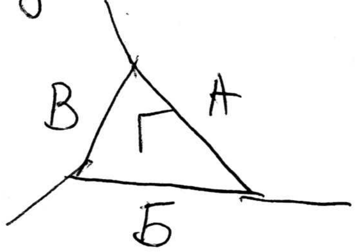
21. Линеен дискриминантен анализ. D о пак да я видя с конспекта

Тази статистическа процедура се използва, като се нуждаем от „прогнозиране" стойностите на групираща променлива. Понякога това се нарича класификация или разпознаване на образи. Нека нашата извадка е нееднородна. т.е състои се от няколко групирани наблюдения с различни вероятностни характеристики. Целта ни е да научим от тази извадка, по зададени параметри на дадено наблюдение, да определим принадлежността му към класа, от който произлиза.

В линейния дискриминантен анализ се строят линейни дискриминантни функции от предикторите. За всеки клас има точно една такава функция. Правилото за класификация е: Наблюдението се класифицира към класа с максимална дискриминантна ф-я.

Областта от стойностите на предикторите, при попадане в която наблюденията се класифицират към даден клас е изпъкнал многоъгълник. Тя се нарича класифика- класа класификационните множества биха могли да изглеждат така



Линейният дискриминантен анализ предполага, че разпределенията на количествените променливи (предикторите) вътре в класовете са нормални и се различават само по средните си стойности. Тогава процедурата произвежда оптимални ~~решения~~ решаващи правила. Тя може да се използва и при случайна (по групирането) извадка

но появата на празен клас е недопустима.
Когато броят на количествените променливи е по-голям се отбират за предиктори само тези променливи, които носят важна за разделянето информация. В това помага статистиката на Махаланобис. Тя позволява да се провери хипотеза за съвпадение на груповите средни на предикторите като цяло. Вместо критичната област за статистиката се използва вероятността съответната разпределената сл. величина да не надхвърля стойността на статистиката. Тази вероятност расте докато променливите допринасят за по-добро разпределение на класовете и започва да намалява, когато предикторите станат твърде много. Добро разделяне може да се очаква, когато хипотезата отхвърля с ~~добро~~ висока вероятност.
Когато избрани най-добрите променливи за предиктори, не означава още, че са построени класификационните множества. Основна цел на дискриминантния анализ е да се получи за причисляване на едно ново наблюдение към даден клас. За това наблюдение може да съществува априорна информация за неговата възможна принадлежност към класовете. Такава и-я е приeто да се формулира в термини на априорни вероятности ~~на класовете да се приемат за равни~~, които са необходими за отреждане на оптимални класификационни правила.
Ако такава информация не съществува, априорните вероятности на класовете се приемат за равни.
Когато извадката е случайна и новото наблюдение се избира по същия начин, може да се приемат за пропорционални на обема на класовете в обучаващата извадка.
Изборът на априорни вероятности фиксира оптимални дискретни функции и класификационни правила.

Не е удобно за ∀ ново наблюдение да се въвеждат априорни вероятности. (~~Свързано е с големи значи-~~ ~~телни~~ изчислителни ~~труд~~ Възможен начин за заобикалянето на това неудобство е представянето на класовете с помощта на няколко групиращи променливи.

Бейсов подход.

Ако са известни вероятностите $\{p(g)\}$, груповите средни $\{m(g)\}$ и вътрешно груповата ковариационна матрица

$C(g) = C$ $\forall g$ $(g = 1, 2, \ldots G)$ тогава по

формулата на Бейс, апостериорната вероятност за класификация в клас g на наблюдението $(x, .)$ ще бъде

$q(p) = c.p(g) \cdot f(x, m(g), C)$

Тук f е плътността на нормалното разпределение със средна стойност $m(g)$ и ковариационна матрица C, а c е нормираща константа такава, че $\sum q(g) = 1$

Класификационни правила

Съгласно принципа за максимално правдоподобие, класифицира се по правилото

$\hat{g} = \max h; q(h)$

Класификационните правила могат да бъдат записани във вида:

$p(\hat{g}) \cdot f(x, m(\hat{g}), C) \geq p(h) \cdot f(x, m(h), C)$

$(h = 1, 2 \ldots G)$

След логаритмуване и съкращаване получаваме:

$$b(\hat{g})'x + a(\hat{g}) \geq b(h)'x + a(h), \quad (h = 1, 2 \ldots G)$$

Векторът $b(g)$ и числото $a(g)$ се получават по формулите:

$$b(h) = m(h)'C^{-1}, \quad a(h) = \log p(h) - m(h)'C^{-1}m(h)$$

Априорни вероятности. Модели

За априорните вероятности $\{p(g)\}$ може да използваме най-добрите им оценки $\{n(g)/N\}$ при случайна извадка или друга априорна информация. За оценка на $\{m(g)\}$ и C се използват вътрешно-груповите средни обединената извадъчна вътрешно-групова ковариация.

Когато групиращите променливи са повече от една бр. на класовете G нараства. Вероятността се появява на празни клетки $(n(g) = 0)$ при случайна извадка с ограничен обем рязко се увеличава. Затруднява се и оценката за $\{m(g)\}$ в такива случаи се препоръчва използването на оценки получени от линеен модел като се направят съответните проверки с методите на дискриминантния анализ. Съответно ще се промени и оценката за C.

Стъпков дискриминантен анализ.

И тук е възприета концепцията за изборно подходящ набор от количествени променливи, с които да построим модела. Те се строят аналогично на регресията, но роля на сумите от квадратите играят

+ вътрешно-груповата ковариационна матрица

– между-груповата ковариационна матрица.

Както в едномерния случай, така и в многомерния е вярно следното равенство:

$$\sum_i \sum_j (x_{ij}-\bar{x})(x_{ij}-\bar{x})' =$$

$$= \sum_i \sum_j (x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)' + \sum_i n_i(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})'$$

$$SS = SS_{in} + SS_{mod}$$

Матрицата SS_{in} се тълкува като „сума от квадрати" отговаряща на разсейването на данните около техните локални средни и с нейна помощ се строи оценка за вътрешно-груповата ковариационна матрица C

$$\hat{C} = \frac{1}{N-G+1} SS_{in}.$$

Матрицата SS_{mod} се тълкува като „сума от квадрати" отговаряща на разсейването на груповите средни и с нейна помощ се строи оценка за между-груповата ковариационна матрица C_{mg}:

$$\hat{C}_{mg} = \frac{1}{G-1} SS_{mg}.$$

Съгласно Фишер един дискриминантен анализ е по-добър колкото е по-малко частното на двете детерминанти: $\Lambda = \frac{|SS_{in}|}{|SS|}$ наричано критерий на Уилкс.

При преминаване от зададена размерност k към размерност $k+1$ имаме изменение: $\lambda = \frac{\Lambda_{k+1}}{\Lambda_k}$

Ако новата променлива не ~~на~~ оказвало влияе на разпределението статистиката

$$F = \left(\frac{n-m}{k}\right)\left(\frac{1-\lambda}{\lambda}\right)$$ има разпределение на

Фишер.