Pokročilé statistické metody
7. cvičení

Diskriminační analýza (DA)

- kanonická DA
- kanonická lineární DA
- popisná DA
- Fisherova DA

Diskriminační analýza – PROČ?

- Jak se liší diskriminační analýza od shlukové analýzy? („unsupervised“ vs. „supervised“)
- Doplňte: „Nové osy diskriminační analýzy jsou tvořeny tak, aby … “
- Co vyjadřuje vlastní číslo osy diskriminační analýzy?
- Jaké jsou předpoklady diskriminační analýzy?

Diskriminační analýza – PROČ?

- Jak se liší diskriminační analýza od shlukové analýzy? („unsupervised“ vs. „supervised“)
- Doplňte: „Nové osy diskriminační analýzy jsou tvořeny tak, aby … “
- Co vyjadřuje vlastní číslo osy diskriminační analýzy?
  - Popisují variabilitu spjatou s kanonickými osami
- Jaké jsou předpoklady diskriminační analýzy?

Diskriminační analýza - cíle

1. **Vytvoření zástupných proměnných**, které nejlépe odliší skupiny objektů.
2. **Vytvoření pravidla pro klasifikaci** objektů do skupin.
   a) Identifikace proměnných diskriminujících mezi předem danými skupinami objektů.
   b) Vyhodnocení klasifikace pro objekty, u kterých známe zařazení do skupin.
3. **Klasifikace** nových objektů do skupin.

Využití:

- v antropologii pro klasifikaci koster,
- v medicíně k určení rizikovosti pacientů,
- ve finančnictví k předvídání krachů firem,
- v biologii ke klasifikaci rostlin,
- v sociologii u psychologických testů.

Výběr proměnných do modelu

- Výběr provádíme na základě:

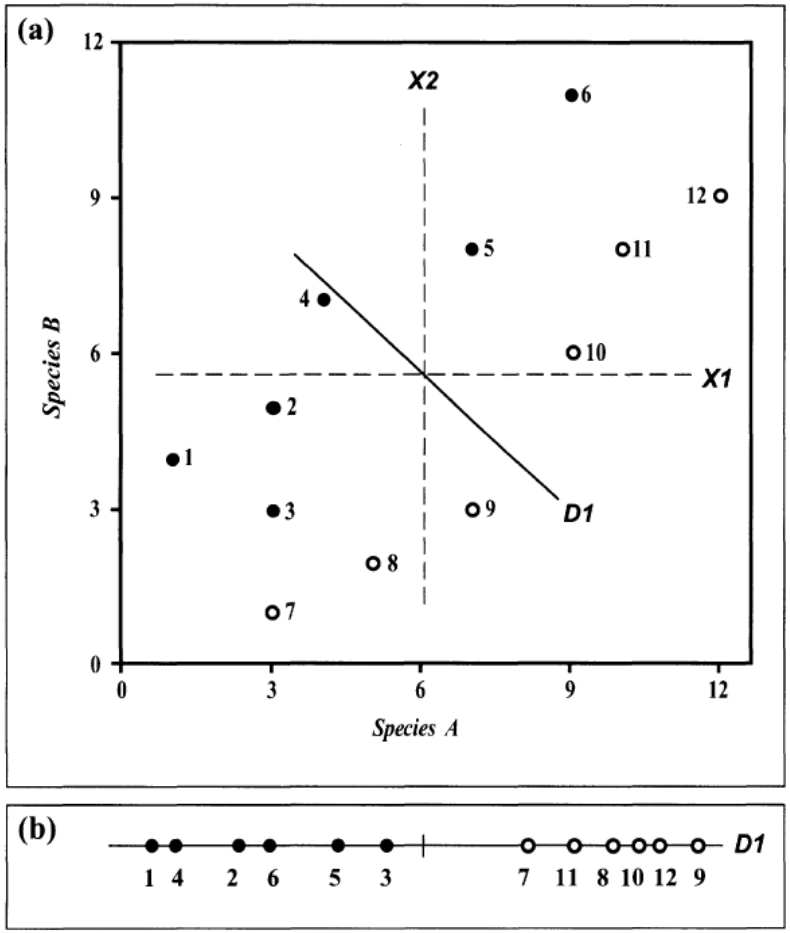
1. Expertní znalosti proměnných (zohledňujeme např. finanční zátěž, chybovost měření, vyplněnost).
2. Pozorovaných dat (hodnotíme korelace proměnných, přínos unikátní informace - % rozptylu, které popisuje, příspěvek k diskriminaci, atd.).
3. Dopředné/zpětné eliminace (proměnné jsou postupně přidávány/odebírány tak, aby došlo k významnému „zlepšení“ modelu).

Diskriminační analýza – algoritmus

2 fáze výpočtu:

1. Vytvoření kanonických os

   - Z původně vysokého počtu parametrů vytvoříme nové osy, které odliší shluky v datech.
   - Osy nejsou v prostoru původních proměnných ortogonální (jako tomu bylo u PCA).
   - Maximální počet os je roven počtu skupin mínus jedna.



Kenkel et al. (2002)

Diskriminační analýza – algoritmus

2 fáze výpočtu:

2. Klasifikace objektů do skupin.
   - Na vstupu definujeme apriorní pravděpodobnosti zařazení objektů do skupin.
   - Pro každý objekt je spočítána vzdálenost od centroidu dané skupiny.
   - Kombinací apriorní pravděpodobnosti a Mahalanobisovy vzdálenosti jsou spočítány posteriorní pravděpodobnosti zařazení objektu do dané skupiny.
   - Pro každou ze skupin je definována diskriminační funkce. Při klasifikaci nových objektů zařadíme objekt do té skupiny, kde diskriminační funkce nabývá maxima.

Diskriminační analýza – OMEZENÍ?

- Předpoklad vícerozměrného normálního rozdělení prediktorů v každé ze skupin.
- Citlivá na přítomnost odlehlých hodnot.
- Citlivá na redundantní proměnné v modelu.
- Rovnice modelu je v základní verzi lineární a tedy i hodnocený problém musí mít lineární řešení.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu - analogické s ANOVA – hodnotí podíl vnitroskupinového a celkového rozptylu (rozsah: 0–1; hodnoty blízké nule značí dobrou diskriminaci skupin),
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných - wilksovo lambda celého modelu při vyřazení dané proměnné (naopak: čím větší, tím je proměnná důležitější pro diskriminaci),
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda: unikátní příspěvek dané proměnné k diskriminaci (čím nižší je hodnota, tím větší unikátní diskriminační sílu prediktor nese),
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance: unikátní variabilita proměnné nevysvětlená ostatními proměnnými v modelu (1 - tolerance = R^2 variabilita proměnné, kterou lze vysvětlit kombinací ostatních proměnných).

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:** vytváří nové osy tak, aby jejich diskriminační funkce byla co největší (počet nových os = min(počet skupin, počet proměnných) -1)
  a) Vlastní vektory: určují směr nových os (definovány jako lineární kombinace proměnných v modelu).
  b) Vlastní čísla: popisují podíl variability mezi a v rámci skupin objektů na nových osách. Osy s nízkou hodnotou vlastního čísla nepřispívají k popisu rozdílu mezi skupinami.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost: pravděpodobnost výskytu objektu ve shluku (rovnoměrná/proporcionální/nastavená uživatelem na základě znalostí dané problematiky)
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost: Používána pro popis vzdáleností objektů od centroidů skupin a následně pro výpočet posteriorních pravděpodobností,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce: pro každou skupinu jedna rovnice, objekt je zařazen do skupiny s maximální hodnotou skóre klasifikační funkce.
  d) Posteriorní pravděpodobnost.

Výstup diskriminační analýzy

- **Popis významu proměnných v modelu:**
  a) Wilksovo lambda modelu,
  b) Wilksovo lambda proměnných,
  c) Parciální lambda,
  d) Tolerance.

- **Kanonická analýza:**
  a) Vlastní vektory,
  b) Vlastní čísla.

- **Klasifikace objektů:**
  a) Apriorní pravděpodobnost,
  b) Mahalanobisova vzdálenost,
  c) Diskriminační funkce,
  d) Posteriorní pravděpodobnost: pravděpodobnost klasifikace objektu do dané skupiny (kombinace Mahalanobisových vzdáleností objektů od centroidů shluků s apriorní pravděpodobností).

Validace modelu

- Maximální predikční síla vs. minimální složitost
- Ideálně na nezávislém datovém souboru, na kterém nebyl model vyvinut. Může se stát, že na naše data bude model sedět perfektně a na jiném souboru zcela selže (bude přetrénovaný).
- Pokud nemáme takový další datový soubor, lze využít validačních technik:
    a) Krosvalidace,
    b) „Leave one out“,
    c) Permutační metody.