Cvičení 5: Lineární diskriminační analýza

Úkol 1.: Třídění do dvou skupin
Použijte datový soubor SIDS.sta, který obsahuje údaje o 65 novorozencích, z nichž někteří zemřeli na syndrom náhlého úmrtí kojence. Obsahuje tyto proměnné :
ID … má hodnotu 1, když novorozenec žije , hodnotu 2, když umřel na syndrom náhlé smrti kojence (SIDS)
X1 … počet tepů za minutu
X2 … porodní hmotnost v gramech
X3 … popisuje funkci srdce a plic
X4 … počet týdnů těhotenství (všichni se narodili aspoň v 37 týdnu, což je považováno za ukončené období zdravého vývoje plodu)

Ověřte předpoklady pro provedení LDA.
(S-W test normality prokázal porušení normality u proměnné X4 ve skupině 1, Boxův test nezamítl shodu variančních matic na hladině významnosti 0,05, linearita vztahů mezi proměnnými je v obou skupinách přibližně splněna.)

Zjistěte význam jednotlivých proměnných v modelu:

	Výsledky diskriminační funkční analýzy (SIDS.sta) Počet prom. v modelu: 4; grupovací: ID (2 skup) Wilk. lambda: ,68278 přibliž F (4,60)=6,9691 p< ,0001					
N=65	Wilk. Lambda	Parc. Lambda	F na vyj (1,60)	p-hodn.	Toler.	1-toler. R^2
X1	0,682851	0,999893	0,00641	0,936461	0,953604	0,046396
X2	0,757725	0,901089	6,58610	0,012792	0,849030	0,150970
X3	0,831482	0,821157	13,06763	0,000616	0,917977	0,082023
X4	0,686567	0,994481	0,33299	0,566064	0,835387	0,164613

Test hypotézy o shodě vektorů středních hodnot v obou skupinách je na hladině významnosti 0,05 průkazný. Největší vliv na diskriminaci mají proměnné X2 a X3.

Vypočtěte Mahalanobisovy vzdálenosti skupin a odpovídající p-hodnoty (2,4267, p = 0,000112).

Jaké jsou apriorní pravděpodobnosti příslušnosti objektů ke skupinám? (p_1 = 0,75385, p_2 = 0,24615).

Stanovte odhad Fisherovy lineární diskriminační funkce:
L(**x**) = -0,00178322553X1+0,00178850321X2-15,5253107X3+0,214815554X4-7,35265591

Posuďte účinnost diskriminace resubstituční metodou.
81,54 % objektů je správně zařazeno.

Které objekty byly zařazeny chybně?
(V 1. skupině objekty č. 14, 32, 34, ve 2. skupině 50, 52, 54, 57, 59, 60, 62, 64, 65)

Odhadněte celkovou pravděpodobnost mylné klasifikace při náhodném zařazování (0,26).

Dále proveďte LDA krokovou dopřednou metodou:
(Do modelu byly zařazeny proměnné X3 a X2, odhad Fisherovy lineární diskriminační funkce je:
L(**x**) = -16,0770541X3+0,00194756243+0,613039967, úspěšnost diskriminace je 81,54 %.)

Proveďte klasifikaci objektů do dvou skupin pomocí neuronové sítě (úspěšnost diskriminace je 100 %).

Úkol 2.: Třídění do tří skupin
Použijte datový soubor ropa.sta, který byl popsán ve cvičení 3.

Zjistěte význam jednotlivých proměnných v modelu:

N=45	Výsledky diskriminační funkční analýzy (ropa.sta) Počet prom. v modelu: 4; grupovací: ID (3 skup) Wilk. lambda: ,17959 přibliž F (8,78)=13,257 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,39)	p-hodn.	Toler.	1-toler. R^2
X1	0,229700	0,781858	5,44059	0,008241	0,730601	0,269399
X2	0,213007	0,843133	3,62803	0,035890	0,736104	0,263896
X3	0,219437	0,818427	4,32621	0,020096	0,648482	0,351519
X4	0,321952	0,557825	15,45717	0,000011	0,662875	0,337126

Test hypotézy o shodě vektorů středních hodnot v obou skupinách je na hladině významnosti 0,05 průkazný. Největší vliv na diskriminaci mají proměnné X1 a X4.

Vypočtěte Mahalanobisovy vzdálenosti skupin a odpovídající p-hodnoty.

ID	Mahalanobisovy vzdálenosti^2 (ropa.sta)		
	G_1:1	G_2:2	G_3:3
G_1:1	0,00000	10,15239	17,55756
G_2:2	10,15239	0,00000	6,78236
G_3:3	17,55756	6,78236	0,00000

ID	p-hodnot (ropa.sta)		
	G_1:1	G_2:2	G_3:3
G_1:1		0,000037	0,000000
G_2:2	0,000037		0,000012
G_3:3	0,000000	0,000012	

Jaké jsou apriorní pravděpodobnosti příslušnosti objektů ke skupinám? (p_1 = 0,15556, p_2 = 0,17778, p_3 = 0,66667).

Najděte odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu:

Proměnná	Klasifikační funkce; grupovací : ID (ropa.sta)		
	G_1:1 p=,15556	G_2:2 p=,17778	G_3:3 p=,66667
X1	0,3645	0,4964	0,5373
X2	0,9792	0,8364	0,6613
X3	0,0499	0,0688	0,0534
X4	0,0085	-0,0064	-0,0042
Konstant	-49,0185	-50,0807	-38,9736

Posuďte účinnost diskriminace resubstituční metodou:
93,33 % objektů je správně zařazeno.

Které případy byly zařazeny chybně?
(V 1. skupině objekt č. 2, ve 2. skupině 10, ve 3. skupině 34)

Dále proveďte LDA krokovou zpětnou metodou:
(Do modelu byly zařazeny proměnné X1 a X4, odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu jsou:

	Klasifikační funkce; grupovací : ID (ropa.sta)		
Proměnná	G_1:1 p=,15556	G_2:2 p=,17778	G_3:3 p=,66667
X1	0,0773	0,18557	0,2940
X4	0,0155	0,00478	0,0044
Konstant	-11,6852	-7,66368	-12,9968

Úspěšnost diskriminace je 88, 89 %.)

Proveďte klasifikaci objektů do tří skupin pomocí neuronové sítě (úspěšnost diskriminace je 100 %).

Nepovinný úkol: Na datovém souboru Irisdat.sta, který obsahuje údaje o délce a šířce okvětních a kališních lístků 150 rostlin tří druhů kosatců (Setosa, Virginic, Versicola) proveďte kanonickou diskriminační analýzu a lineární diskriminační analýzu, eventuálně použijte pro klasifikaci kosatců automatické neuronové sítě.