Projektaufgabe

Brustkrebs-Diagnose

Schreiben Sie ein dokumentiertes Notebook zur Vorhersage von Brustkrebs.

Hinweis: Die Aufgabe lässt sich einfacher mit Hilfe des Moduls `numpy` lösen. Ziel der Aufgabe ist es jedoch, die Konzepte aus der Vorlesung zu üben.

Daten: Breast Cancer Coimbra Data Set [1]

Laden Sie die Datei `cancer.txt` von der PGKI-Kurseite herunter. Der Datensatz enthält klinische Daten von 116 Patientinnen. Der folgende Ausschnitt zeigt die ersten Zeilen des Datensatzes:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
48	23.50	70	2.71	0.47	8.81	9.70	8.00	417.11	1
83	20.69	92	3.12	0.71	8.84	5.43	4.06	468.79	1
82	23.12	91	4.50	1.01	17.94	22.43	9.28	554.70	1
68	21.37	77	3.23	0.61	9.88	7.17	12.77	928.22	1

x_1: Age x_2: BMI x_3: Glucose x_4: Insulin x_5: HOMA x_6: Leptin
x_7: Adiponectin x_8: Resistin x_9: MCP.1 y: Classification

Die Tabelle besteht aus 116 Zeilen und 10 Spalten. Jede Zeile entspricht einer Teilnehmerin. Die ersten neun Spalten $x_1, \ldots, x_9$ beschreiben Merkmale und die letzte Spalte y die Klassenzugehörigkeit der jeweiligen Patientin. Bei den Merkmalen handelt es sich um klinische Daten, die von den Patientinnen erhoben wurden. Die Klasse beschreibt die Diagnose. Eine Patientin gehört der Klasse 1 an, wenn bei ihr Brustkrebs diagnostiziert wurde und andernfalls der Klasse 0. Von den 116 Patientinnen haben 64 Brustkrebs und 52 kein Brustkrebs.

Beispiel: Die erste Zeile beschreibt die Merkmale und die Klasse der ersten Patientin. Ihr Alter ist $x_1 = 48$, ihr BMI-Wert ist $x_2 = 23.50$, ihr Glukose-Wert ist $x_3 = 70$, usw. Ihre Klasse ist $y = 1$ (Brustkrebs).

Einlesen der Datei

Lesen Sie die Datei ein. Die Datei besteht aus einer Überschrift und aus 116 Zeilen mit je 10 numerischen Werten, die durch Leerzeichen getrennt sind.

Datentransformation

Wandeln Sie die eingelesenen Zeilen in eine geeignete numerische Repräsentation um. Verwenden Sie dazu den Konstruktor `float()`.

Klassifikation

Schätzen Sie die Fehlerrate der nächsten Nachbarin Regel unter Verwendung der Euklidischen Distanz (`math.dist()`). Testprotokoll ist leave-one-out validation:

Zerlegen Sie den Datensatz in eine Trainings- und Testmenge. Die Testmenge besteht aus genau einer Patientin. Alle anderen Patientinnen bilden die Trainingsmenge. Sagen Sie die Klasse der Testpatientin vorher. Wiederholen Sie dieses Experiment, so das jede Patientin genau einmal Testpatientin ist. Ermitteln Sie die Fehlerrate über alle 116 Experimente.

Z-Transformation

Die Vorhersagen der nächsten-Nachbarin Regel sind unbefriedigend. Wir können versuchen, ihre Vorhersagequalität zu verbessern. Ein Blick auf die Daten zeigt, dass die Werte in unterschiedlichen Spalten eine unterschiedliche Größenordnung besitzen. Zum Beispiel sind die Werte der Spalte x_9 um etwa 500- bis 1000-fach größer als die Werte der Spalte x_5. Das bedeutet, dass die euklidische Distanz von x_9-Werten dominiert wird, während x_5-Werte keine bedeutende Rolle spielen. Die Idee ist, die Daten so zu transformieren, dass die Spalten in ihrer Größenordnug vergleichbar werden und nicht mehr einzelne Spalten die euklidische Distanz dominieren. Das lässt sich mit der Z-Transformation realisieren.

Aufgabe: Implementieren Sie eine Z-Transformation (siehe Tutorial unten). Transformieren Sie die Daten und wiederholen Sie die leave-one-out Validierung mit den transformierten Daten.

Tutorial: Um die Z-Transformation zu beschreiben, betrachten wir den folgenden vereinfachten Datensatz:

x_1	x_2	x_3	y
1	30	100	1
3	10	200	0
2	15	150	1
1	15	200	0

Für die Z-Transformation sind nur die x_j-Spalten relevant. Die y-Spalte bleibt unverändert. Wir betrachten also die Matrix $X = (x_{ij})$ bestehend aus 4 Zeilen und 3 Spalten.

Als erstes berechnen wir den Mittelwert und die Standardabweichung jeder Spalte j von X:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mu_i - x_{ij})^2}$$

Dabei ist n die Anzahl der Zeilen (hier: $n = 4$), μ_j ist der Mittelwert der j-ten Spalte und σ_j ist die Standardabweichung der j-ten Spalte von X. Für die erste Spalte sieht die Berechnung folgendermaßen aus (gerundete Werte, keine Indizes):

	x	$x - \mu$	$(x-\mu)^2$
	1	$1 - 1.75 = -0.75$	0.56
	3	$3 - 1.75 = +1.25$	1.56
	2	$2 - 1.75 = +0.25$	0.06
	1	$1 - 1.75 = -0.75$	0.56
sum	7	0	2.75
μ	$7/4 = \mathbf{1.75}$	$0/4 = 0$	$2.75/4 = 0.69$
σ			$\sqrt{0.69} = \mathbf{0.83}$

Somit ist der Mittelwert der ersten Spalte $\mu_1 = 1.75$ und die Standarabweichung ist $\sigma_1 = 0.83$. Die folgende Tabelle enthält die Mittelwerte und Standardabweichungen aller Spalten von X:

	x_1	x_2	x_3
	1	30	100
	3	10	200
	2	15	150
	1	15	200
μ	1.75	17.5	162.5
σ	0.83	7.50	41.46

Die Z-Transformation transformiert jeden Wert x_{ij} zum Wert

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Die folgende Tabelle zeigt die Werte der transformierten Daten:

z_1	z_2	z_3
-0.90	1.67	-1.51
1.51	-1.00	0.90
0.30	-0.33	-0.30
-0.90	-0.33	0.90

Die transformierten Daten besitzen die Eigenschaft, dass jede Spalte (näherungsweise) den Mittelwert 0 mit Standardabweichung 1 hat. Mit anderen Worten, die Größenordnung der Werte in den verschiedenen Spalten ist nun vergleichbar.

References

[1] Patricio, M., Pereira, J., Crisostomo, J., Matafome, P., Gomes, M., Seica, R., and Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer, 18(1).