応用統計学 2017 第 4 回 判別分析

2017 年 10 月 18 日 (水)

清 智也 sei@mist.i.u-tokyo.ac.jp
http://ur0.pw/yTzt

- 判別分析：マハラノビスの距離，線形判別関数[1]。
- 後日扱う内容：一般化線形モデル，特にロジスティック回帰モデル。
- 本講義で扱った手法を含め，多変量解析の基礎については例えば以下の本が参考になる。
  - 永田靖，棟近雅彦「多変量解析法入門」サイエンス社
- 第 1 回レポートについて：モンテカルロ法，標準誤差[2]。レポート課題は別紙参照。

演習問題

問題 4-1. 以下の 3 変量データからマハラノビス距離を構成したとき，平均ベクトルから最も遠いデータはどれか。ただし $h > 0$ は定数とする。

$$\boldsymbol{X} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ -1 & -1 & h \\ -1 & -1 & -h \end{pmatrix}.$$

問題 4-2. マハラノビスの距離はアフィン変換に関して不変であることを説明せよ。

問題 4-3. 以下の 2 群の 2 変量データ（人工データ）について，これらを判別する線形判別関数を求めよ。

群 1	変量	
	1	2
個体 1	1	0
2	0	1
3	−1	0
4	0	−1

群 2	変量	
	1	2
個体 1	0	0
2	2	0
3	3	0
4	2	1
5	0	2

問題 4-4. フィッシャーのあやめ (iris) データについて web などで調べよ。またそのデータを入手し，3 品種のうちの 2 品種を適当に選んだ上で，それらを判別する線形判別関数を構成せよ。

[1] Discriminant analysis: Mahalanobis distance, linear discriminant function,
[2] Monte Carlo method, standard error.

問題 4-5. モンテカルロ法を用いて，次の事象が起こる確率を推定し，またその標準誤差も推定せよ。

事象：「10 個のサイコロを同時に投げたとき，現れる目の最大度数が 4 となる」

また，モンテカルロ法を用いずに計算するプログラムも書き，結果を確認せよ。

宿題 4

問題 4-6. 前回の宿題で扱った「世界の上位 50 大学の performance breakdown」に対して，アメリカの大学とそれ以外を判別する線形判別関数を求めよ。また，この判別関数を用いて 51 位から 100 位の大学を判別したときの正答率を求めよ。

先週の講義の補足　別紙参照。