

Classificação de Dados de Elevada Dimensão Ignorar ou Incorporar Correlações ?

A. PEDRO DUARTE SILVA

**Faculdade de Economia e Gestão/CEGE
Universidade Católica Portuguesa
Centro Regional do Porto**

(* Supported by: FEDER / POCI 2010



**Ciência.Inovação
2010**

Programa Operacional Ciência e Inovação 2010
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA

Correlações e classificação em grandes dimensões

Índice

1. Classificação com $p \gg n$
2. Métodos de “Análise Discriminante Diagonal”
3. Métodos modernos de selecção de variáveis
 - 3.1. False Discovery e Non-Discovery Rates
 - 3.2. Donoho e Jin’s “Higher Criticism”
4. Como Incorporar Correlações ?
 - 4.1. Thomaz e Gilles’s “Novas FDL”
 - 4.2. Estimadores “encolhidos” e regularizados
 - 4.3. Estimadores baseadas em modelos factoriais
5. Resultados empiricos
6. Conclusões e perspectivas

Correlações e classificação em grandes dimensões

Classificação em grandes dimensões: O problema

$$Y_i; X_i \quad i = 1, 2, \dots, n \quad Y_i \in \{1, \dots, k\}$$

$$X_i \in \mathbb{R}^p \quad p \gg n$$

Pretende-se determinar uma regra capaz de prever Y_i dado X_i

Pressuposto habitual: $\mathbf{X}_i | Y_i \sim \mathbf{N}_p(\boldsymbol{\mu}_{(Y_i)}, \boldsymbol{\Sigma})$

\Rightarrow Regra de Bayes:

$$\begin{aligned} \mathbf{Y}_i &= \arg \min_{\mathbf{g}} (0.5 (\mathbf{X}_i - \boldsymbol{\mu}_{(\mathbf{g})})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i - \boldsymbol{\mu}_{(\mathbf{g})}) - \ln \pi_{\mathbf{g}}) := \\ &\arg \min_{\mathbf{g}} (0.5 \Delta_{i(\mathbf{g})}^T \boldsymbol{\Sigma}^{-1} \Delta_{i(\mathbf{g})} - \ln \pi_{\mathbf{g}}) = \\ &\arg \min_{\mathbf{g}} (0.5 \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{X}_i - \mathbf{X}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{(\mathbf{g})} - \ln \pi_{\mathbf{g}}) \end{aligned}$$

Correlações e classificação em grandes dimensões

Análise Discriminante Diagonal

Naive Bayes

$$\hat{\Delta}_{i(g)} = \mathbf{X}_i - \hat{\boldsymbol{\mu}}_{(g)} = \mathbf{X}_i - \bar{\mathbf{X}}_{(g)} \quad \hat{\Sigma} = \hat{\mathbf{D}} = \text{diag}(\mathbf{S}) \quad \mathbf{S} = \frac{\sum_{g=1}^k \sum_{Y_i=g} (\mathbf{X}_i - \bar{\mathbf{X}}_{(g)}) (\mathbf{X}_i - \bar{\mathbf{X}}_{(g)})^T}{n - k}$$

Nearest Shruken Centroids Tibshirani, Hastie, Narasimhan e Chu (2003)

$$\hat{\boldsymbol{\mu}}_{(g)}^* = \hat{\boldsymbol{\mu}} + \sqrt{\frac{1}{n_g} - \frac{1}{n}} \hat{\mathbf{D}}^{0.5} \mathbf{d}_g^* \quad \mathbf{d}_g^*(\mathbf{j}) = \text{sign}(\mathbf{d}_g(\mathbf{j})) (\mathbf{d}_g(\mathbf{j}) - \alpha)_+$$

$$\mathbf{d}_g = \frac{\hat{\boldsymbol{\mu}}_{(g)} - \hat{\boldsymbol{\mu}}}{\theta_g \sqrt{\frac{1}{n_g} - \frac{1}{n}}} \hat{\mathbf{D}}^{-0.5} \quad \hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{g=1}^k n_g \hat{\boldsymbol{\mu}}_{(g)}$$

$\alpha, \theta_1, \dots, \theta_k$ obtidos por validação cruzada

Correlações e classificação em grandes dimensões

Métodos modernos de seleção de variáveis

Taxas (locais) de falsas descoberta e não descoberta

Dada uma sucessão de p testes e estatísticas ordenadas, z_1, \dots, z_p com

$$P_0 = P(H_0) ; P_1 = P(H_1) \quad f_0(z) ; f_1(z) ; f(z) = P_0 f_0(z) + P_1 f_1(z)$$

Taxa local de falsas descobertas: $\text{fdr}(z) = P_0 f_0(z) / f(z)$

Taxa local de falsas não-descobertas: $\text{fndr}(z) = 1 - \text{fdr}(z)$

Higher Criticism

Dada uma sucessão de p testes e valores de prova, π_1, \dots, π_p ordenados

HC – máxima diferença estandarizada entre π_j e o seu valor esperado se a distribuição de todos os π fosse uniforme

Correlações e classificação em grandes dimensões

Como incorporar correlações ?

Thomaz e Gilles's "Novas FDL"

$$S = \sum_{m=1}^p \lambda_m \mathbf{v}_m \mathbf{v}_m^T \quad \hat{\Sigma} = \sum_{m=1}^p \max(\lambda_m, \bar{\lambda}) \mathbf{v}_m \mathbf{v}_m^T$$

Estimadores "encolhidos" e regularizados

$$\hat{\Sigma} = \rho_1 I_p + \rho_2 S \quad \text{Guo, Hastie e Tibshirani (2007)}$$

Xu, Brock e Parrish (2009)

ou:

$$\Sigma = \mathbf{D}^{0.5} \mathbf{R} \mathbf{D}^{0.5}$$

$$\hat{\mathbf{R}}^* = (1 - \rho_1) \hat{\mathbf{R}}$$

$$\hat{\mathbf{D}}^*(\mathbf{j}, \mathbf{j}) = \rho_2 \text{me}_j(\hat{\mathbf{D}}(\mathbf{j}, \mathbf{j})) + (1 - \rho_2) \hat{\mathbf{D}}(\mathbf{j}, \mathbf{j})$$

$$\hat{\pi}_g^* = \rho_3 \frac{1}{\mathbf{k}} + (1 - \rho_3) \frac{\mathbf{n}_g}{\mathbf{n}}$$

Ahdesmaki e Strimmer (2009)

Correlações e classificação em grandes dimensões

Como incorporar correlações ?

Covariâncias estimadas por modelos factoriais Duarte Silva (2009)

$$\mathbf{X}_i = \mu_{(Y_i)} + \mathbf{B} \mathbf{f}_i + \varepsilon_i \quad \mathbf{f}_i \in \mathcal{R}^q \quad \varepsilon_i \in \mathcal{R}^p \quad q \ll p$$

$$\mathbf{f}_i \sim N_q(\mathbf{0}, \mathbf{I}_q) \quad \varepsilon_i \sim N_p(\mathbf{0}, \mathbf{D}_\varepsilon)$$

$$\Rightarrow \Sigma = \mathbf{B} \mathbf{B}^T + \mathbf{D}_\varepsilon$$

$$\Sigma^{-1} = \mathbf{D}_\varepsilon^{-1} - \mathbf{D}_\varepsilon^{-1} \mathbf{B} [\mathbf{I}_q + \mathbf{B}^T \mathbf{D}_\varepsilon^{-1} \mathbf{B}]^{-1} \mathbf{B}^T \mathbf{D}_\varepsilon^{-1}$$

$$\hat{\Sigma}_{\text{Fctq}} = \hat{\mathbf{B}} \hat{\mathbf{B}}^T + \hat{\mathbf{D}}_\varepsilon ; \hat{\mathbf{B}}, \hat{\mathbf{D}}_\varepsilon = \arg \min_{\hat{\mathbf{B}}, \hat{\mathbf{D}}_\varepsilon} \|\hat{\Sigma}_{\text{Fctq}} - \mathbf{S}\|^2$$

Correlações e classificação em grandes dimensões

Singh's Prostate Cancer Data – p=6033; n=50+52

Rule	Selection Criterion*	Error Estimate	± 2 std errors
AhdesStri	C_HC	0.0561	± 0.0094
AhdesStri	C_FNDR	0.0569	± 0.0058
AhdesStri	I_HC	0.0658	± 0.0105
AhdesStri	I_FNDR	0.0681	± 0.0117
Fctq1	I_HC	0.0635	± 0.0107
Fctq1	I_FNDR	0.0646	± 0.0104
ThomGil NFDL	I_HC	0.0674	± 0.0108
NSC (Pam)	-----	0.0812	± 0.0116
Naïve Bayes	I_HC	0.0668	± 0.0112
Support Vector Machines	I_HC	0.0614	± 0.0107
Fisher's FDL	I_HC	0.2393	± 0.0219

* C – Correlation Adjusted T-scores ; I – Independence based T-scores
 HC – Higher Criticism ; FNDR – False Non-Discovery Rates

Correlações e classificação em grandes dimensões

Simulation Experiment -- Guo, Hastie and Tibshirani (2007) setup

p=10 000; n=100+100 ; 100 Independent Blocks ; $\rho = 0.90$

Rule	Selection Criterion	Error Estimate	± 2 std errors
AhdesStri	C_HC	0.0073	± 0.0012
AhdesStri	C_NFDR	0.0144	± 0.0038
AhdesStri	I_HC	0.0029	± 0.0006
AhdesStri	I_NFDR	0.0075	± 0.0018
Fctq1	I_HC	0.0122	± 0.0020
Fctq1	I_NFDR	0.0196	± 0.0031
ThomGil NFDL	I_HC	0.0055	± 0.0008
NSC (Pam)	-----	0.0149	± 0.0014
Naïve Bayes	I_HC	0.0136	± 0.0019
Support Vector Machines	I_HC	0.0052	± 0.0008
Fisher's FDL	I_HC	0.0435	± 0.0133

Correlações e classificação em grandes dimensões

Simulation Experiment -- Guo, Hastie and Tibshirani (2007) setup

$p=10\ 000$; $n=100+100$; 100 Independent Blocks ; $\rho = 0.99$

Rule	Selection Criterion	Error Estimate	± 2 std errors
AhdesStri	C_HC	0.0033	± 0.0026
AhdesStri	C_NFDR	0.0148	± 0.0049
AhdesStri	I_HC	0.0025	± 0.0011
AhdesStri	I_NFDR	0.0417	± 0.0201
Fctq1	I_HC	0.1124	± 0.0239
Fctq1	I_NFDR	0.1197	± 0.0272
ThomGil NFDL	I_HC	0.0052	± 0.0012
NSC (Pam)	-----	0.0349	± 0.0048
Naïve Bayes	I_HC	0.0781	± 0.0160
Support Vector Machines	I_HC	0.0228	± 0.0071
Fisher's FDL	I_HC	0.0524	± 0.0139

Correlações e classificação em grandes dimensões

Simulation Experiment -- Guo, Hastie and Tibshirani (2007) setup

$p=10\ 000$; $n=100+100$; 10 Independent Blocks ; $\rho = 0.90$

Rule	Selection Criterion	Error Estimate	± 2 std errors
AhdesStri	C_HC	0.4959	± 0.0008
AhdesStri	C_NFDR	0.4942	± 0.0010
AhdesStri	I_HC	0.4973	± 0.0006
AhdesStri	I_NFDR	0.4964	± 0.0009
Fctq1	I_HC	0.4955	± 0.0008
Fctq1	I_NFDR	0.4928	± 0.0011
ThomGil NFDL	I_HC	0.0186	± 0.0037
NSC (Pam)	-----	0.4951	± 0.0009
Naïve Bayes	I_HC	0.4942	± 0.0011
Support Vector Machines	I_HC	0.4963	± 0.0006
Fisher's FDL	I_HC	0.0293	± 0.0131

Correlações e classificação em grandes dimensões

Simulation Experiment -- Guo, Hastie and Tibshirani (2007) setup
 $p=10\ 000$; $n=100+100$; 10 Independent Blocks ; $\rho = 0.99$

Rule	Selection Criterion	Error Estimate	± 2 std errors
AhdesStri	C_HC	0.4878	± 0.0009
AhdesStri	C_NFDR	0.4845	± 0.0013
AhdesStri	I_HC	0.4930	± 0.0010
AhdesStri	I_NFDR	0.4918	± 0.0018
Fctq1	I_HC	0.4759	± 0.0027
Fctq1	I_NFDR	0.4777	± 0.0031
ThomGil NFDL	I_HC	0.0154	± 0.0099
NSC (Pam)	-----	0.4797	± 0.0024
Naïve Bayes	I_HC	0.4790	± 0.0029
Support Vector Machines	I_HC	0.4856	± 0.0016
Fisher's FDL	I_HC	0.0315	± 0.0163

Correlações e classificação em grandes dimensões

Conclusões

- ✓ **A escolha do numero adequado de predictores é critica**
 - ❖ Donoho e Jin's "Higher Criticism" parece produzir os melhores resultados
- ✓ **Podem (e devem-se) incorporar correlações, mesmo com $p \gg n$**
- ✓ **Estimadores de Covariâncias/Correlações baseados em "alvos de referência" são fortemente dependentes da razoabilidade dos alvos adoptados**
- ✓ **Desenfatizar a importância dos ultimos vectores próprios da matriz de covariâncias é uma forma eficaz de regularização para uma grande variedade de condições**

Correlações e classificação em grandes dimensões

Perspectivas e questões em aberto

- ✓ **Em que condições é que se pode confiar em “alvos de referência” ?**
- ✓ **Devem-se incorporar correlações na selecção de variáveis ?**
 - ❖ Quando e Como ?
- ✓ **Qual a importância de regularizar também os estimadores de médias e de probabilidades à priori ?**
- ✓ **Quais as propriedades assintóticas de métodos regularizados ?**
 - ❖ Qual a relevância dessas propriedades ?

Correlações e classificação em grandes dimensões

Referências

Ahdesmaki, P. and Strimmer, K. (2009). Feature selection in "omics" prediction problems using cat scores and non-discovery rate control. *rXiv,stat.AP:0903.2003v1*.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci, USA* 105, 14790-14795.

Duarte Silva, A.P. (2009). Linear Discriminant Analysis with more Variables than Observations. A not so Naïve Approach. To appear In: *Classification as a Tool for Research. Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation*. Dresden, Germany.

Guo, Y., Hastie, T. and Tibshirani, T. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics* 8, 86-100.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9, 303.

Tibshirani, R., Hastie, B., Narismhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids with applications to DNA microarrays. *Statistical Science*, 18, 104-117.

Thomaz, C.E. and Gillies, D.F. (2005). A maximum uncertainty lda-based approach for limited sample size problems with application to face recognition. In: *18th Brazilian Symposium on computer Graphics and Image Processing. SIBGRAPI 2005*, 89-96.

Xu, P., Brock, G.N., and Parrish, R. (2009). Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics and Data Analysis*, 53, 1674-1687.