Plant Biosystematics Group

Masaryk University
Brno, Czech Republic

PŘÍRODOVĚDECKÁ FAKULTA · MASARYKOVA UNIVERZITA · BRNO

ÚSTAV BOTANIKY A ZOOLOGIE

Bi6589

Laboratorní a bioinformatické metody rostlinné biosystematiky

Analýza a vizualizace více dimenzionálních dat

ORDINACE

Ordinace

- Naměřili jsme u vzorků spoustu údajů (PROMĚNNÝCH) = vícedimenzionální data

ID	Taxon	Taxon_group	Taxon_id	Delka_list	Sir_list_prohluben	Sir_list_lalok	Ubory	Ochlupeni	ostnitost_lodyhy	Sir_uboru	sir_zakrovu	del_zakrovu	PRUM_del_kor_trubky	PRUM_sir_listenu	PRUM _del_listenu	GS	viabilita	PRUD_prumer	pomer-DL-SLP	pomer-DL-SLL	pomer-DZ-SZ	pomer-DLn-SLn	pohlavi
1.	CAN	CAN	CAN1	266	52	60	4	0	0	47	26	15	12	2,2	16,4	2,03	99	18,58	5,115	4,433	0,577	7,455	H
2.	CAN	CAN	CAN2	218	25	45	3	0	0	40	21	14	10,6	3	11,4	2,04	100	20,08	8,72	4,844	0,667	3,8	H
3.	CAN	CAN	CAN3	246	29	37	3	0	0	46	28	15	9,2	2,2	12,8	2,06	97	20,12	8,483	6,649	0,536	5,818	H
4.	CAN	CAN	CAN4	236	35	40	3	0	0	48	32	18	15,8	2,2	14	2,02	95	18,72	6,743	5,9	0,563	6,364	H
5.	CAN	CAN	CAN5	264	40	48	4	0,5	0	48	24	13	10,6	2,6	8	1,99	90	18,15	6,6	5,5	0,542	3,077	H
6.	CAN	CAN	CAN6	222	28	41	4	0	0	45	24	14	12	2,4	11,4	2,02	96	18,56	7,929	5,415	0,583	4,75	H
7.	CAN	CAN	CAN7	243	36	46	3	0	0	47	26	15	12	2	13,8	2,03	95	19,89	6,75	5,283	0,577	6,9	H
11.	PAL	PAL	PAL11	96	5	19	25	1	1	20	14	12	5	2	7,2	2,27	89	19,85	19,2	5,053	0,857	3,6	H
12.	PAL	PAL	PAL12	124	7	19	28	1	1	21	15	9	5,2	2	6	2,31	100	18,89	17,71	6,526	0,6	3	F
13.	PAL	PAL	PAL13	111	9	20	27	1	1	21	14	9	5,8	1,6	9,4	2,25	92	18,83	12,33	5,55	0,643	5,875	F

Ordinace

- Naměřili jsme u vzorků spoustu údajů (PROMĚNNÝCH) = vícedimenzionální data
- Identifikace vhodných znaků pro diskriminační analýzu
- Ordinace = redukce vícedimenzionálních dat a vizualizace vztahů mezi vzorky
- Více ordinačních metod (např. PCA, PCoA, DCA, NMDS) - rozdíly spočívají v metodologii, předpokladech a typech dat

Přehled vybraných ordinačních metod

Metoda	Typ vzdáleností	Lineárnost	Model vztahů	Hlavní použití
PCA	Euklidovská	Lineární	Ortogonální	Kontinuální data, lineární vztahy
PCoA	Různé metriky	Nemusí být	Generalizované	Diskrétní/číselná data, různé metriky
DCA	Unimodální	Nelineární	Unimodální	Ekologická data s gradienty
NMDS	Nemetrické	Nelineární	Pořadí vzdáleností	Komplexní a nestrukturovaná data

Přehled vybraných ordinačních metod

Metoda	Typ vzdáleností	Lineárnost	Model vztahů	Hlavní použití
PCA	Euklidovská	Lineární	Ortogonální	Kontinuální data, lineární vztahy
PCoA	Různé metriky	Nemusí být	Generalizované	Diskrétní/číselná data, různé metriky
DCA	Unimodální	Nelineární	Unimodální	Ekologická data s gradienty
NMDS	Nemetrické	Nelineární	Pořadí vzdáleností	Komplexní a nestrukturovaná data

Tranformace dat a jejich standardizace umožňuje využití více metod

Přehled vybraných ordinačních metod

Metoda	Typ vzdáleností	Lineárnost	Model vztahů	Hlavní použití
PCA	Euklidovská	Lineární	Ortogonální	Kontinuální data, lineární vztahy
PCoA	Různé metriky	Nemusí být	Generalizované	Diskrétní/číselná data, různé metriky
DCA	Unimodální	Nelineární	Unimodální	Ekologická data s gradienty
NMDS	Nemetrické	Nelineární	Pořadí vzdáleností	Komplexní a nestrukturovaná data

NMDS (*Non-Metric Multidimensional Scaling*) - nemetrické více-dimenzionální škálování; umožňuje zkoumat komplexní ekologická a taxonomická data tím, že je redukuje do menšího počtu dimenzí při zachování vzorců podobností (nebo rozdílů).

NMDS (*Non-Metric Multidimensional Scaling*)

Důležitá je hladina stresu:

Stress	Fit	Description
<0.05	Excellent	considered best for NMDS interpretation
<0.1	Good	good ordination with little risk of misinterpreation
<0.2	Fair	usable but higher values approach poor interpretation
>0.2	Poor	poorly represents the data

Analýza a vizualizace více dimenzionálních dat

Diskriminační analýza

Diskriminační analýza

- statistické metody používané k rozlišování mezi dvěma nebo více skupinami na základě hodnot vstupních proměnných
- cílem je klasifikovat nové pozorování do správné skupiny nebo pochopit vztahy mezi proměnnými a skupinami.

Volba metody závisí

- **Povaze dat** (např. lineární vs. nelineární vztahy).
- **Předpokladech** (např. normalita, homogenita kovariančních matic).
- **Účelu analýzy** (např. predikce, interpretace).

Přehled vybraných metod Diskriminační analýzy (dle GPT chat)

Metoda	Cíl	Předpoklady	Použití	Výhody
Lineární diskriminační analýza (LDA)	Najít lineární kombinace proměnných pro separaci skupin.	Normalita dat, homogenní kovarianční matice.	Klasifikace, redukce dimenze dat.	Jednoduchost, efektivní pro lineárně separovatelné skupiny.
Kvadratická diskriminační analýza (QDA)	Umožnit různé kovarianční matice mezi skupinami.	Normalita dat, rozdílné kovarianční struktury.	Nelineárně separovatelné skupiny.	Flexibilita, zvládá složitější separace.
Kanonická diskriminační analýza (CDA)	Maximalizace separace mezi skupinami pomocí kanonických funkcí.	Podobné jako LDA.	Redukce dimenze, analýza odlišností mezi skupinami.	Vizuální interpretace separace skupin.
Naivní Bayesova metoda	Klasifikace na základě nezávislých proměnných.	Nezávislost mezi proměnnými.	Rychlá klasifikace, textová analýza.	Rychlost, jednoduchost implementace.
Flexibilní diskriminační analýza (FDA)	Nelineární klasifikace pomocí spline nebo jiných metod.	Žádné přísné předpoklady.	Nelineární vztahy mezi proměnnými a skupinami.	Flexibilita v modelování.
Diskriminační analýza s penalizací	Zvládnout vysokorozměrná data a snížit overfitting.	Závisí na konkrétní penalizaci (ridge, lasso).	Genomika, chemometrie.	Řešení problémů s vysokou dimenzionalitou.
Smíšené modely DA	Zohlednit hierarchickou strukturu nebo opakovaná měření.	Korelace mezi daty nebo hierarchická struktura.	Situace s opakovanými pozorováními.	Zvládá složité datové struktury.
Neuronové sítě v DA	Klasifikace nelineárních a složitých dat pomocí hlubokého učení.	Velká datová množství a výkonný hardware.	Obrázky, texty, komplexní úlohy.	Vysoce flexibilní, robustní.
Random Forest Diskriminace	Klasifikace a odhad pravděpodobností pomocí rozhodovacích stromů.	Robustní vůči šumu, zvládá různé typy proměnných.	Predikce, analýza s velkým šumem v datech.	Robustní, zvládá velké množství proměnných.

Přehled vybraných metod Diskriminační analýzy (dle GPT chat)

Metoda	Cíl	Předpoklady	Použití	Výhody
Lineární diskriminační analýza (LDA)	Najít lineární kombinace proměnných pro separaci skupin.	Normalita dat, homogenní kovarianční matice.	Klasifikace, redukce dimenze dat.	Jednoduchost, efektivní pro lineárně separovatelné skupiny.
Kvadratická diskriminační analýza (QDA)	Umožnit různé kovarianční matice mezi skupinami.	Normalita dat, rozdílné kovarianční struktury.	Nelineárně separovatelné skupiny.	Flexibilita, zvládá složitější separace.
Kanonická diskriminační analýza (CDA)	Maximalizace separace mezi skupinami pomocí kanonických funkcí.	Podobné jako LDA.	Redukce dimenze, analýza odlišností mezi skupinami.	Vizuální interpretace separace skupin.
Naivní Bayesova metoda	Klasifikace na základě nezávislých proměnných.	Nezávislost mezi proměnnými.	Rychlá klasifikace, textová analýza.	Rychlost, jednoduchost implementace.
Flexibilní diskriminační analýza (FDA)	Nelineární klasifikace pomocí spline nebo jiných metod.	Žádné přísné předpoklady.	Nelineární vztahy mezi proměnnými a skupinami.	Flexibilita v modelování.
Diskriminační analýza s penalizací	Zvládnout vysokorozměrná data a snížit overfitting.	Závisí na konkrétní penalizaci (ridge, lasso).	Genomika, chemometrie.	Řešení problémů s vysokou dimenzionalitou.
Smíšené modely DA	Zohlednit hierarchickou strukturu nebo opakovaná měření.	Korelace mezi daty nebo hierarchická struktura.	Situace s opakovanými pozorováními.	Zvládá složité datové struktury.
Neuronové sítě v DA	Klasifikace nelineárních a složitých dat pomocí hlubokého učení.	Velká datová množství a výkonný hardware.	Obrázky, texty, komplexní úlohy.	Vysoce flexibilní, robustní.
Random Forest Diskriminace	Klasifikace a odhad pravděpodobností pomocí rozhodovacích stromů.	Robustní vůči šumu, zvládá různé typy proměnných.	Predikce, analýza s velkým šumem v datech.	Robustní, zvládá velké množství proměnných.

Random Forest Diskriminace

Předpoklady

- **Závislá proměnná:** Musí být kategorická (= vzorky, resp. jména vzorků).
- **Nezávislé proměnné:** Kontinuální nebo kategorické prediktory, které vysvětlují závislou proměnnou (= studované znaky).
- **Vyvážení skupin:** vzorky z různých skupin by v datovém souboru měly být vyvážené.

Interpretace výsledků:

- **Význam proměnných (Feature Importance):** Identifikace proměnných s největším přínosem pro predikci.
- **Predikce (Class probabilities):** Pravděpodobnosti přiřazení k jednotlivým třídám.