Memórias Autossociativas de Projeção em Subespaço Baseadas em Máquinas de Vetores de Suporte

Emely Pujólli da Silva
Orientador: Prof. Dr. Marcos E. Valle

Universidade Estadual de Campinas
Instituto de Matemática, Matemática Aplicada e Computação Científica

Campinas, 23 de Abril de 2015

Memórias Associativas

Baseado na capacidade do cérebro humano em armazenar e recordar informações temos as **Memórias Associativas** - AM.

Baseado na capacidade do cérebro humano em armazenar e recordar informações temos as **Memórias Associativas** - AM.

Assim uma memória associativa deve recuperar uma informação corrompida ou incompleta.

Baseado na capacidade do cérebro humano em armazenar e recordar informações temos as **Memórias Associativas** - AM.

Assim uma memória associativa deve recuperar uma informação corrompida ou incompleta.

Exemplo

Escutamos: "É o amor"

Memórias Associativas

Baseado na capacidade do cérebro humano em armazenar e recordar informações temos as **Memórias Associativas** - AM.

Assim uma memória associativa deve recuperar uma informação corrompida ou incompleta.

Exemplo

Escutamos: "É o amor"
Associamos: "É o amooorrrr que mexe com minha cabeça e me deixa assim..."

Formalmente temos

$$\mathcal{F} = \{(\mathbf{x}^{\xi}, \mathbf{y}^{\xi}) / \xi = 1, ..., p\}$$

o conjunto de pares de vetores composto dos itens memorizados (padrão de entrada) e os itens relembrados (padrão de saída) é chamado o **conjunto das memórias fundamentais**

Formalmente temos

$$\mathcal{F} = \{(\mathbf{x}^{\xi}, \mathbf{y}^{\xi})/\xi = 1, ..., p\}$$

o conjunto de pares de vetores composto dos itens memorizados (padrão de entrada) e os itens relembrados (padrão de saída) é chamado o **conjunto das memórias fundamentais**

Uma memória associativa corresponde à uma aplicação $\mathcal{M}$ de A em B espaços vetoriais, tal que

$$\mathcal{M}(\mathbf{x}^{\xi}) \approx \mathbf{y}^{\xi}$$
$$\mathcal{M}(\hat{\mathbf{x}}^{\xi}) \approx \mathbf{y}^{\xi}$$

para $\hat{\mathbf{x}}^{\xi}$ versão ruidosa de $\mathbf{x}^{\xi}$.

Podemos classificar uma memória associativa por:

Podemos classificar uma memória associativa por:

- heteroassociativa quando $\mathbf{y}^{\xi} \neq \mathbf{x}^{\xi}$ ou

Podemos classificar uma memória associativa por:

- heteroassociativa quando $\mathbf{y}^\xi \neq \mathbf{x}^\xi$ ou
- autoassociativa quando $\mathbf{y}^\xi = \mathbf{x}^\xi$.

Seja o conjunto das memórias fundamentais da forma

$$\mathcal{F} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^p\}$$

onde os vetores $\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^p \in \mathbb{R}^n$ são chamados **memórias fundamentais**, e são linearmente independentes.

Memória Autoassociativa

Seja o conjunto das memórias fundamentais da forma

$$\mathcal{F} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^p\}$$

onde os vetores $\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^p \in \mathbb{R}^n$ são chamados **memórias fundamentais**, e são linearmente independentes.

Então buscamos uma aplicação $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^n$ que satisfaça

$$\mathcal{M}(\mathbf{x}^\xi) = \mathbf{x}^\xi$$
$$\mathcal{M}(\hat{\mathbf{x}}^\xi) = \mathbf{x}^\xi$$

para $\hat{\mathbf{x}}^\xi$ versão ruidosa de $\mathbf{x}^\xi$.

Definição

Seja

$$\mathcal{S} = \{\mathbf{s} \in \mathbb{R}^n; \mathcal{M}(\mathbf{s}) = \mathbf{s}\}$$

o conjunto dos pontos fixos da aplicação $\mathcal{M}$. *Uma memória associativa* $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^n$ *é uma memória autoassociativa de projeção (PAM) se*

$$\mathcal{M}(\mathbf{x}) \in \mathcal{S} \quad e \quad d(\mathbf{x}, \mathcal{M}(\mathbf{x})) \leq d(\mathbf{x}, \mathbf{s}), \quad \forall \mathbf{s} \in \mathcal{S} \tag{1}$$

com d *uma certa medida de distância em* $\mathbb{R}^n$.

Definição

Seja

$$\mathcal{S} = \{\mathbf{s} \in \mathbb{R}^n ; \mathcal{M}(\mathbf{s}) = \mathbf{s}\}$$

o conjunto dos pontos fixos da aplicação $\mathcal{M}$. *Uma memória associativa* $\mathcal{M} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ *é uma memória autoassociativa de projeção (PAM) se*

$$\mathcal{M}(\mathbf{x}) \in \mathcal{S} \quad e \quad d(\mathbf{x}, \mathcal{M}(\mathbf{x})) \leq d(\mathbf{x}, \mathbf{s}), \quad \forall \mathbf{s} \in \mathcal{S} \tag{1}$$

com d *uma certa medida de distância em* $\mathbb{R}^n$.

Definição

Se $\mathcal{M} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ *é uma memória autoassociativa de projeção e* $\mathcal{S}$ *subespaço de* $\mathbb{R}^n$, *então dizemos que* $\mathcal{M}$ *é uma memória autoassociativa de projeção em subespaço (SPAM).*

Definição

Seja

$$\mathcal{S} = \{\mathbf{s} \in \mathbb{R}^n; \mathcal{M}(\mathbf{s}) = \mathbf{s}\}$$

o conjunto dos pontos fixos da aplicação $\mathcal{M}$. *Uma memória associativa* $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^n$ *é uma memória autoassociativa de projeção (PAM) se*

$$\mathcal{M}(\mathbf{x}) \in \mathcal{S} \ \text{ e } \ d(\mathbf{x}, \mathcal{M}(\mathbf{x})) \leq d(\mathbf{x}, \mathbf{s}), \quad \forall \mathbf{s} \in \mathcal{S} \tag{1}$$

com d *uma certa medida de distância em* $\mathbb{R}^n$.

Definição

Se $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^n$ *é uma memória autoassociativa de projeção e* $\mathcal{S}$ *subespaço de* $\mathbb{R}^n$, *então dizemos que* $\mathcal{M}$ *é uma memória autoassociativa de projeção em subespaço (SPAM).*

Note que $\mathfrak{F} \subseteq \mathcal{S}$ para uma memória autoassociativa.

Vamos estudar algumas memórias, para isso faremos a seguinte hipótese:

Vamos estudar algumas memórias, para isso faremos a seguinte hipótese:

Suponhamos $\mathcal{S}$ o subespaço gerado pelas memórias fundamentais, isto é $\mathcal{S} = [\mathfrak{F}]$.

Vamos estudar algumas memórias, para isso faremos a seguinte hipótese:

Suponhamos $\mathcal{S}$ o subespaço gerado pelas memórias fundamentais, isto é $\mathcal{S} = [\mathfrak{F}]$.

Dado o padrão de entrada $\mathbf{x} \in \mathbb{R}^n$ sua saída $\mathcal{M}(\mathbf{x}) \in \mathcal{S}$ é combinação linear dos elementos de $\mathfrak{F}$ isto é, existem $\alpha_1, \alpha_2, ..., \alpha_p \in \mathbb{R}$ escalares tais que

$$\mathcal{M}(\mathbf{x}) = \alpha_1 \mathbf{x}^1 + \alpha_2 \mathbf{x}^2 + ... + \alpha_p \mathbf{x}^p$$

Uma maneira de descrever esse problema é como um problema de minimização

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} d(\mathbf{x}, X\boldsymbol{\alpha}).$$

Uma maneira de descrever esse problema é como um problema de minimização

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} d(\mathbf{x}, X\boldsymbol{\alpha}).$$

Em forma matricial

$$\mathbf{x} \approx X\boldsymbol{\alpha}$$

com $X = [\mathbf{x}^1, ..., \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ a matriz de posto completo denominada *matriz das memórias fundamentais* e $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_p]$ o vetor dos coeficientes.

Estimadores de Regressão

Em um problema de Regressão Multilinear procuramos estimar $f : \mathcal{X} \to \mathcal{Y}$ no ponto x dado a observação y, ou seja $y \approx f(x)$.

Em um problema de Regressão Multilinear procuramos estimar $f : \mathcal{X} \to \mathcal{Y}$ no ponto x dado a observação y, ou seja $y \approx f(x)$.

Em outras palavras, buscamos $f(x)$ de modo que o erro ε na expressão

$$y = f(x) + \varepsilon \tag{2}$$

se torne o menor possível.

Em um problema de Regressão Multilinear procuramos estimar $f : \mathcal{X} \to \mathcal{Y}$ no ponto x dado a observação y, ou seja $y \approx f(x)$.

Em outras palavras, buscamos $f(x)$ de modo que o erro ε na expressão

$$y = f(x) + \varepsilon \tag{2}$$

se torne o menor possível.

Uma maneira de resolver (2) é utilizar uma função perda para descrever quanto o erro irá influenciar na estimativa de $f(x)$.

Definição

Denote por $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ *a tripla consistindo do padrão* x*, uma observação* y *e uma previsão* $f(x)$*. Então a aplicação* $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ *com a propriedade* $\rho(x, y, y) = 0$ *para todo* $x \in \mathcal{X}$ *e* $y \in \mathcal{Y}$ *é chamada de função perda.*

Definição

Denote por $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ *a tripla consistindo do padrão* x*, uma observação* y *e uma previsão* $f(x)$*. Então a aplicação* $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ *com a propriedade* $\rho(x, y, y) = 0$ *para todo* $x \in \mathcal{X}$ *e* $y \in \mathcal{Y}$ *é chamada de função perda.*

Os algoritmos de regressão são estruturados de acordo com o risco. Ou seja, a solução do problema é encontrada minimizando alguma forma do funcional risco.

Definição

Denote por $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ *a tripla consistindo do padrão* x*, uma observação* y *e uma previsão* $f(x)$*. Então a aplicação* $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ *com a propriedade* $\rho(x, y, y) = 0$ *para todo* $x \in \mathcal{X}$ *e* $y \in \mathcal{Y}$ *é chamada de função perda.*

Os algoritmos de regressão são estruturados de acordo com o risco. Ou seja, a solução do problema é encontrada minimizando alguma forma do funcional risco. Estimadores minimizam o risco empírico.

Definição

Denote por $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ *a tripla consistindo do padrão* x*, uma observação* y *e uma previsão* $f(x)$*. Então a aplicação* $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ *com a propriedade* $\rho(x, y, y) = 0$ *para todo* $x \in \mathcal{X}$ *e* $y \in \mathcal{Y}$ *é chamada de função perda.*

Os algoritmos de regressão são estruturados de acordo com o risco. Ou seja, a solução do problema é encontrada minimizando alguma forma do funcional risco. Estimadores minimizam o risco empírico.

Definição

O risco empírico é definido como

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i, f(x_i))$$

No caso da regressão, a função perda toma a forma particular

$$\rho(x, y, f(x)) = \tilde{\rho}(f(x) - y)$$

com $\tilde{\rho} : \mathbb{R} \to [0, \infty)$.

No caso da regressão, a função perda toma a forma particular

$$\rho(x, y, f(x)) = \tilde{\rho}(f(x) - y)$$

com $\tilde{\rho} : \mathbb{R} \to [0, \infty)$.

Então um Estimador de regressão resolve o seguinte problema

$$\min \sum_{i=1}^{n} \rho(f(x_i) - y_i).$$

No caso da regressão, a função perda toma a forma particular

$$\rho(x, y, f(x)) = \tilde{\rho}(f(x) - y)$$

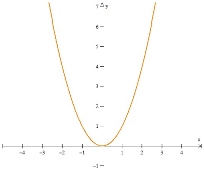
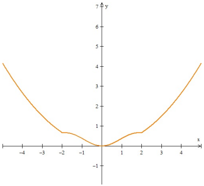
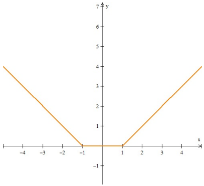
com $\tilde{\rho} : \mathbb{R} \to [0, \infty)$.

Então um Estimador de regressão resolve o seguinte problema

$$\min \sum_{i=1}^{n} \rho(f(x_i) - y_i).$$

Cada estimador resolve o problema de regressão colocando hipóteses sobre a função perda ρ.

Exemplos de funções perda

Perda Quadrada	Função perda de Tukey	Função perda $\varepsilon-$insensível
$\rho_Q(u) = (u)^2$	$\rho_T(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2k^2} + \frac{u^6}{6k^4} & \text{se } \lvert u\rvert \leq k \\ \frac{u^2}{6} & \text{se } \lvert u\rvert > k \end{cases}$	$\rho_\varepsilon(u) = \begin{cases} \lvert u\rvert - \varepsilon & \text{se } \lvert u\rvert \geq \varepsilon \\ 0 & \text{c. c.} \end{cases}$
		

Colocando em termos das memórias autoassociativas

$$\arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho \left(\sum_{j=1}^{p} X_{ij}\alpha_j - x_i \right)$$

Resolver o problema de regressão com estimadores, resulta na solução do problema de memórias autoassociativas com a hipótese de que $\mathcal{S}$ é subespaço gerado pelas memórias fundamentais.

Seja o caso em que $\rho(u) = u^2$ ou seja, a função perda toma a forma da métrica euclideana

$$\boldsymbol{\alpha}^* = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} ||\mathbf{x} - X\boldsymbol{\alpha}||_2$$

que é o problema de quadrados mínimos, cuja a solução de norma mínima pode ser obtido por

$$\boldsymbol{\alpha}^* = X^\dagger \mathbf{x}$$

onde $X^\dagger$ representa a pseudo-inversa de X.

O padrão pode ser recuperado utilizando a expressão

$$\mathcal{M}(\mathbf{x}) = X\boldsymbol{\alpha}^*$$

Este é o modelo introduzido por Kohonen e Ruohonen, chamado **Memória Associativa Linear Ótima** - OLAM.

O padrão pode ser recuperado utilizando a expressão

$$\mathcal{M}(\mathbf{x}) = X\boldsymbol{\alpha}^*$$

Este é o modelo introduzido por Kohonen e Ruohonen, chamado **Memória Associativa Linear Ótima** - OLAM.

Ainda, $\mathcal{M}(\mathbf{x}) = M_0\mathbf{x}$ com $M_0 = XX^\dagger$ chamada matriz peso sináptico, o que faz da aplicação $\mathcal{M}$ linear.

Apesar da simplicidade da solução de quadrados mínimos esta tem sido criticada por sua falta de robusteza, ou seja, insensibilidade à pequenas variações nas hipóteses.

Apesar da simplicidade da solução de quadrados mínimos esta tem sido criticada por sua falta de robusteza, ou seja, insensibilidade à pequenas variações nas hipóteses.

Uma melhora pode ser obtida ao substituir a aproximação de quadrados mínimos por um estimador robusto e elegante para o problema de regressão multilinear $X\boldsymbol{\alpha} \approx \mathbf{x}$.

Apesar da simplicidade da solução de quadrados mínimos esta tem sido criticada por sua falta de robusteza, ou seja, insensibilidade à pequenas variações nas hipóteses.

Uma melhora pode ser obtida ao substituir a aproximação de quadrados mínimos por um estimador robusto e elegante para o problema de regressão multilinear $X\boldsymbol{\alpha} \approx \mathbf{x}$.

A $\varepsilon-$regressão de vetor de suporte tem como objetivo encontrar o $\boldsymbol{\alpha}$ tal que $X\boldsymbol{\alpha}$ tenha no máximo $\boldsymbol{\varepsilon}$ de diferença do padrão desejado $\mathbf{x}$ e ao mesmo tempo, ser o mais plana possível.

Apesar da simplicidade da solução de quadrados mínimos esta tem sido criticada por sua falta de robusteza, ou seja, insensibilidade à pequenas variações nas hipóteses.

Uma melhora pode ser obtida ao substituir a aproximação de quadrados mínimos por um estimador robusto e elegante para o problema de regressão multilinear $X\boldsymbol{\alpha} \approx \mathbf{x}$.

A $\varepsilon-$regressão de vetor de suporte tem como objetivo encontrar o $\boldsymbol{\alpha}$ tal que $X\boldsymbol{\alpha}$ tenha no máximo $\boldsymbol{\varepsilon}$ de diferença do padrão desejado $\mathbf{x}$ e ao mesmo tempo, ser o mais plana possível.

A planitude neste caso, significa que procuramos um $\boldsymbol{\alpha}$ com magnitude pequena.

O funcional risco tem a forma

$$R[\boldsymbol{\alpha}; X, \mathbf{x}] = \frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_{i=1}^{n}\rho_{\boldsymbol{\varepsilon}}\left(\sum_{j=1}^{p}X_{ij}\alpha_j - x_i\right) \tag{3}$$

onde $C > 0$ é a constante que determina a troca entra a planitude e a quantidade de desvios maiores que ε são permitidos.

O funcional risco tem a forma

$$R[\boldsymbol{\alpha}; X, \mathbf{x}] = \frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_{i=1}^{n}\rho_{\boldsymbol{\varepsilon}}\left(\sum_{j=1}^{p}X_{ij}\alpha_j - x_i\right) \tag{3}$$

onde $C > 0$ é a constante que determina a troca entra a planitude e a quantidade de desvios maiores que ε são permitidos.

Para minimizar o risco (3) colocamos na forma de um problema de otimização convexa, e para permitir alguns erros incluimos variáveis de folga.

O funcional risco tem a forma

$$R[\boldsymbol{\alpha}; X, \mathbf{x}] = \frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_{i=1}^{n}\rho_{\boldsymbol{\varepsilon}}\left(\sum_{j=1}^{p}X_{ij}\alpha_j - x_i\right) \tag{3}$$

onde $C > 0$ é a constante que determina a troca entra a planitude e a quantidade de desvios maiores que ε são permitidos.

Para minimizar o risco (3) colocamos na forma de um problema de otimização convexa, e para permitir alguns erros incluimos variáveis de folga.

$$\begin{array}{ll} \text{minimizar} & \frac{1}{2}||\boldsymbol{\alpha}||^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) \\ \text{sujeito à} & \mathbf{x} - X\boldsymbol{\alpha} \leq \varepsilon + \boldsymbol{\xi}^+ \\ & X\boldsymbol{\alpha} - \mathbf{x} \leq \varepsilon + \boldsymbol{\xi}^- \\ & \boldsymbol{\xi}^+, \boldsymbol{\xi}^- \geq 0 \end{array} \tag{4}$$

Muitas vezes, o problema (4) é mais fácil ser resolvido na sua forma dual. A idéia é contruir a Função de Lagrange utilizando a função objetivo e colocando pesos nas restrições. Assim

Muitas vezes, o problema (4) é mais fácil ser resolvido na sua forma dual. A idéia é contruir a Função de Lagrange utilizando a função objetivo e colocando pesos nas restrições. Assim

$$
\begin{aligned}
L(\boldsymbol{\alpha},\lambda^+,\lambda^-,\eta^+,\eta^-) \quad = \quad & \frac{1}{2}\sum_{j=1}^{p}\boldsymbol{\alpha}_j^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) \\
& -\sum_{i=1}^{n}\lambda_i^+(\varepsilon - y_i + \xi_i^+ + \sum_{j=1}^{p}X_{ij}\alpha_j) \\
& -\sum_{i=1}^{n}\lambda_i^-(\varepsilon + y_i + \xi_i^- - \sum_{j=1}^{p}X_{ij}\alpha_j) \\
& -\sum_{i=1}^{n}(\eta_i^+\xi_i^+ + \eta_i^-\xi_i^-)
\end{aligned}
\qquad (5)
$$

Aqui $\lambda^-,\lambda^+,\eta^-,\eta^+ \geq 0$ são os multiplicadores de Lagrange.

É possível verificar que essa função tem ponto de sela com respeito às variáveis primais e duais na solução (SMOLA, SCHOLKOPF).

É possível verificar que essa função tem ponto de sela com respeito às variáveis primais e duais na solução (SMOLA, SCHOLKOPF).

Derivamos a função de Lagrange com respeito à α_k, ξ^+ e ξ^-, obtendo as expressões

$$\alpha_k = \sum_{i=1}^{n} (\lambda_i^+ - \lambda_i^-) X_{ik}$$
$$\eta_k^+ = C - \lambda_k^+$$
$$\eta_k^- = C - \lambda_k^-$$

E das duas últimas expressões, obtemos ainda

$$C - \lambda_k^+ = \eta_k^+ \geq 0 \Leftrightarrow \lambda_k^+ \leq C$$
$$C - \lambda_k^- = \eta_k^- \geq 0 \Leftrightarrow \lambda_k^- \leq C, \quad \forall k = 1, ..., n.$$

Encontramos o seguinte problema dual

$$\begin{aligned}
\text{Maximizar} \quad & -\frac{1}{2}\sum_{j=1}^{p}\left[\sum_{i=1}^{n}(\lambda_i^{+}-\lambda_i^{-})X_{ij}\right]^2 \\
& -\sum_{i=1}^{n}(\lambda_i^{+}+\lambda_i^{-})\varepsilon+\sum_{i=1}^{n}(\lambda_i^{+}-\lambda_i^{-})y_i \qquad (6)\\
\text{sujeito à} \quad & 0 \leq \boldsymbol{\lambda}^{+} \leq C \\
& 0 \leq \boldsymbol{\lambda}^{-} \leq C
\end{aligned}$$

E a solução do problema

$$\boldsymbol{\alpha}^* = X^T(\boldsymbol{\lambda}^+ - \boldsymbol{\lambda}^-)$$

conhecida como expansão do vetor de suporte e determina a estimativa $\boldsymbol{\alpha}^*$ em função dos multiplicadores de Lagrange.

E a solução do problema

$$\boldsymbol{\alpha}^* = X^T(\boldsymbol{\lambda}^+ - \boldsymbol{\lambda}^-)$$

conhecida como expansão do vetor de suporte e determina a estimativa $\boldsymbol{\alpha}^*$ em função dos multiplicadores de Lagrange.

Dessa maneira o padrão recordado toma a forma

$$\mathcal{M}(\mathbf{x}) = X\boldsymbol{\alpha}^*$$

Experimentos computacionais

Próxima apresentação...

Obrigada!