Modelos estadísticos de Clasificación con alta dimensión en el número de covariables

Laura Freijeiro González

Universidad de Santiago de Compostela

25 de octubre de 2018

Introducción

- Contexto de *Alta Dimensión* o *Big Data*.

Introducción

- Contexto de *Alta Dimensión* o *Big Data*.

p $>$ **n**: mayor número de características que de muestras.

Introducción

- Contexto de *Alta Dimensión* o *Big Data*.

  p $>$ **n**: mayor número de características que de muestras.

- Problemas en las **reglas de clasificación del Análisis Discriminante**.

Introducción

- Contexto de *Alta Dimensión* o *Big Data*.

  p $>$ **n**: mayor número de características que de muestras.

- Problemas en las **reglas de clasificación del Análisis Discriminante**.

- Ejemplo ilustrativo con una base de datos de cáncer
  - Cáncer de ovarios: *National Cancer Institute (NCI)*.
  - Cáncer de próstata: *National Cancer Institute (NCI)* y *Eastern Virginia Medical School (EVMS)*.

Guyon, I. (July 2003) Design of experiments for the NIPS 2003 variable selection benchmark

Índice

1. Reglas usuales del Análisis Discriminante

2. Métodos Alternativos
   - Regresión logística regularizada
   - Algoritmo de K-vecinos más cercanos y algoritmo de K-medias
   - Support Vector Machine (SVM)
   - Aplicación a una base de datos

3. Referencias

El **Análisis Discriminante** persigue

- Buscar las coincidencias y discrepancias entre $L \geq 2$ grupos determinados de antemano para conseguir caracterizarlos.
- Construir **reglas de clasificación** que permitan determinar a qué grupo pertenece una nueva muestra.

Se ilustrará como proceder en el caso de $L = 2$ grupos, mientras que cuando $L > 2$ se puede recurrir a algoritmos como

- **OVA (*one versus all*)**
- **OVO (*one versus one*)** o ***majority voting***

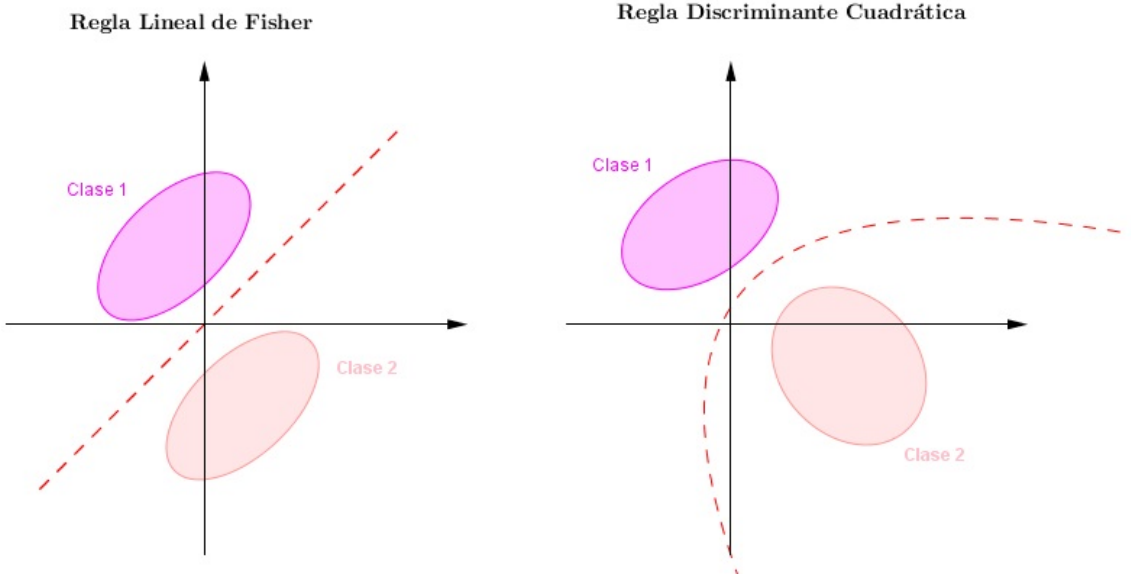
Reglas LDA y QDA

- La **Regla Lineal de Fisher** ($\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$) clasifica en G_1 cuando:

$$\lambda^t \left[x - \tfrac{1}{2}(\mu_1 + \mu_2)\right] > \log\left(\frac{\pi_2}{\pi_1}\right) \quad \text{siendo } \lambda = \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2)$$

- La **Regla Discriminante Cuadrática**:

$$\text{máx}_l\, \delta_l(x) = -\tfrac{1}{2}\log(|\boldsymbol{\Sigma}_l|) - \tfrac{1}{2}(x - \mu_l)^t \boldsymbol{\Sigma}_l^{-1}(x - \mu_l) + \log(\pi_l)$$



En el contexto de $p > n$, la matriz de covarianzas $\boldsymbol{\Sigma}_j$ es singular.

Aplicación a una base de datos

- Datos de cáncer de ovarios y próstata de personas de Virginia y otras partes de Estados Unidos, correspondientes al año 2002.
- Cada una de las variables x_{ij} toma un valor entero mayor o igual que cero $(0, 1, 2, \ldots, 924)$ y la variable respuesta y toma el valor 0 o 1 según si la muestra pertenece a una persona sana o enferma respectivamente.
- $\mathbf{X} \in \mathcal{M}_{100\times 10000}, \quad y \in \mathcal{M}_{100\times 1}.$

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Uso de R

- Se emplea el **porcentaje de clasificación incorrecta** para ver como de bien clasifican.
- Es necesario programar a mano la **versión regularizada de las reglas LDA y QDA**.
- Dificultad en la inversión de las matrices de covarianzas regularizadas: necesidad de recurrir a la **factorización de Cholesky** a través de la función ***chol2inv()***.
- Los tiempos computacionales son grandes.

%Clasif. incorrecta	MÉTODOS	
	LDA regul.	QDA regul.
$\mathbf{G_1}$ $(\mathbf{Y}=\mathbf{0})$	27 %	15 %
$\mathbf{G_2}$ $(\mathbf{Y}=\mathbf{1})$	21 %	28 %
$0{,}56\mathbf{G_1}+0{,}44\mathbf{G_2}$	24,36 %	20,72 %
tiempo (s)	967,03	2583,65

Cuadro: Reglas *LDA* y *QDA* regularizadas, sumando 10^{-7} y 1 respectivamente a los autovalores nulos de $\mathbf{\Sigma}$.

Índice

1 Reglas usuales del Análisis Discriminante

2 Métodos Alternativos
- Regresión logística regularizada
- Algoritmo de K-vecinos más cercanos y algoritmo de K-medias
- Support Vector Machine (SVM)
- Aplicación a una base de datos

3 Referencias

Caso particular del modelo GLM: Regresión logística regularizada

- La **regresión logística** estudia problemas de regresión donde la variable respuesta Y es una **variable binaria (o dicotómica)**, es decir, sólo puede tomar dos valores. Se representarán estos valores por 0 y 1 respectivamente.
- Se construye un modelo para la probabilidad de éxito condicionada, $\pi(x) = \mathbb{P}(\, Y = 1 \mid X = x\,)$, empleando la **función logit**, $g(p) = \log\left(\frac{p}{1-p}\right) \quad \forall p \in [0,1]$:

$$\log\left(\frac{\pi(x,\beta)}{1-\pi(x,\beta)}\right) = x^t\beta.$$

- Para estimar β hay que resolver el problema

$$\max_{\beta}\left\{\sum_{i=1}^{n}\left[y_i(\beta_0+\beta^t x_i) - \log(1+e^{\beta_0+\beta^t x_i})\right]\right\}$$

.

Caso particular del modelo GLM: Regresión logística regularizada

- Se recurre a una **versión regularizada** puesto que la matriz $\mathbf{X}^t\mathbf{V}\mathbf{X}$ del esquema iterativo de estimación de parámetros es singular

$$\beta_{k+1} = \beta_k + (\mathbf{X}^t\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^t(y - \pi(\mathbf{X}, \beta)) \quad k \in \{0, 1, 2, \dots\},$$

- **Regularización Elastic Net**:

$$\max_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta^t x_i) - \log(1 + e^{\beta_0 + \beta^t x_i}) \right] - \lambda \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\|\beta_j\|_2^2 \right) \right\}$$

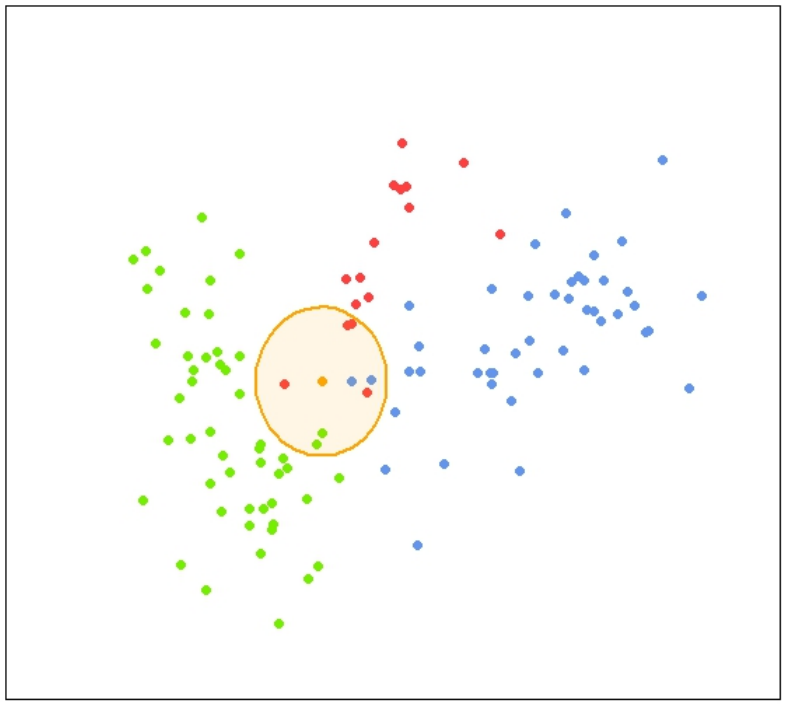
con $\alpha \in [0, 1]$, aplicando una regularización de tipo L_1 cuando $\alpha = 1$ y de tipo L_2 para $\alpha = 0$.

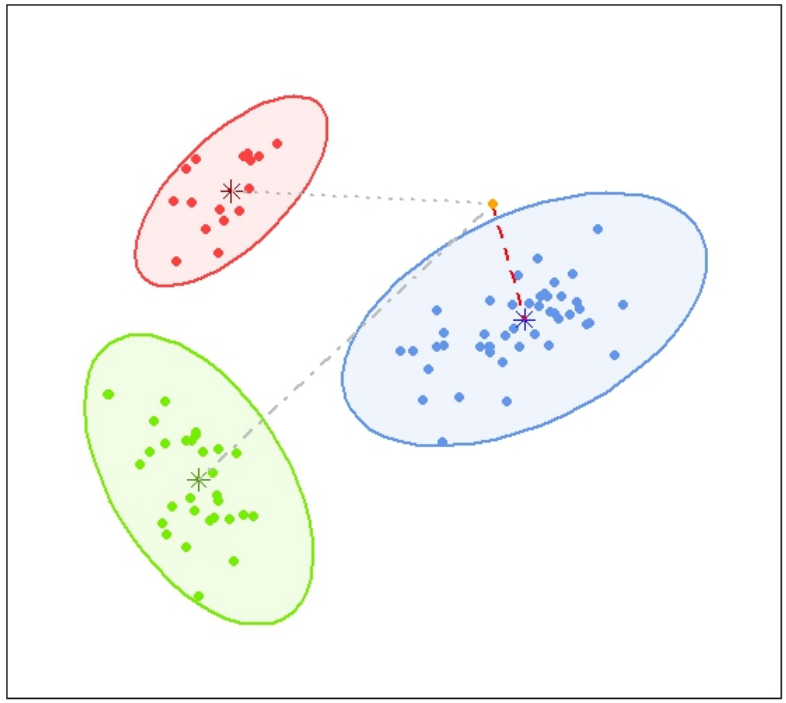
Uso de R

- **glmnet library:** *cv.glmnet(x, y, family="binomial", alpha, weights, offset, lambda, type.measure, nfolds, foldid, grouped, . . .)*
  - Con *alpha* $= 0$ penalización L_2, *alpha* $= 1$ penalización L_1 y *alpha* $\in (0, 1)$ penalización de tipo Elastic Net.
  - Calcula internamente el valor óptimo del factor de penalización λ.
  - Ya tiene implementada la validación cruzada en su algoritmo.
  - Paquete actualizado y mejorado este año

Noah Simon and Jerome Friedman and Trevor Hastie and Rob Tibshirani (2011) *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent*. Journal of Statistical Software, 39(5):1-13.

K-vecinos más cercanos y K-medias





Uso de R

- **class library:** *knn(train, test, cl, k, l, prob, use.all)*
  - Lleva a cabo todo el algoritmo de k-vecinos.
  - Sólo tiene en cuenta la métrica euclídea, no permite emplear métricas adaptadas a cada contexto, como en el caso de la regla **DANN (Discriminant Adaptative Nearest-Neighbour).**

- **stats library:** *kmeans(x, centers, iter.max, nstart, algorithm, trace)*
  - Lleva a cabo todo el algoritmo de k-medias.
  - Sólo tiene en cuenta la métrica euclídea.


W. N. Venables and B. D. Ripley (2002) *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.


R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Método SVM: Support Vector Machine

- **Clases sin solapamiento**

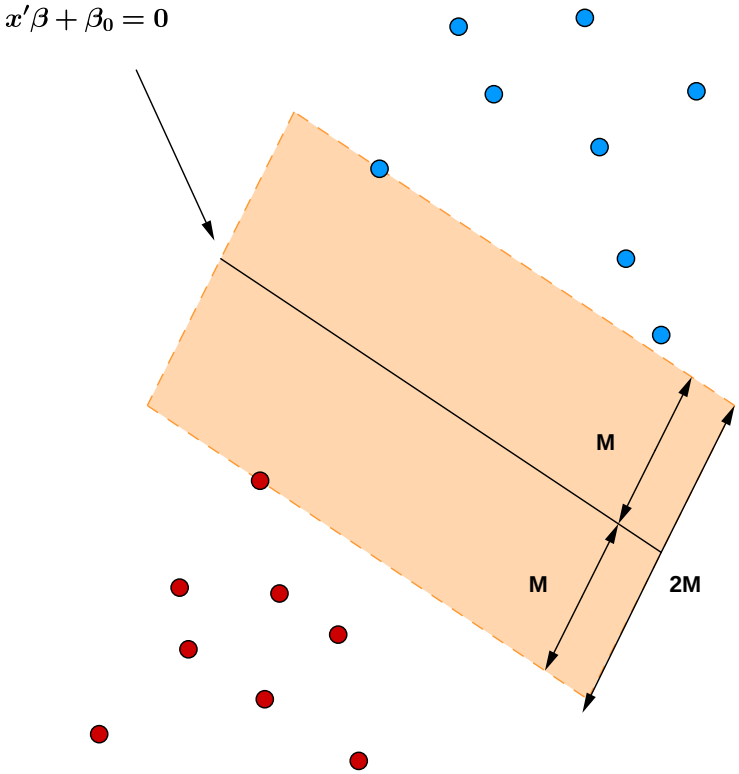
$$\begin{cases} \min\limits_{\beta,\beta_0} \|\beta\| \\ \text{sujeto a } y_i(x_i^t\beta + \beta_0) \geq 1, \quad i = 1, \ldots, n \end{cases}$$

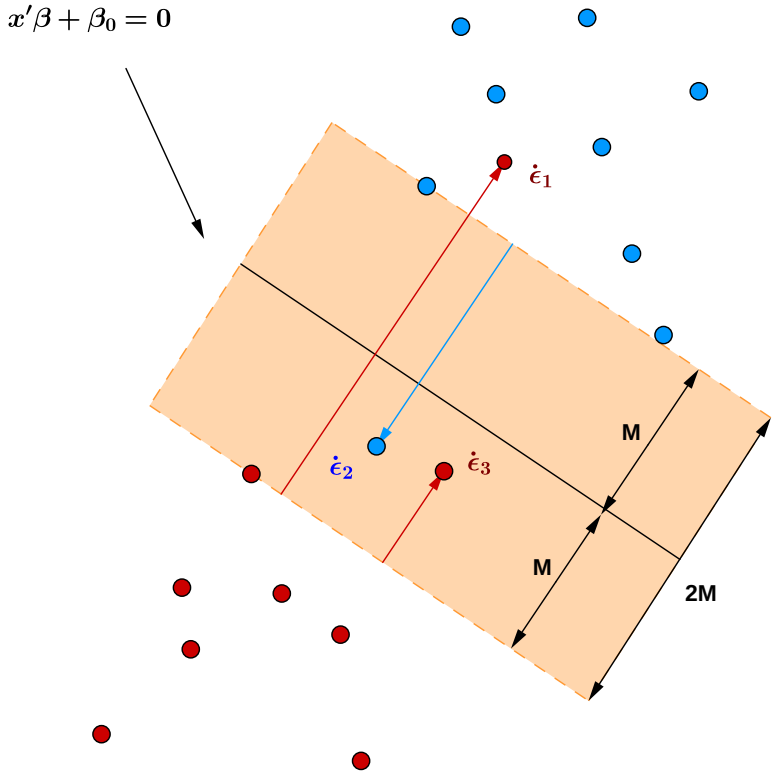
- **Clases con solapamiento**

$$\min \|\beta\| \quad \text{sujeto a} \begin{cases} y_i(x_i^t\beta + \beta_0) \geq 1 - \xi_i \quad \forall i, \\ \xi_i \geq 0, \quad \sum \xi_i \leqslant cte. \end{cases}$$

- **Regla discriminante:**

$$\hat{G}(x) = \text{sign}[\hat{f}(x)] = \text{sign}[x^t\hat{\beta} + \hat{\beta}_0]$$


$x'\beta + \beta_0 = 0$
M
M
2M


$x'\beta + \beta_0 = 0$
$\dot{\epsilon}_1$
$\dot{\epsilon}_2$
$\dot{\epsilon}_3$
M
M
2M

Uso de R

- **e1071 library:** *svm(formula, data, subset, na.action, scale, . . .)*
  - Lleva a cabo todo el algoritmo.
  - Tiene en cuenta fronteras de decisión lineales como no lineales.

David Meyer and Evgenia Dimitriadou and Kurt Hornik and Andreas Weingessel and Friedrich Leisch (2018) *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-0.

Aplicación a una base de datos

- Datos de cáncer de ovarios y próstata de personas de Virginia y otras partes de Estados Unidos, correspondientes al año 2002.
- Cada una de las variables x_{ij} toma un valor entero mayor o igual que cero $(0, 1, 2, \ldots, 924)$ y la variable respuesta y toma el valor 0 o 1 según si la muestra pertenece a una persona sana o enferma.
- $\mathbf{X} \in \mathcal{M}_{100\times 10000}, \quad y \in \mathcal{M}_{100\times 1}$.

Contrastando resultados

%Clasif. incorrecta	MÉTODOS					
	LOG L_2	LOG L_1	LOG E.N. $(1-\alpha)L_2+\alpha L_1$			
			$\alpha=0{,}1$	$\alpha=0{,}2$	$\alpha=0{,}5$	$\alpha=0{,}7$
G_1 $(Y=0)$	10 %	14 %	18 %	15 %	15 %	14 %
G_2 $(Y=1)$	8 %	15 %	10 %	12 %	14 %	14 %
$0{,}56G_1+0{,}44G_2$	9,12 %	14,44 %	14,48 %	13,68 %	14,56 %	14 %
tiempo (s)	24,36	8	11,47	8,58	7,20	7,61

Cuadro: Regresión logística regularizada

%Clasif. incorrecta	MÉTODOS					
	K-MEDIAS	K-VECINOS			SVM	
		$K=1$	$K=2$	$K=5$	LINEAL	POLINOM.
G_1 $(Y=0)$	21 %	4 %	8 %	8 %	6 %	6 %
G_2 $(Y=1)$	10 %	8 %	10 %	10 %	11 %	11 %
$0{,}56G_1+0{,}44G_2$	16,16 %	5,76 %	8,88 %	8,88 %	8,2 %	8,2 %
tiempo (s)	1,11	0,58	1,18	1,77	2,05	1,95

Cuadro: Métodos no paramétricos de clasificación.

Índice

1. Reglas usuales del Análisis Discriminante

2. Métodos Alternativos
   - Regresión logística regularizada
   - Algoritmo de K-vecinos más cercanos y algoritmo de K-medias
   - Support Vector Machine (SVM)
   - Aplicación a una base de datos

3. Referencias

- FRIEDMAN, Jerome; HASTIE Trevor; TIBSHIRANI Robert. The elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer, 2009.
- HASTIE, Trevor; TIBSHIRANI, Robert. Discriminant adaptive nearest neighbor classification and regression. En Advances in Neural Information Processing Systems. 1996. p. 409-415.
- HONG, Zi-Quan; YANG, Jing-Yu. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern recognition, 1991, vol. 24, no 4, p. 317-324.
- McCULLAGH, P.; NELDER, J. Generalized Linear Models. Second Edition. Chapman and Hall, 1989.
- NELDER, John A.; BAKER, R. Jacob. Generalized linear models, 1972.
- VAPNIK, V. N.; LERNER, A. Ya. Recognition of Patterns with help of Generalized Portraits, 1963. Avtomat. i Telemekh, 1963, vol. 24, no 6, p. 774-780.

Gracias por vuestra atención