Carlos A. Silva

PATTERN RECOGNITION

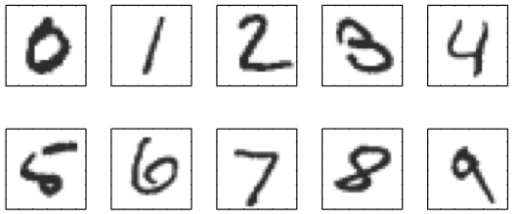
Introdução: Motivação

Pattern Recognition: Motivação

- Aplicações:
  - No século 16, Kepler usou observações do astrónomo Tycho Brahe para descobrir empiricamente as leis que governam o movimento dos planetas.
  - Detecção do uso indevido do cartão de crédito.
  - Como tomar decisões na venda e compra de ações ?
  - Detecção de sequências no DNA indicativas de doença.
  - ...
- Confluência de diversas áreas:
  - Estatística (inferência/predição).
  - Machine Learning (Informática).
  - Reconhecimento de Padrões (Electrónica, Processamento de Sinal).
  - ...

Carlos A. Silva

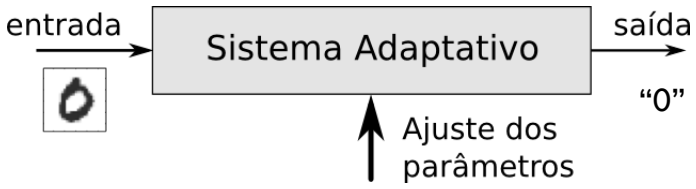
Pattern Recognition: Motivação



- Caso de estudo:
  - Pretendemos desenvolver um sistema que receba uma imagem (28x28) e que como saída indique o digito representado na imagem.
    - Problema de elevada complexidade: a forma dos caracteres variam de pessoa para pessoa.

Carlos A. Silva

Pattern Recognition: Motivação

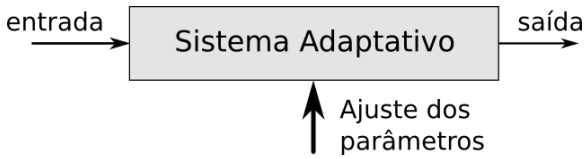


- A imagem (28x28) é decomposta em colunas que são agrupadas (*stacked*) formando o vector de entrada do sistema reconhecedor.

Carlos A. Silva

Pattern Recognition: Nomenclatura

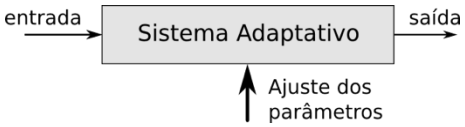
- Estrutura do sistema:



  - **Problema**: Como ajustar os parâmetros do modelo adaptativo de forma a efetuar com sucesso o reconhecimento dos dígitos ?
- Nomenclatura:
  - Variável de entrada (*input*): valores pré-definidos ou medidos.
    - Também chamada variável independente. Terá influência sobre uma ou mais variáveis de saída.
    - Em estatística é chamada de variável preditora (*predictor*). Em *pattern recognition* é chamada de *feature*.

Carlos A. Silva

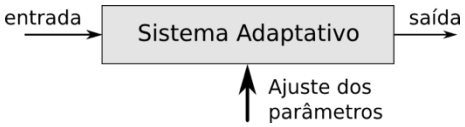
Pattern Recognition: Nomenclatura



- Variável de saída (*output*):
  - Também chamada variável dependente ou resposta.
- Tipo de variável de saída:
  - Medida quantitativa:
    - Uma variável de saída é do tipo quantitativa quando algumas medidas são maiores do que outras e medidas com valores próximos são próximo na sua natureza.
  - Medida qualitativa:
    - A variável assume valores num conjunto finito. P. ex., definimos a variável dígito como assumindo valores no conjunto: {0, 1, 2, ..., 9}.

Carlos A. Silva

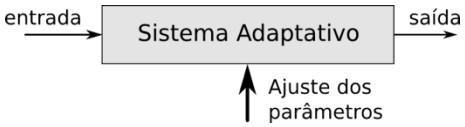
Pattern Recognition: Nomenclatura



- Não existe qualquer ordem explícita nas classes.
  - *Nota: quando a ordem é importante, s = {quente, morno e frio}, então dizemos que estas são do tipo qualitativa (ou categórica) ordinal (não falaremos deste caso).*
- Frequentemente usamos etiquetas (*labels*) para as classes, embora estas sejam codificadas posteriormente, onde o código usado deve ser escolhido de forma a maximizar a estabilidade do algoritmo.

Carlos A. Silva

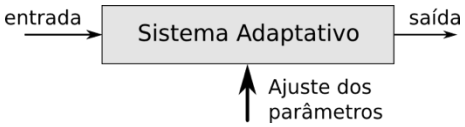
Pattern Recognition: Nomenclatura



- Processo de Aprendizagem:
  - Conjunto de treino (*training set*):
    - Conjunto de vectores de entrada, $x = \{x_1, x_2, ..., x_N\}$, usados no ajuste dos parâmetros do meu modelo adaptativo.
    - Ex., no caso dos dígitos seria um conjunto de imagens de dígitos com repetições do mesmo dígito escrito por diferentes pessoas em circunstâncias diferentes (depende da generalidade pretendida do modelo).
  - Fase de treino ou aprendizagem (*Training/Learning phase*):
    - Nesta fase a função g(t): x(t) -> y(t) é determinada com base nos dados do conjunto de treino. *Equivalente a dizer: determinação dos parâmetros do sistema adaptativo.*

Carlos A. Silva

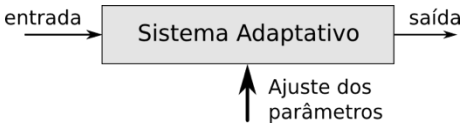
Pattern Recognition: Nomenclatura



- Fase de teste (*test phase*):
  - Após a fase de treino, um novo conjunto de dados, chamado de *test set*, é usado para aferir a qualidade do sistema.
  - O conjunto de teste deve conter entradas que não foram usadas durante a fase de treino.
- Generalização:
  - Propriedade do sistema de reconhecimento de padrões.
  - Esta propriedade indica a capacidade do sistema reconhecer/ categorizar corretamente entradas que não foram usadas na fase de treino; portanto, entradas não consideradas na aprendizagem.

Carlos A. Silva

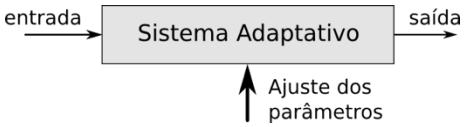
Pattern Recognition: Nomenclatura



- Pré-processamento:
  - Tipicamente as variáveis de entrada são transformadas num novo espaço de variáveis que esperamos seja mais efetivo para a tarefa de classificação.
    - No casos dos dígitos, este pré-processamento poderia consistir em deslocar a imagem e alterar a dimensão da imagem de forma a que todos os exemplos estejam com dimensão idêntica.
    - Este tratamento permitiria reduzir a variabilidade existente nos meus dados de entrada.
  - As variáveis de entrada depois de transformada são as vezes chamadas de características (*features*) e a operação de pré-processamento de extração de características (*feature extraction*).

Carlos A. Silva

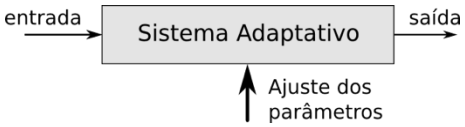
Pattern Recognition: Nomenclatura



- Tipos de Aprendizagem:
  - Aprendizagem supervisionada (*Supervised Learning*):
    - Quando no processo de treino usamos pares conhecidos de vectores de entrada e de vectores de saída (chamado, *target vector*) para determinar os parâmetros do sistema adaptativo, então dizemos que temos um processo de aprendizagem supervisionada.
    - No caso do dígito teríamos um vector de entrada (*features*), x, que caracterizaria o dígito e um vector de saída, y, com um elemento, uma *label*, que indicaria o dígito (elemento do conjunto da classe).

Carlos A. Silva

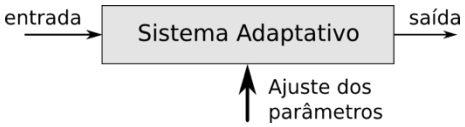
Pattern Recognition: Nomenclatura



- Aprendizagem não supervisionada (*Unsupervised Learning*):
  - Quando temos um conjunto de vectores de entrada sem a respectiva saída alvo durante a fase de treino dizemos então que temos uma aprendizagem *não supervisionada*.
  - A aprendizagem não supervisionada tem como objectivo:
    - Descobrir grupo de elementos semelhantes nos dados de entrada.
    - Descobrir a distribuição dos dados de entrada (*density estimation*)

Carlos A. Silva

Pattern Recognition: Nomenclatura



- Tipos de Problemas:
  - Quando a minha variável de saída é qualitativa dizemos então que temos um *problema de classificação*.
  - Quando a minha variável de saída é quantitativa dizemos que temos um *problema de regressão*.

Carlos A. Silva

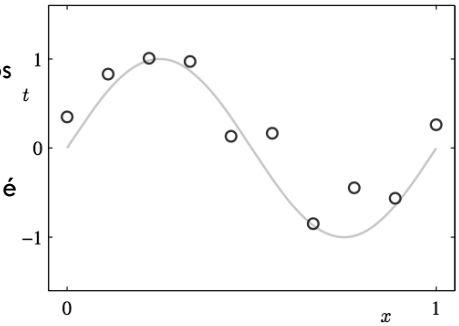
Pattern Recognition: Regressão

- Método dos Mínimos Quadrados (ou Quadrados Mínimos Ordinários):
  - Cenário de estudo:
    - Observamos uma variável real de entrada, x.
    - Com estas observações pretendemos estimar os valores da variável alvo, t.
    - Após o treino pretende-se ser capaz de estimar $\hat{t}$ com base na entrada $\hat{x}$
    - Suponhamos que foi-nos dado um conjunto de N observações da entrada, $x = (x_1, x_2, ..., x_N)^T$, e o respectivo vector de observações do valor alvo $t = (t_1, t_2, ..., t_N)^T$.

Carlos A. Silva

Pattern Recognition: Regressão

- Dados:
  - Os pontos azuis representam os valores das observações dos pares (xi, t_i).
  - Cada amostra de observação é composta por duas componentes.



  - A primeira componente representa o valor real da grandeza subjacente ao problema e a segunda componente ao ruído.
  - Esta segunda componente por representar erros de medida, ou ainda uma componente latente não observável do processo.

Carlos A. Silva

Pattern Recognition: Regressão

- Modelos lineares:
  - Na predição dos valores de saída vamos usar um modelo linear descrito pela seguinte equação polinomial:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

    - **Questão**: *Como o modelo pode ser linear se é descrito por uma equação polinomial ?*
  - A equação polinomial acima é não linear em x, mas é uma função linear dos parâmetros w.

Carlos A. Silva

Pattern Recognition: Regressão

- Determinação dos parâmetros do modelo (*Model fitting*):
  - Os parâmetros do modelo podem ser determinados com base nos dados através da minimização de uma função de custo:

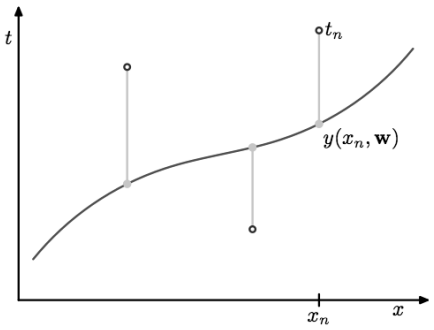
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

  - Esta função de custo é sempre positiva sendo nula apenas quando os valores das predições coincidem com as observações.
  - Resolvemos o problema da adaptação do modelo pela escolha do vector w que minimize a função de custo.
  - Contudo, temos que determinar de igual modo o parâmetro N (grau do polinómio). A este problema chamamos seleção do modelo (*model selection*).

Carlos A. Silva

Pattern Recognition: Regressão

- Minimizamos a soma do quadrado do erro entre as observações do valores alvo e das estimativas.
- Os coeficientes w são obtidos derivando a equação função de custo e estudando os seus zeros:



$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

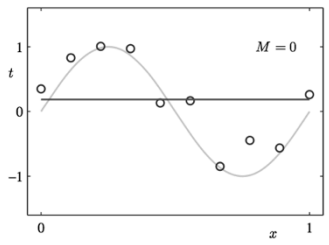
Carlos A. Silva

Pattern Recognition: Regressão

- Modelos lineares:
  - Quão crítico é a ordem no modelo no erro das estimativas ?
  - Qual será o efeito do número de observações sobre o erro das estimativas e sobre a seleção da ordem do modelo ?

Carlos A. Silva

Pattern Recognition: Regressão



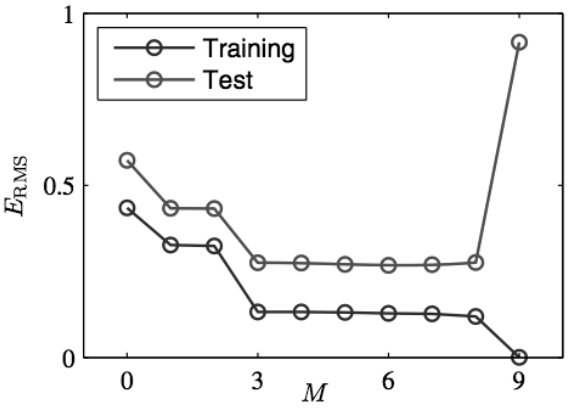
- A este problema chamamos de over-fitting.
- Este acontece quando o modelo se adapta tanto aos dados que perde a capacidade de generalizar para novos casos.

Carlos A. Silva

Pattern Recognition: Regressão

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$

- A raiz quadrada permite que o erro RMS seja medido na mesma escala da variável alvo.
- A divisão por N permite que comparar erros com *data sets* de dimensões diferentes.



Carlos A. Silva

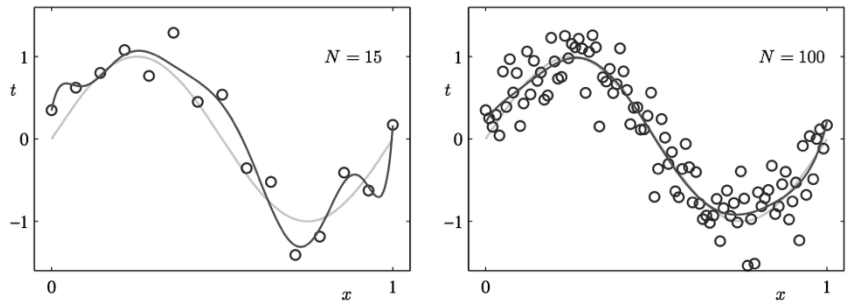
Pattern Recognition: Regressão

- O que acontece aos parâmetros do modelo a medida que a ordem aumenta ?
- Tendo em conta o valor dos parâmetros, porque o erro aumenta ?

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Carlos A. Silva

Pattern Recognition: Regressão



- Nos dois casos acima, a ordem dos polinómios é M= 9 e o número de observações é N = 100.
- **O que podemos concluir** sobre a relação entre o número de observações, a ordem do modelo e o problema de *over-fitting* ?

Carlos A. Silva

Pattern Recognition: Regressão

- Regularização:
  - Podemos minimizar o efeito do *over-fitting*, rescrevendo a função de custo de modo a penalizar soluções para w de valor elevado:

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2,$$

  - A esta técnica denominamos de regularização. No caso do regularizador quadrático chamamos de *ridge regression.*

Carlos A. Silva

Pattern Recognition: Regressão

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

- A resolução da equação acima é semelhante à resolução do seguinte sistema de equações:

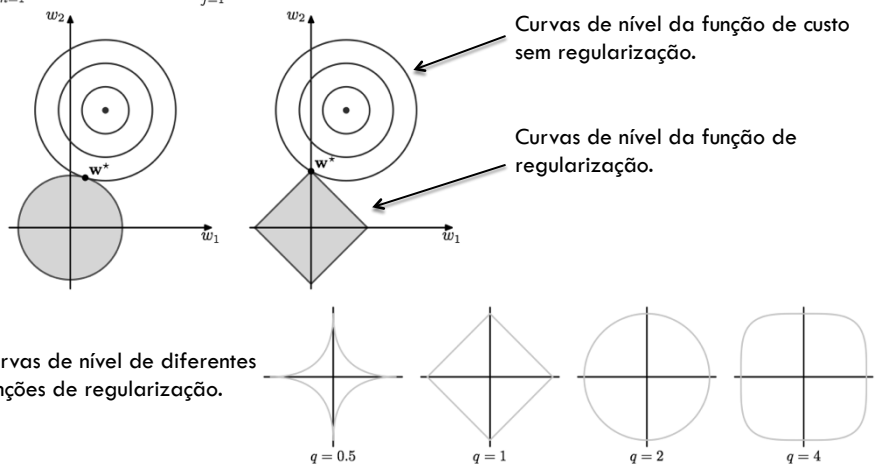
$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2.$$

$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta$$

Carlos A. Silva

Pattern Recognition: Regressão

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



Curvas de nível da função de custo sem regularização.

Curvas de nível da função de regularização.

Curvas de nível de diferentes funções de regularização.

Carlos A. Silva

Pattern Recognition: Regressão

- Caso em que M = 9 para N = 10.
- O caso em que $\lambda = 0$ corresponde a não existir regularização.

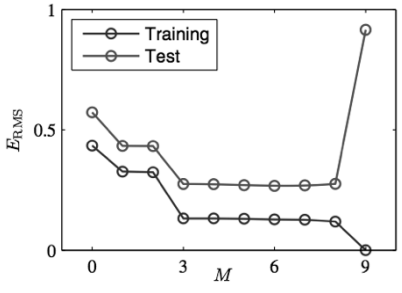
	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

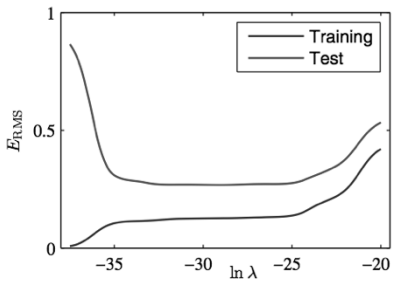
Carlos A. Silva

Pattern Recognition: Regressão

Erro sem regularização



Erro com regularização



Carlos A. Silva

Pattern Recognition: Regressão

- Lasso:
  - Podemos usar um regularizador de primeira ordem l^1:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

  - A esta técnica denominamos de *Lasso*.

Carlos A. Silva

Pattern Recognition: Regressão

- Elastic Net:
  - Podemos de igual modo combinar ambos os regularizadores numa única função de custo:

$$\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \left[\alpha ||\beta||_2^2 + (1 - \alpha) ||\beta||_1 \right].$$

  - A esta técnica denominamos de *Elastic Net*.

Carlos A. Silva

Pattern Recognition: Regressão

- Exercício:
  - Estude o problema da regressão para os quatro métodos estudados para o caso do sinal sinusoidal com M = 10.