die Anzahl aller Klassifikationen) von 20% und auf den Testdaten eine Fehlerquote von 30%. Anschließend wird die Nächste-Nachbar-Klassifikation (d.h. die k-Nächste-Nachbarn-Klassifikation mit $k = 1$) angewendet. Dabei resultiert als durchschnittliche Fehlerquote (gemittelt über Test- und Trainingsdatensatz) von 18%. Welche dieser beiden Verfahren sollte man auf der Grundlage dieser Ergebnisse für die Klassifizierung neuer Beobachtungen verwenden? Begründen Sie ihre Antwort.

Aufgabe 7 †

Betrachtet wird ein Trainingsdatensatz

$$\Big\{(\mathbf{x}_n, t_n) \in \{0,1\}^{p+1} : n = 1, \ldots, N\Big\}$$

mit den Testergebnissen bzgl. der Infekion mit einem bestimmten Virus bei N verschiedenen Patienten in p unterschiedlichen Laboren. Dabei gibt $\mathbf{x}_n = (x_{n1}, \ldots, x_{np})^T \in \{0,1\}^p$ mit

$$x_{ni} := \begin{cases} 0 & \text{Test war negativ} \\ 1 & \text{Test war positiv} \end{cases}$$

an, ob für Patient $n \in \{1, \ldots, N\}$ der Test in Labor $i \in \{1, \ldots, p\}$ positiv oder negativ war, und

$$t_n := \begin{cases} 0 & \text{Patient ist nicht mit dem Virus infiziert} \\ 1 & \text{Patient ist mit dem Virus infiziert} \end{cases}$$

gibt Auskunft darüber, ob Patient n tatsächlich mit dem Virus infiziert ist oder nicht. Es wird angenommen, dass in der Bevölkerung die Wahrscheinlichkeit mit dem Virus tatsächlich infiziert zu sein $\pi \in (0,1)$ beträgt.

a) Ermitteln Sie den Naiven Bayes-Klassifikator, mit dem für einen neuen Patienten bei Vorliegen der Testergebnisse aus den p Laboren entschieden werden kann, ob er wahrscheinlich mit dem Virus infiziert ist oder nicht.

b) Es wird nun zusätzlich angenommen, dass alle p Labore mit der gleichen Wahrscheinlichkeit fälschlicherweise ein positives Testergebnis bzw. fälschlicherweise ein negatives Testergebnis liefern und diese Wahrscheinlichkeiten kleiner als $\frac{1}{2}$ sind. Bestimmen Sie wie viele der p Testergebnisse positiv sein müssen, damit ein Patient durch den Naiven Bayes-Klassifikator als mit dem Virus infiziert eingestuft wird.

c) Betrachtet wird die Situation zu Beginn einer Pandemie, wo die Zuverlässigkeit von Tests häufig gering ist, d.h. negative Testergebnisse oft auftreten. Dazu sei zusätzlich angenommen, dass $\pi = 0{,}2$ gilt und dass für alle p Labore der Anteil der falschen positiven Testergebnisse 1% und der Anteil der falschen negativen Testergebnisse 40% beträgt. Führt dann der Naiven Bayes-Klassifikator bei einem Patienten mit 3 positiven Testergebnissen bei insgesamt $p = 10$ Laboren zu einer Einstufung als mit dem Virus infiziert?

† Zusatzaufgabe