ACP- Do Livro ADM (Daniel Peña)

Introdução

E.F.T[1]

[1]EACH-USP
Universidade de São Paulo

Do Livro ADM (Daniel Peña)

Outline

Outline

Análise Discriminante

A análise discriminante é uma técnica multivariada que cria funções discriminantes, provenientes de combinações lineares das variáveis iniciais, que maximizam as diferenças entre as médias dos grupos e minimizam a probablidade de classificações incorretas dos casos nos grupos.
É aplicada quando a variável dependente é qualitativa (grupos) e as variáveis independentes são quantitativas. As variáveis dicotômicas, como sexo, podem também ser incluídas nas variáveis explicativas.

Objetivos

- A análise discriminante tem por objetivo escolher as variáveis que distinguem os grupos de modo que, conhecendo-se as características de um novo caso, se possa prever a que grupo pertence.
- Avaliar a precisão da classificação
- A análise discriminante pode ser usada também para validar a análise de cluster e confirmar o análise fatorial.

Pressupostos

- Cada grupo é uma amostra aleatória de uma população normal multivariada. A sua violação pode levar a decisões incorretas, principalmente quando as amostras são pequenas. Quando a violação da normalidade se deve apenas à não simetria da distribuição, a potência do teste não é afetada contrariamente ao que acontece se a distribuição não for mesocúrtica e de forma mais acentuada, se for platicúrtica, casi em que devemos optar pela regressão logística.

Pressupostos

- Dentro dos grupos a variabilidade é idêntica, isto é, as matrizes da variância e covariância são iguais para todos os grupos. Caso seja violado, aumenta a probabilidadde dos casos serem classificados no grupo com maior dispersão. A violação deste pressuposto afeta sobretudo o erro tipo I quando os grupos não têm igual dimensão, mesmo que as diferenças sejam moderadas.

Introdução
Divisão das principais técnicas

- Dispõe-se de um amplo conjunto de elementos que podem vir de duas ou mais populações distintas
- em cada elemento tem-se observado uma v.a. p-dimensional x, com distribuição conhecida nas populações consideradas.
- Pretende-se classificar um novo elemento, com valores das variáveis conhecidas em uma das populações.

Introdução

A primeira aplicação da análisis discriminante consistiu em classificar os restos de uma crânio descoberto numa excavação como humano, utilizando a distribuição de medidas físicas para os crânios humanos e de antropoides. Outros exemplos de classificação são por exemplo os sistemas automáticos de concessão de crédito (credit scoring).
A técnica de análisis discriminante é conhecida também como classificação supervisionada.

Outline

problema

Sejam P_1 e P_2 duas populações nas quais temos definida uma v.a. x p-variante. Suponha que x é absolutamente contínua, com f.d.p. f_1 e f_2 conhecidas.
Estudamos o problema de classificar um novo elemento x_0 com valores conhecidos das $p-$variáveis em uma destas populações.
Se conhecermos as probabilidades a priori π_1 e π_2 como $\pi_1 + \pi_2 = 1$ de que o novo elemento venha de cada uma das populações, a distribuição de probabilidade será uma mistura:

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

problema

Uma vez observado x_0 podemos calcular as probabilidades a possteriori de que o elemento tenha sido gerado por cada uma das populações, $P(i|x_0)$, com $i = 1, 2$.
Pelo teorema de Bayes:

$$P(1|x_0) = \frac{P(x_0|1)\pi_1}{P(x_0|1)\pi_1 + P(x_0|2)\pi_2}$$

Como $P(x_0|1) = f_1(x_0)\Delta x_0$, temos:

$$P(1|x_0) = \frac{f_1(x_0)\pi_1}{f_1(x_0)\pi_1 + f_2(x_0)\pi_2}$$

e para a segunda população:

$$P(2|x_0) = \frac{f_2(x_0)\pi_2}{f_1(x_0)\pi_1 + f_2(x_0)\pi_2}$$

problema

Classificamos x_0 na população mais provável a posteriori. Como os denominadores são iguais, classificaremos x_0 em P_2 se

$$\pi_2 f_2(x_0) > \pi_1 f_1(x_0)$$

se as probabilidades a priori são iguais, a classificação em P_2 será se:

$$f_2(x_0) > f_1(x_0)$$

Consideração e consequências

Os erros cometidos podem ter diferentes consequências que podem ser quantificadas. Exemplos:

- Uma máquina classifica equivocadamente uma nota de 10 euros como sendo de 20 e devolve o troco equivocado, o custo de classificação é 10 euros.
- Se não concedermos o crédito que seria devolvido (cliente bom), perdemos o cliente e o retorno que poderia gerar, no entanto, se o crédito não é devolvido, o custo é o valor não pago.
- Se classificarmos um processo produtivo como estando sob controle, o custo será a produção defeituosa, e se por error, pararmos um processo que funciona adequadamente, o custo será o da parada e da revisão.

Representação de um problema de classificação

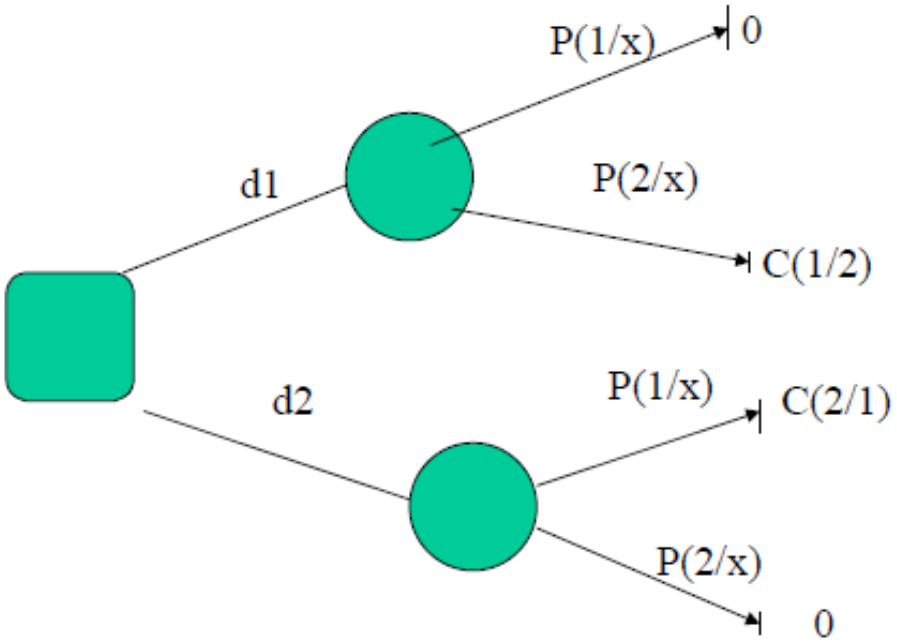

Figura: Classificação entre dois grupos como um problema de decisão

Classificação

Em geral, supomos que as decisões possíveis são dois: Atribuir em P_1 ou em P_2. Uma regra de decisão é uma partição do espaço amostral E_x (que em geral é $\mathcal{R}^p$) em dois regiões A_1 e $A_2 = E_x - A_1$ tais que

$$\text{se} \quad x_0 \in A_1 \Rightarrow d_1 \quad (\text{classificar em } P_1)$$

$$\text{se} \quad x_0 \in A_2 \Rightarrow d_2 \quad (\text{classificar em } P_2)$$

Consequências

Suponhamos que:

- 1. As consequências associadas aos erros de classificação são $c(2|1)$ e $c(1|2)$, onde $c(i|j)$ é o custo de classificar em P_i de uma unidade que pertence a P_j. (Suporemos custos conhecidos)
- 2. O decisor quer maximizar sua utilidade, o que equivale a minimizar o custo esperado

Consequências

A melhor decisão é a que minimiza os custos esperados o funções de perda de oportunidade. Os resultados de cada decisão (figura), teremos as possíveis consequências:

- a) acertar com probabilidade $P(2|x_0)$, nesse caso, não há custo de penalidade.
- b) errar com probabilidade $P(1|x_0)$, nesse caso o custo associado é $c(2|1)$

Custo médio

O custo médio, ou valor esperado da decisão d_2: classificar x_0 em P_2 será:

$$E(d_2) = c(2|1)P(1|x_0) + 0P(2|x_0) = c(2|1)P(1|x_0)$$

análogamente, o custo esperado da decisão d_1: classificar no grupo P_1 é:

$$E(d_1) = 0P(1|x_0) + c(1|2)P(2|x_0) = c(1|2)P(2|x_0)$$

Custo médio

Atribuiremos o elemento ao grupo 2, se seu custo esperado for menor, isto é, se:

$$\frac{f_2(x_0)\pi_2}{c(2|1)} > \frac{f_1(x_0)\pi_1}{c(1|2)}$$

esta condição indica que a igualdade nos outros termos, classificaremos em P_2 se:

- a) Sua probabilidade a priori é maior
- b) a verossimilhança de que x_0 venha de P_2 é maior
- c) o custo de errarmos ao classificar o elemento em P_2 é menor

Outline

AD e Análise de Cluster

- As fronteiras de alguns clusters não são nítidas e a classificação nem sempre é obvia, já que muitos indivíduos podem ser enquadrados em um ou em outro cluster
- Tanto a análise de clusters quanto a análise discriminante se referem a classificação
- Entretanto, a analise discriminante exige o conhecimento prévio da composição do grupo ou cluster para cada objeto ou caso incluídos para então definir uma regra de classificação;
- Em contrapartida, na analise de cluster não há informações a priori sobre a composição do grupo ou cluster para qualquer um de seus objetos. Os grupos ou clusters são sugeridos pelos dados, não definidos a priori.

Dicas

- O número mínimo de observações por variável independente: 5
- Número recomendado de observações por variável: 20
- Número de observações por grupo: o menor grupo deve ter um tamanho que exceda o número de variáveis independentes.
- É recomendável que o número mínimo de casos em cada grupo seja 20, e que os grupos tenham dimensões semelhantes.