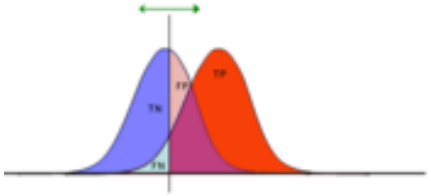
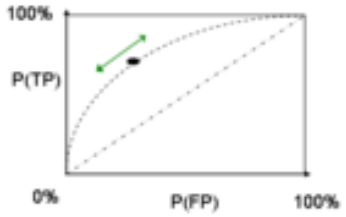
Análise e Comparação de Desempenho

André E. Lazzaretti

UTFPR/CPGEI



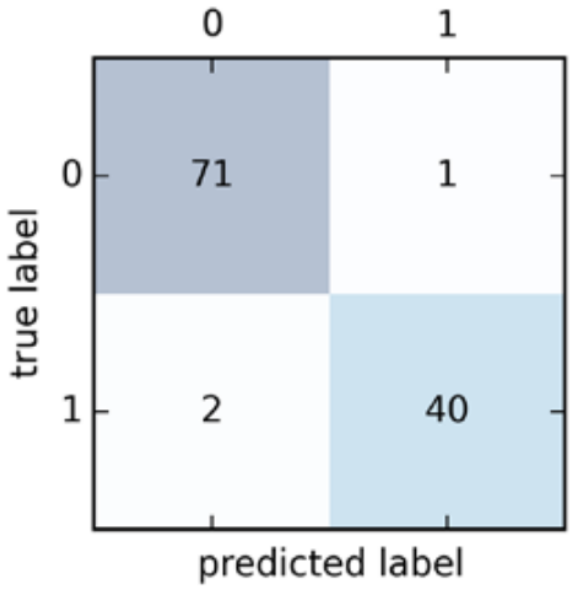
TP	FP
FN	TN
1	1



Matriz de Confusão

- Supondo que você esteja trabalhando com um problema de classificação de duas classes e tem em mãos $\mathbf{y}_{pred}$ e $\mathbf{y}_{true}$:

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)



Erro, acurácia, taxas TP e FP, *precision*, *recall*, f1:

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} = 1 - ERR$$

$$ERR = \frac{FP + FN}{FP + FN + TP + TN}$$

$$TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

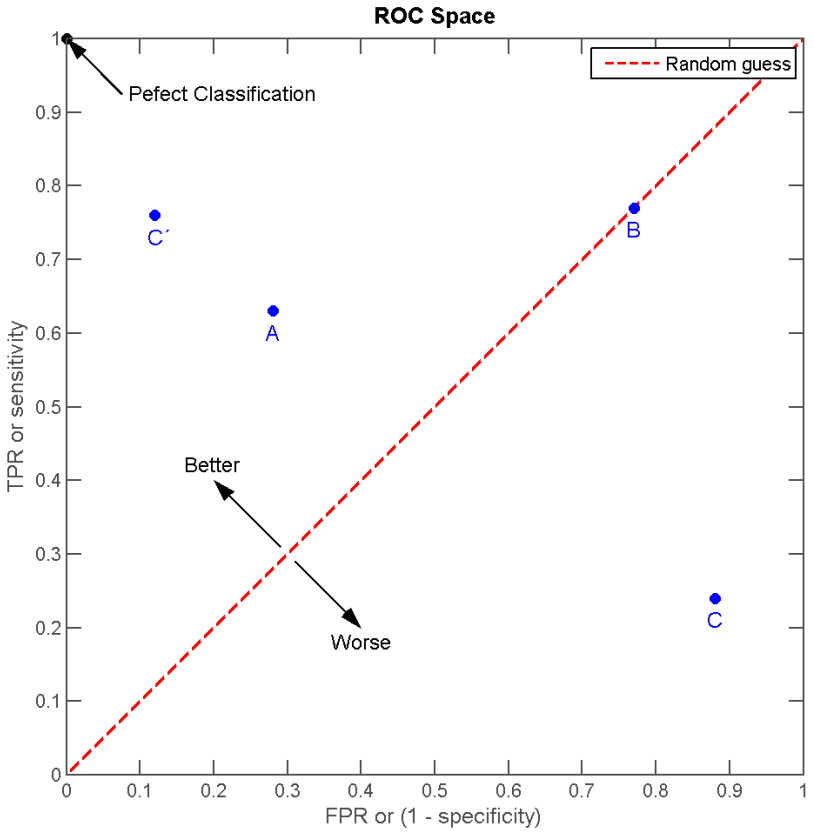
$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

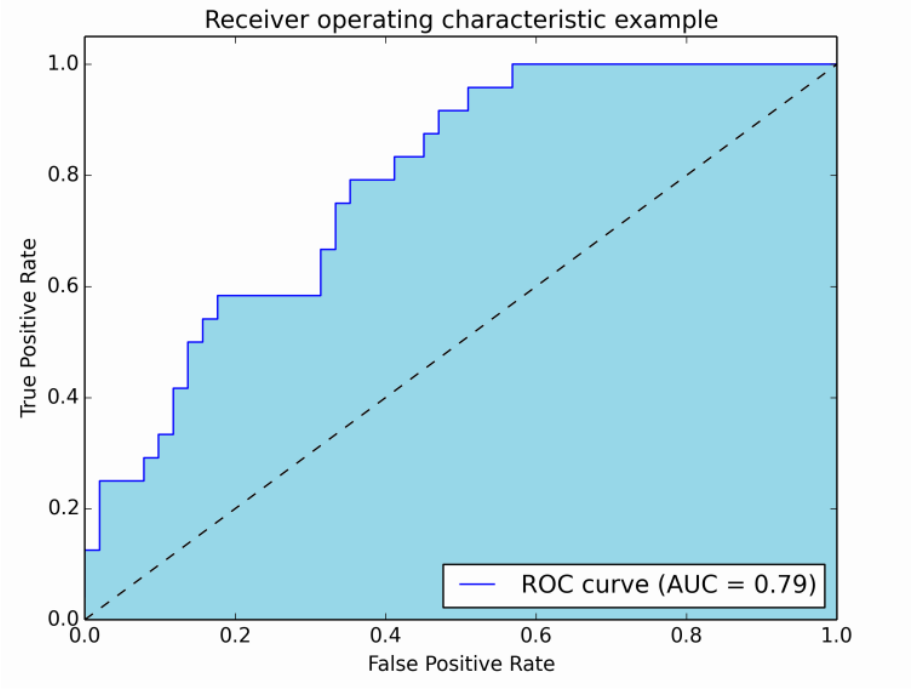
$$PRE = \frac{TP}{TP + FP}$$

$$F1 = 2\frac{PRE \times REC}{PRE + REC}$$

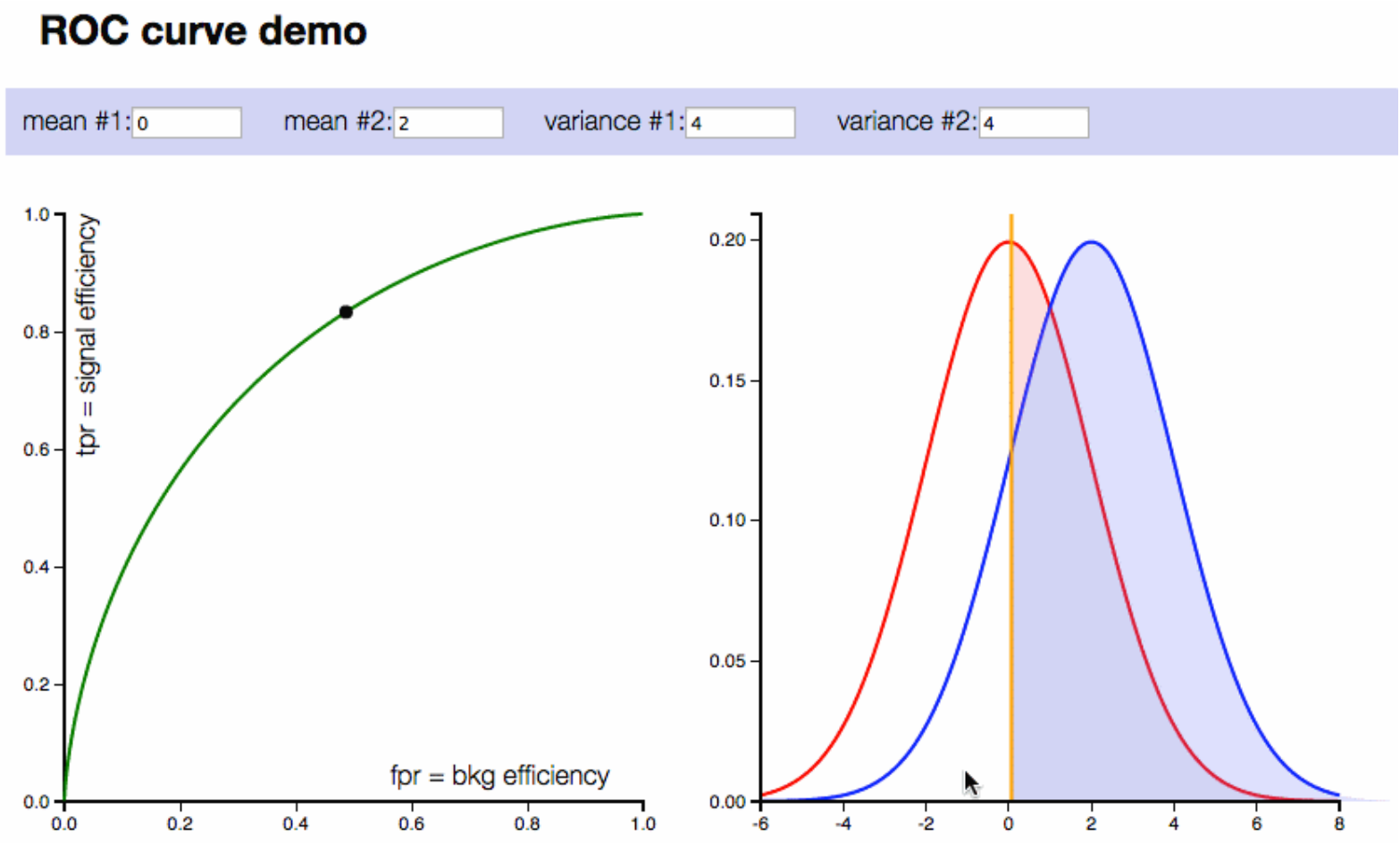
Tudo isso pode ser expandido para multiclass!
Cuidado com a acurácia no caso desbalanceado!

Receiver Operating Curve (ROC)





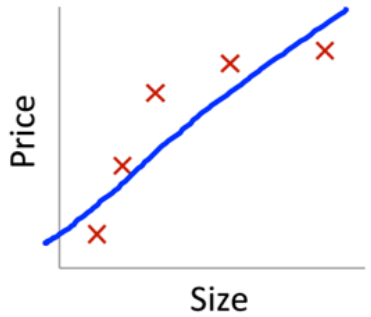
Receiver Operating Curve (ROC)



Mínimos Quadrados - Regressão

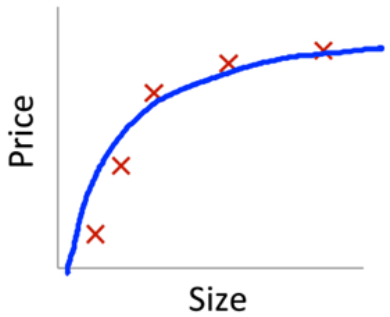
- Regressão: minimizar o seguinte funcional em relação à θ:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$



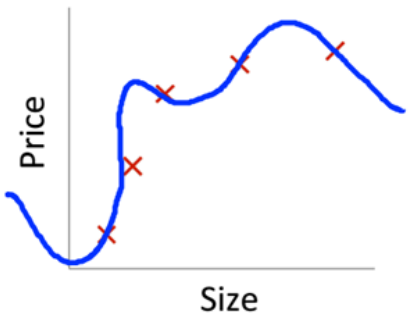
h_θ $\theta_0 + \theta_1 x$

High bias (underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

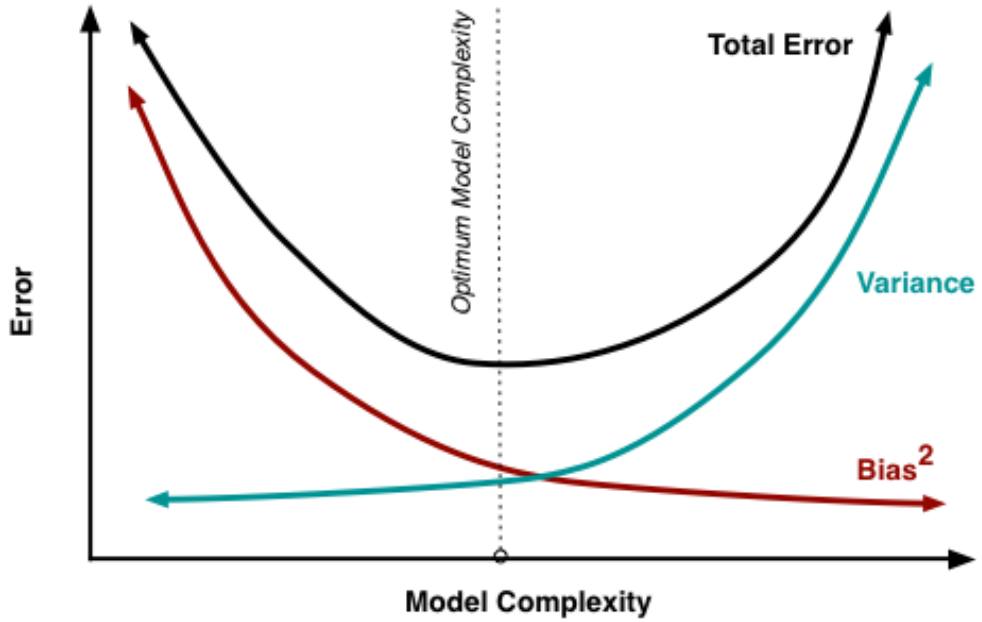


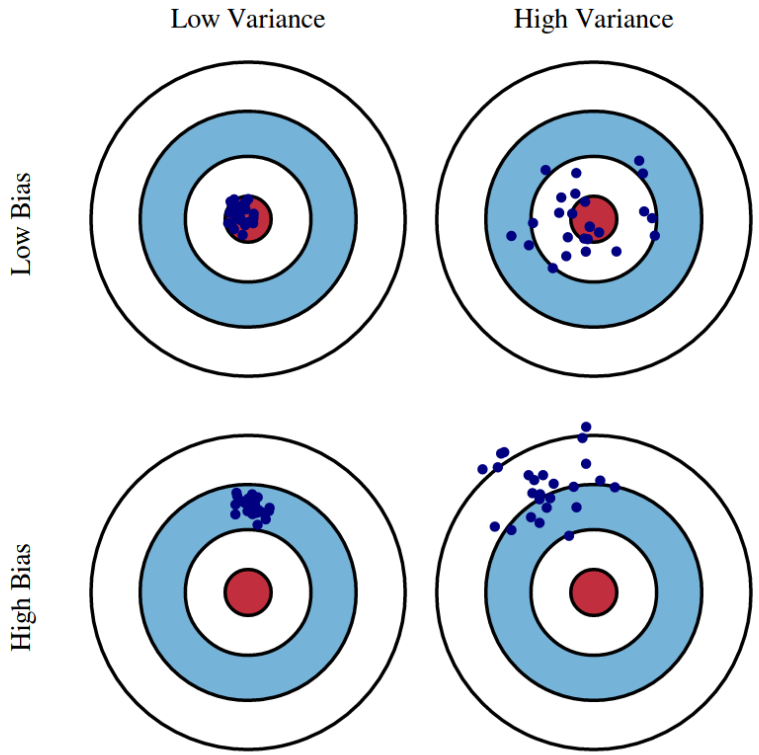
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance (overfit)

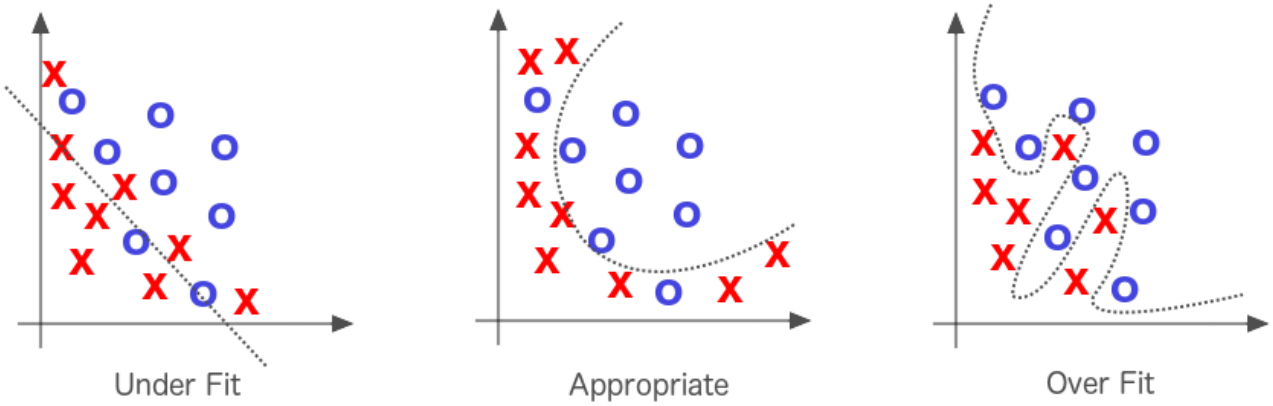
Aulas Prof. Andrew Ng

Bias x Variância

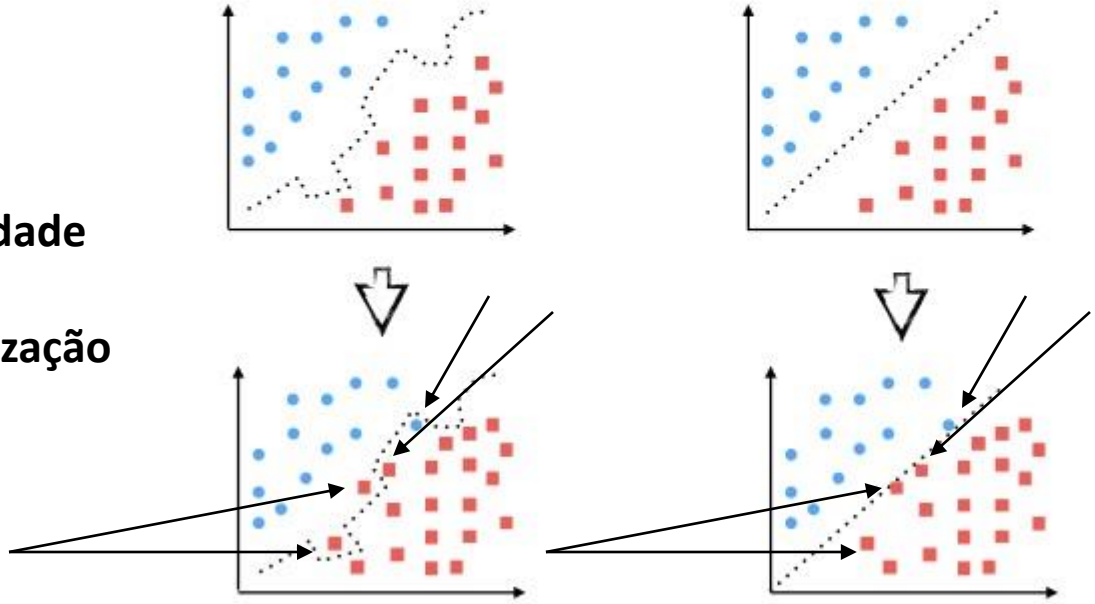




Overfitting - Classificação



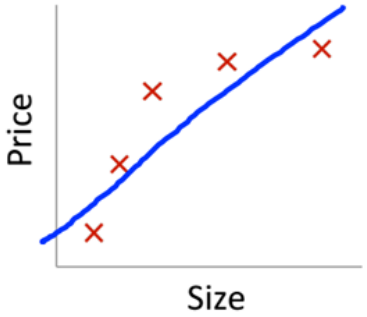
Capacidade de Generalização



Regularização

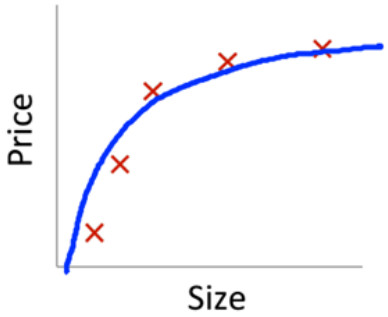
- Sem conhecimento prévio do que se classifica como “verdadeiro”, é possível julgar qual modelo é melhor?
- Na prática não é possível, porém pode-se utilizar um modelo de penalização de modelos excessivamente complexos.
- Isso normalmente se chama **regularização**.

Regularização



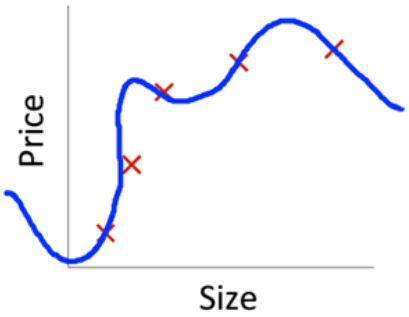
h_θ $\theta_0 + \theta_1 x$

High bias
(underfit)



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"



Eliminar!

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

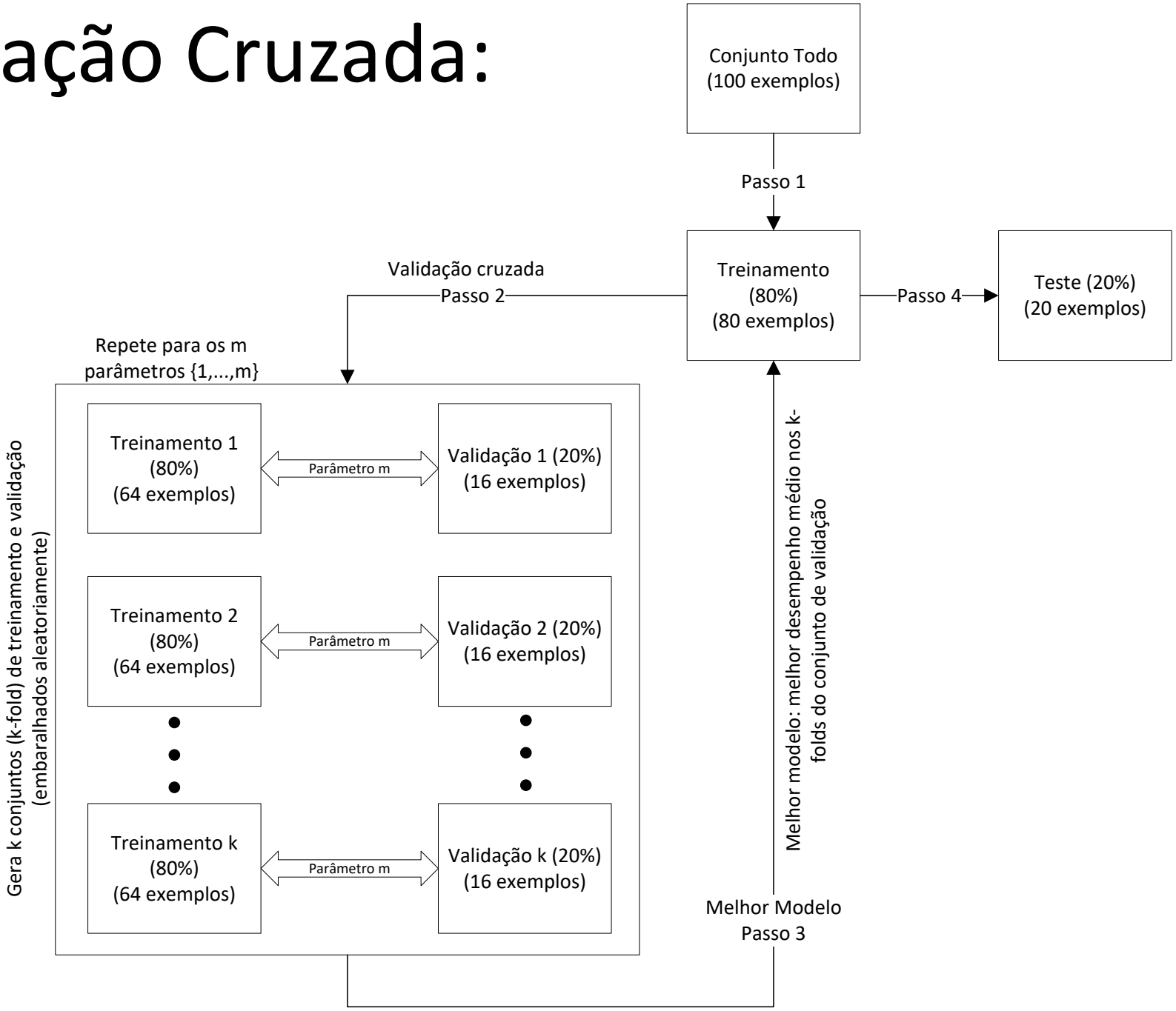
High variance
(overfit)

Alternativa: alterar a minimização do seguinte funcional em relação à θ:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$$

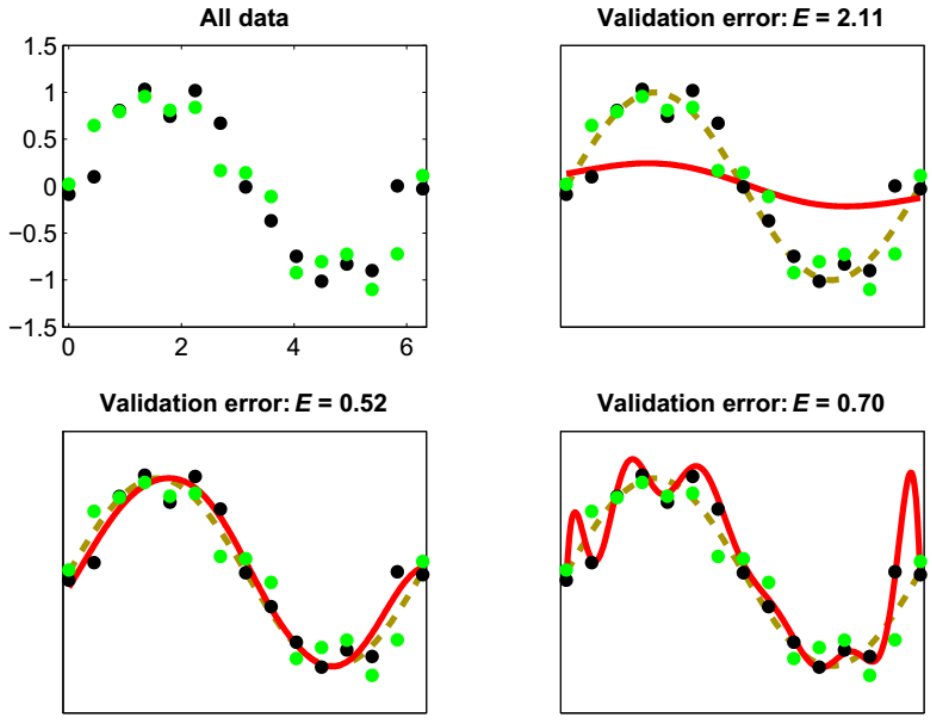
Aulas Prof. Andrew Ng

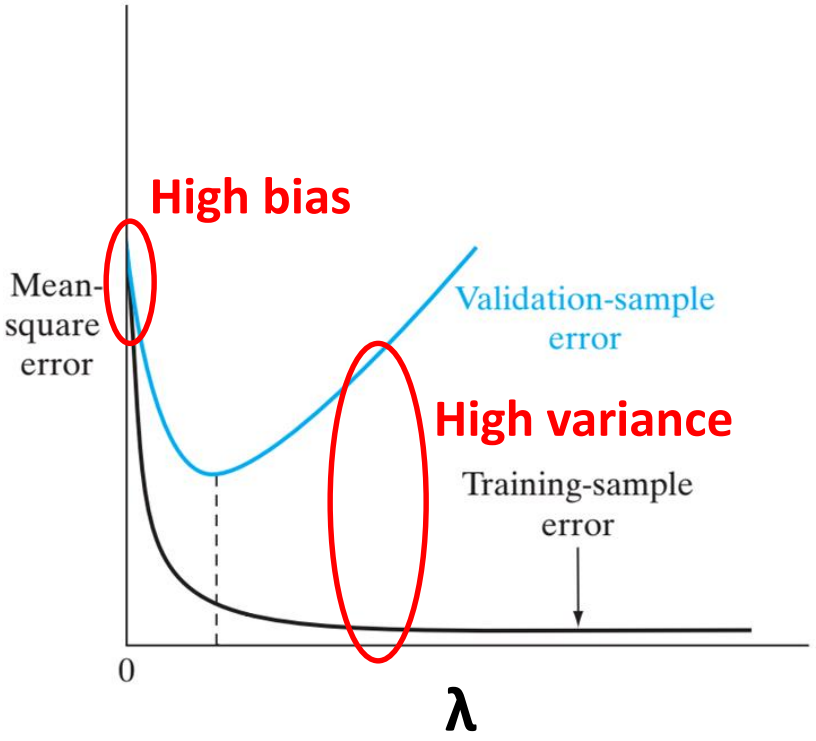
Validação Cruzada:



Regularização

Selecionar λ com o menor erro no conjunto de validação:

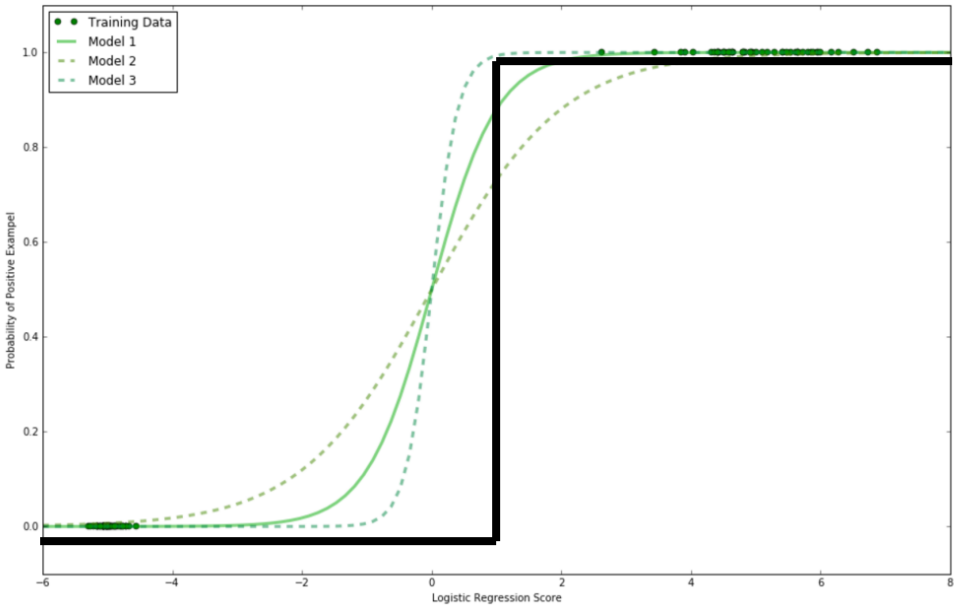




Complexidade (ordem do modelo)

Regularização – Logistic Regression

- Por ser um problema de otimização baseado em likelihood – overfitting!
- Caso linear: no limite pode ser uma função degrau, localizada em diferentes pontos (limiares).



- Regularização quadrática:

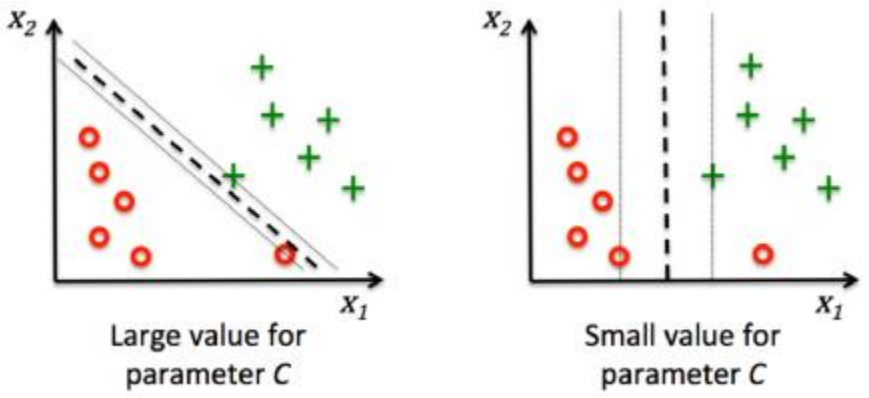
$$
\begin{aligned}
f'(\mathbf{w}) &= \mathrm{NLL}(\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w} \\
\mathbf{g}'(\mathbf{w}) &= \mathbf{g}(\mathbf{w}) + \lambda \mathbf{w} \\
\mathbf{H}'(\mathbf{w}) &= \mathbf{H}(\mathbf{w}) + \lambda \mathbf{I}
\end{aligned}
$$

- Ou Bayesian Logistic Regression!

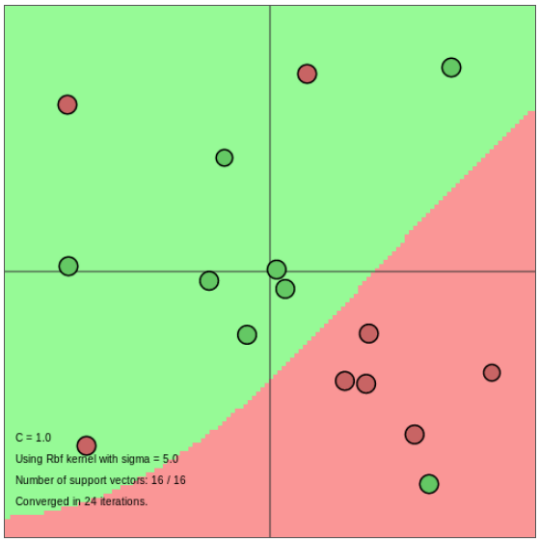
Regularização - SVM

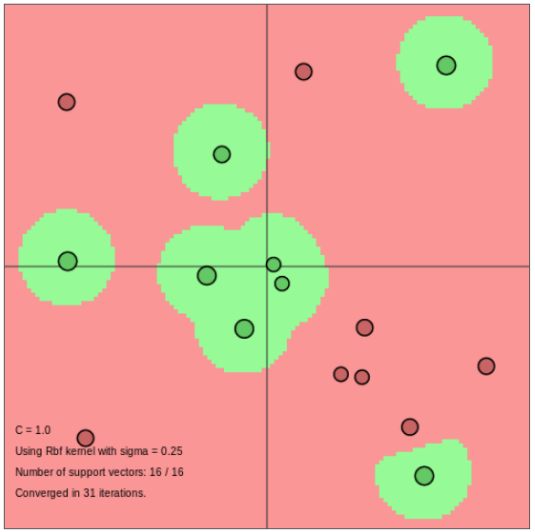
- Da aula anterior: $\text{minimize} \quad J(\boldsymbol{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$

Influência do parâmetro C:



O que isso significa? Relação com regularização!

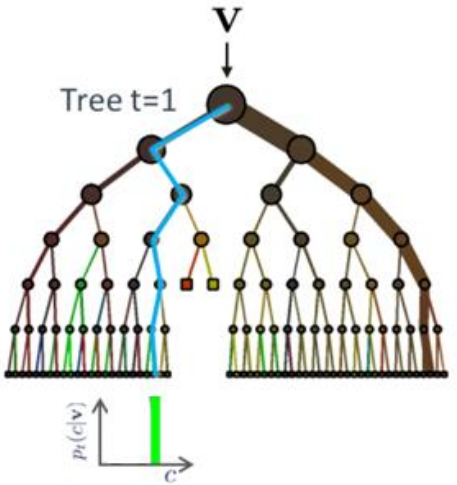


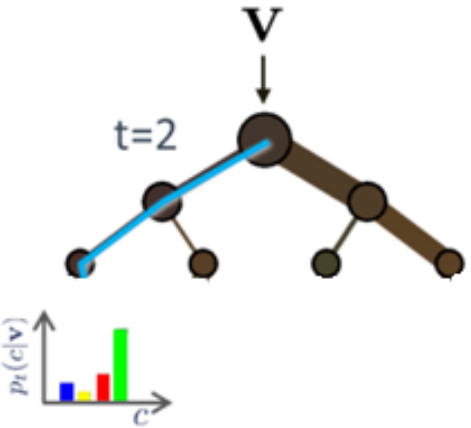


- Maior o sigma → mais "suave" o classificador → reduz overfitting → classificador global
- Menor o sigma → mais "sharp" o classificador → overfitting → classificador local

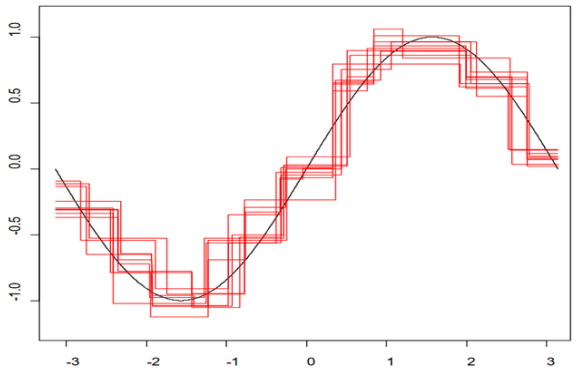
Regularização – *Decision Tree*

- Limitando máximo comprimento das árvores:





- Ensembles – Random Forest:

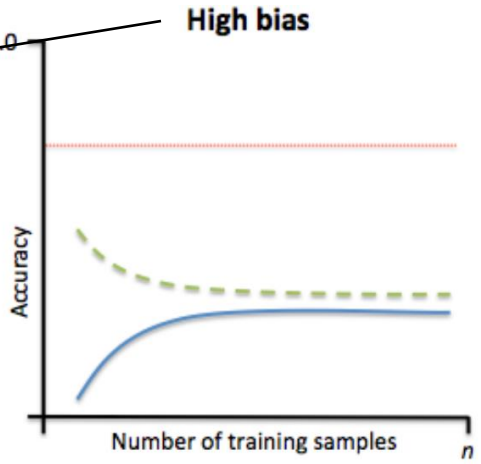


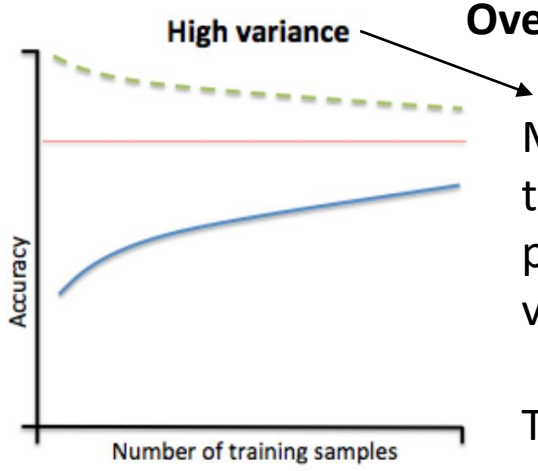
Análise de Curvas

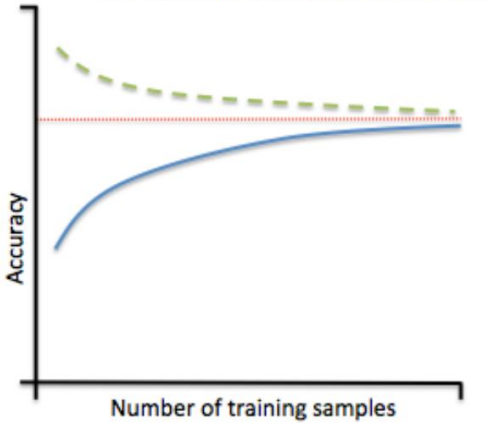
Mais dados de treinamento provavelmente não vão ajudar significativamente

Tente adicionar novas *features!*

Verificar a validação cruzada. Aumentar *range* das variáveis.







Mais dados de treinamento provavelmente vão ajudar!

Tente reduzir núm. de *features*

Verificar a validação cruzada. Diminuir *range* das variáveis.

Outras Dicas

- Se você é um especialista no problema, faça uma inspeção visual dos dados e verifique se você consegue diferenciar o que o classificador está errando;
- Checar a anotação (rótulos) dos dados;
- Verifique os erros principais da matriz de confusão;

Para o trabalho final

- Muito importante analisar os pontos discutidos aqui:
  - Análise de curvas
  - Curva ROC ou Matriz de confusão
  - O que você fez (como agiu) para contornar um problema

Referências

- Aulas Prof. Andrew Ng
- Aulas Prof. David Tax
- Livro Python Machine Learning - Sebastian Raschka - Capítulo 6