Uloha5

Kdy pro hledání klasifikačního pravidla použijete lineární diskriminační funkci a kdy logistickou regresi a proč?

Obě metody slouží k nalezení klasifikačního pravidla, které odděluje třídy vícerozměrných dat.

Rozdíl spočívá v jejich statistických předpokladech a způsobu, jakým modelují podmíněnou pravděpodobnost.

Kdy použít Lineární diskriminační funkci (LDF)

LDF je generativní model, který modeluje rozdělení vstupních dat X v rámci každé třídy Y ($P(X|Y)$).

Použití LDF je vhodné, když:

- Předpoklad normality je splněn: datové body pro každou třídu pocházejí z vícerozměrného normálního rozdělení.
- Předpoklad homoskedasticity je splněn: třídy sdílejí stejnou kovarianční matici (tj. rozptyl a korelace jsou ve všech třídách stejné, jen se liší střední hodnoty).
- Počet vzorků je malý: LDF je stabilnější a efektivnější než Logistická regrese, zejména pokud je počet trénovacích vzorků malý a počet prediktorů je velký.
- Více než dvě třídy (Multiklasifikace): LDF je navržena pro libovolný počet tříd ($K > 2$)

Proc?
Pokud jsou předpoklady o normalitě a stejné kovarianci splněny, je LDF teoreticky optimální a poskytuje stabilnější a přesnější odhady, což vede k nižší míře chyby klasifikace než LR.

Kdy použít Logistickou regresi (LR)

Logistická regrese je diskriminační model, který přímo modeluje podmíněnou pravděpodobnost třídy Y za daných vstupních dat X ($P(Y|X)$), aniž by dělala jakékoli předpoklady o rozdělení X.

Použití LR je vhodné, když:

- Předpoklad normality je porušen: data nejsou normálně rozložena (jsou např. diskrétní, exponenciální, nebo silně šikmá)
- Předpoklad homoskedasticity je porušen (Heteroskedasticita): Kovarianční matice se mezi třídami liší (různé třídy mají různý rozptyl a tvar).
- Binární klasifikace: LR je nejčastěji používána pro dvě třídy.

- Chceme robustní odhad pravděpodobnosti: LR poskytuje přímý a snadno interpretovatelný odhad pravděpodobnosti příslušnosti ke třídě $P(Y = 1|X)$.

Proč?

- Robustnost vůči rozdělení dat: LR je robustnější, protože nevyžaduje, aby vícerozměrná data pocházela z normálního rozdělení.
- Odhad pravděpodobnosti: Přímý výstupní odhad pravděpodobnosti je užitečný pro řízení prahů rozhodování.

Pokud můžeme předpokládat, že data jsou přibližně normálně rozložena a mají podobné rozptyly (kovariance) v rámci tříd použijeme **LDF**.

Pokud jsou data silně nenormální, nebo pokud se rozptyly mezi třídami výrazně liší, použijeme **LR**.