- Super-ajuste (*Overfitting*)
- Validação Cruzada

Eduardo Raul Hruschka

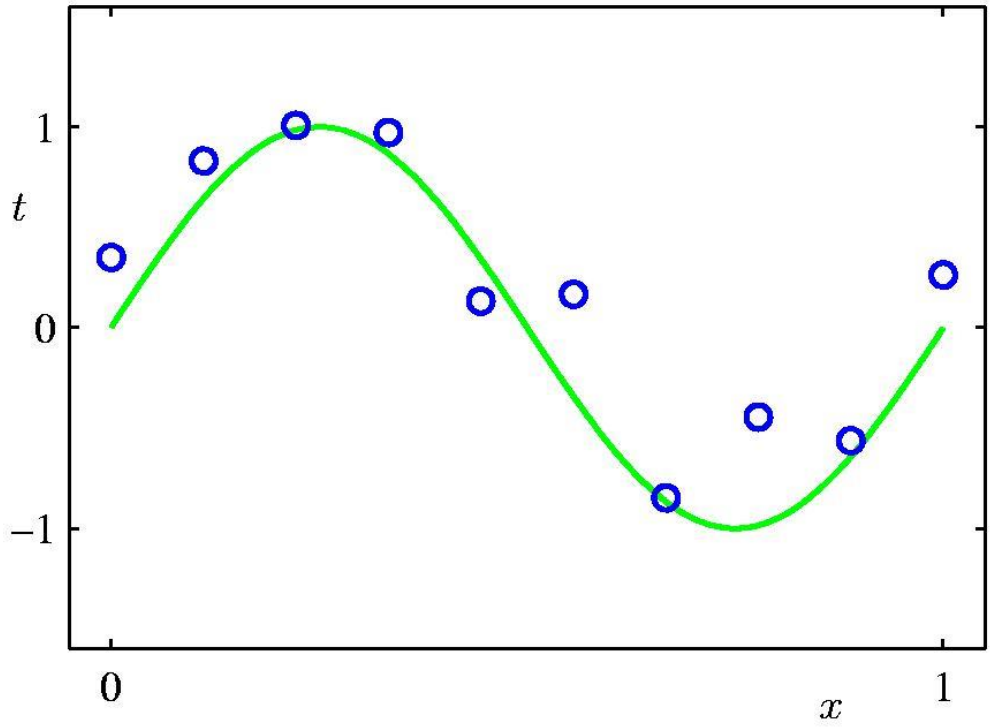
Agenda:

- Conceitos de Classificação
- Técnicas de Classificação
  - *One Rule* (1R)
  - Naive Bayes (com seleção de atributos)
  - Árvores de Decisão
  - K-Vizinhos Mais Próximos (K-NN)
- Super-ajuste (*Overfitting*)
- Validação cruzada
- Combinação de Modelos

Dados de Treinamento, Validação e Teste

- Avaliar desempenho de um classificador ou regressor nos dados de treinamento é útil, mas insuficiente;
- Modelos sofisticados podem simplesmente memorizar dados de treinamento;
- Abordagem recomendada envolve separar os dados disponíveis em 3 conjuntos: Treinamento, Validação e Teste;
- Antes de abordar o assunto em detalhes, vejamos alguns exemplos que servirão de motivação…

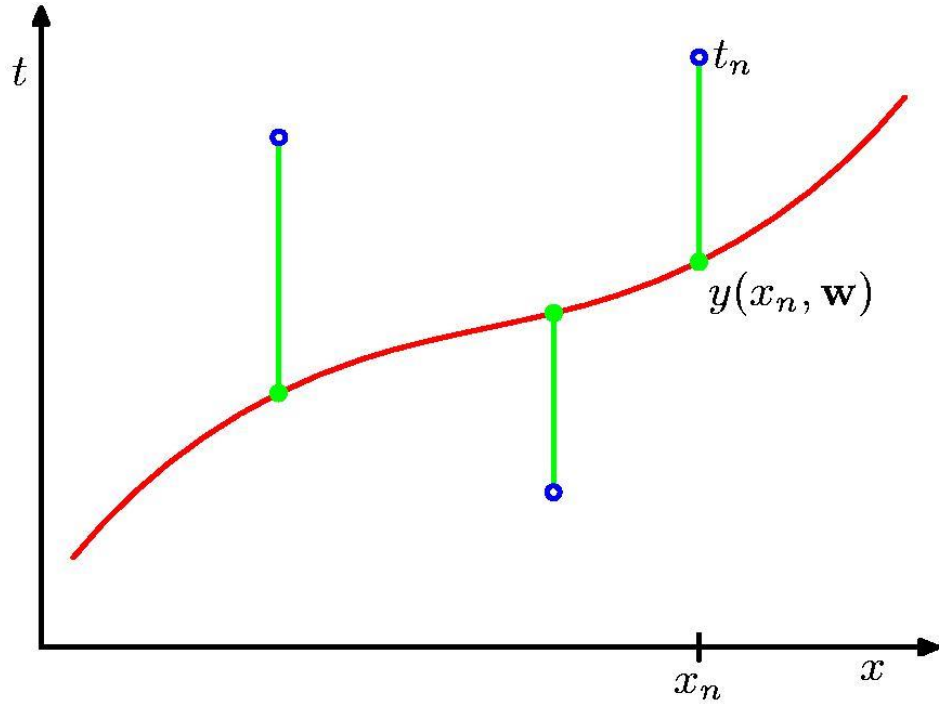
Ajuste de curvas via polinômios



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

Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

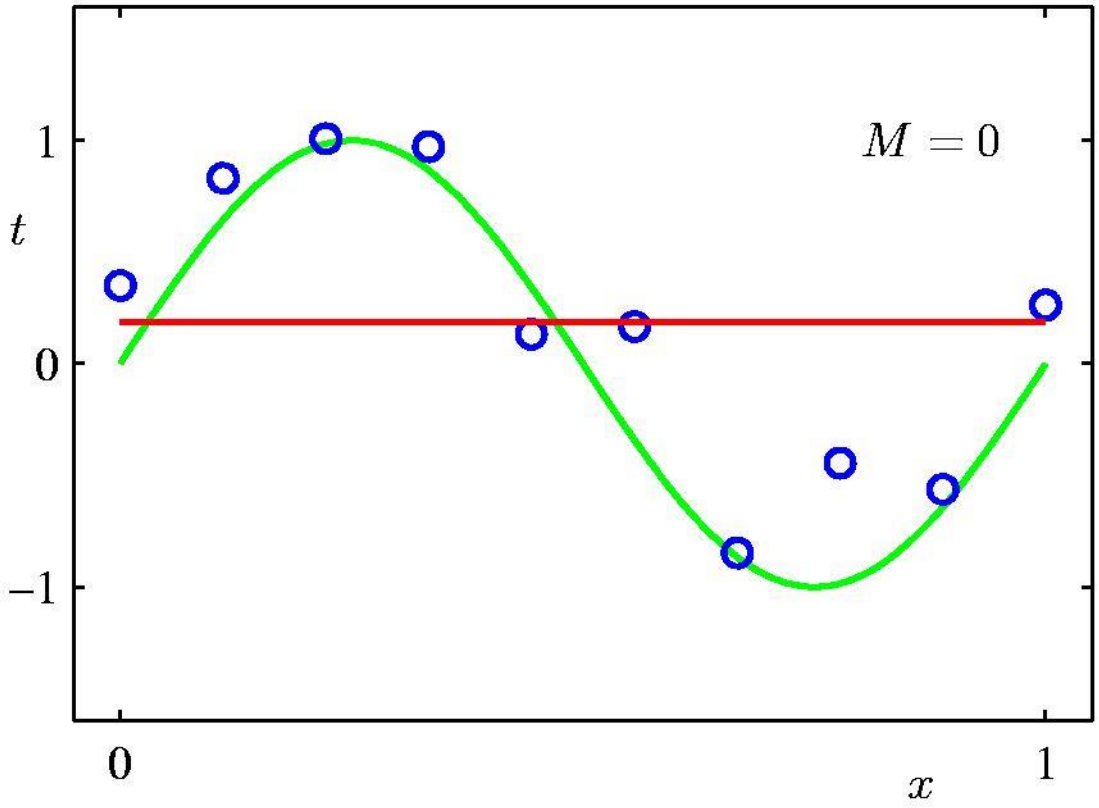
Erro baseado em mínimos quadrados



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

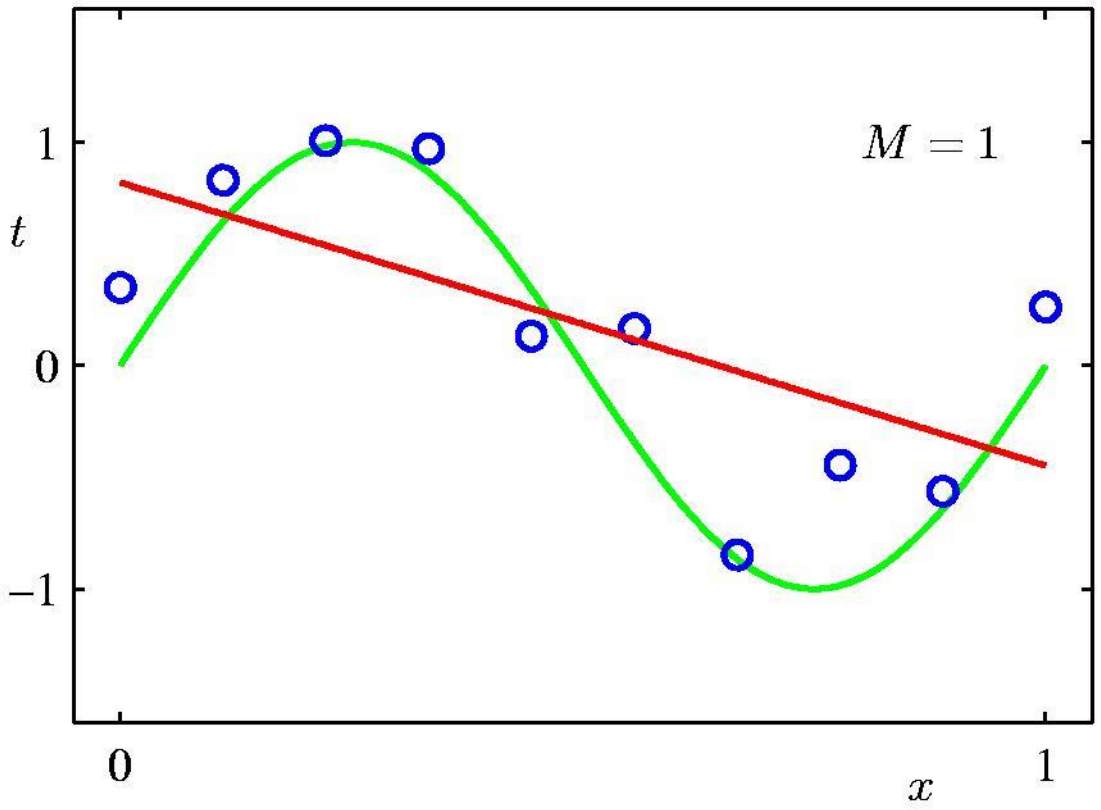
Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

Polinômio de grau zero



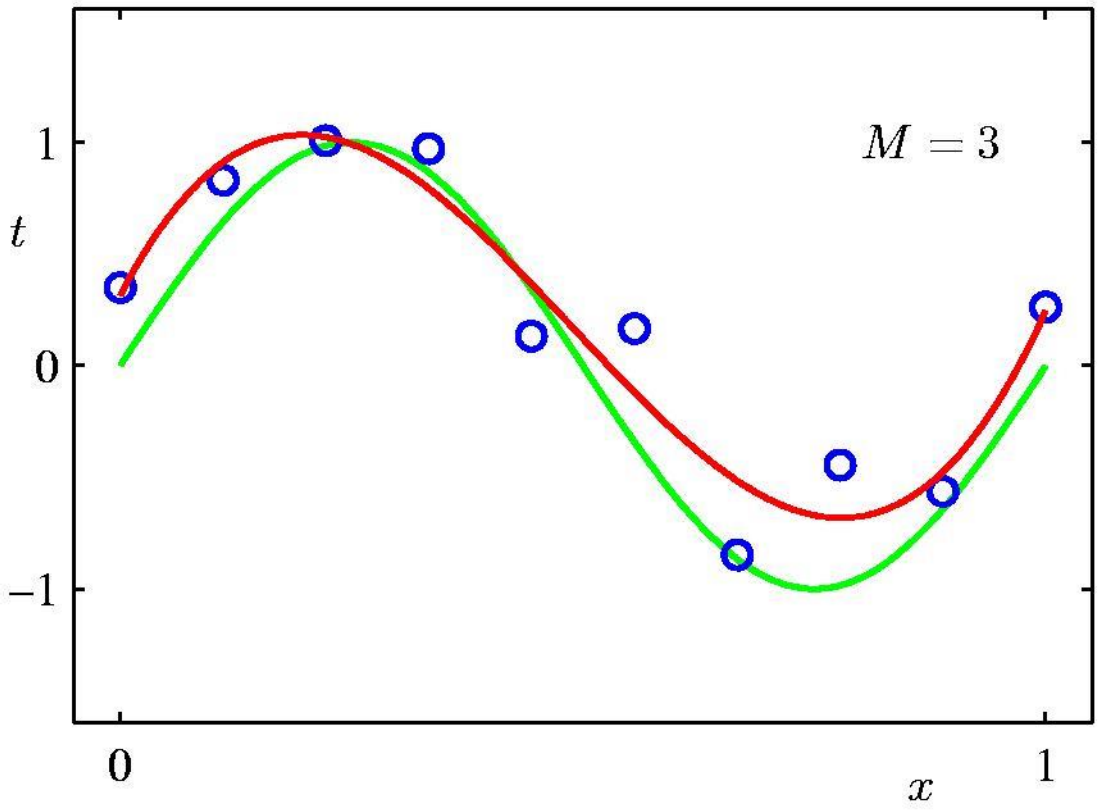
Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

Polinômio de 1º grau



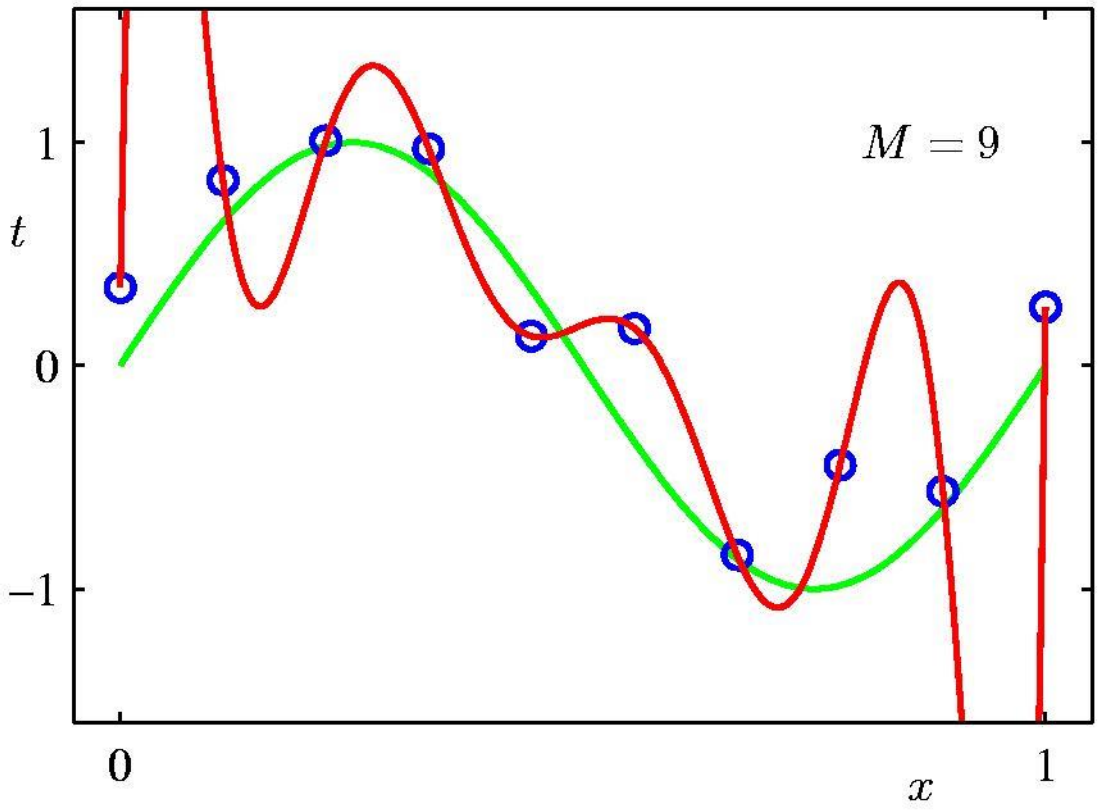
Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

Polinômio de 3º grau



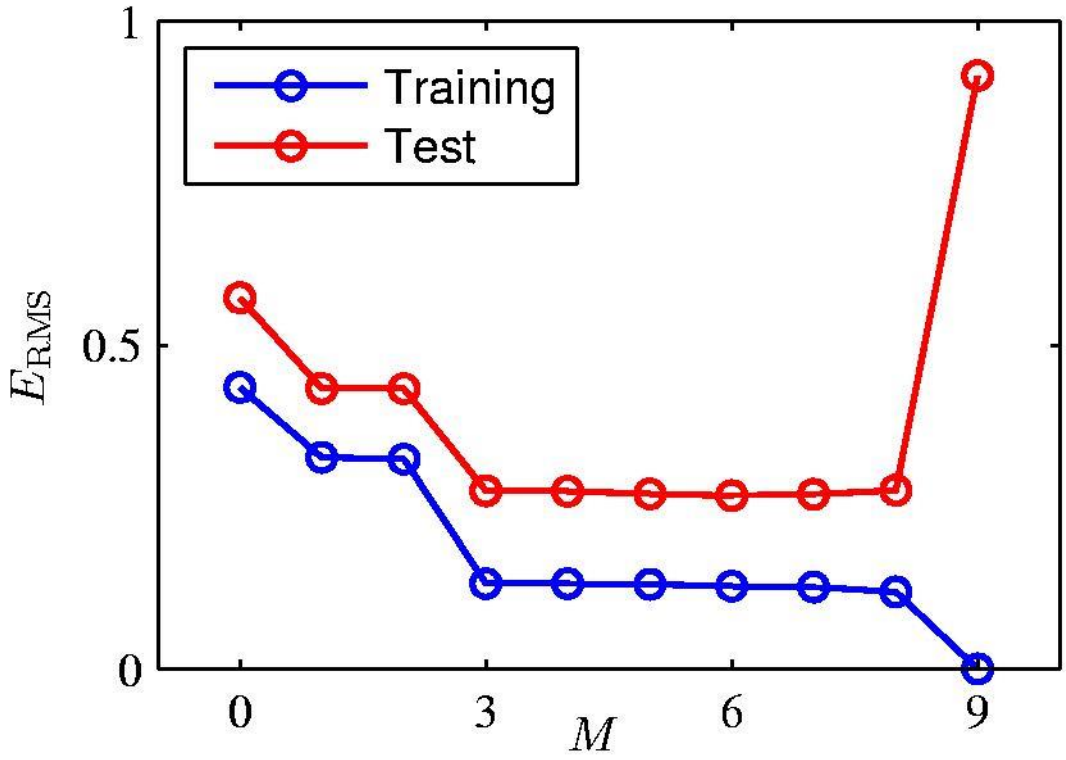
Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

Polinômio de 9º grau



Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

Super-ajuste (*Over-fitting*)



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

Baseado em Bishop, C., *Pattern Recognition and Machine Learning*, *Springer*, 2009.

Coeficientes dos polinômios

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	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

Baseado em Bishop, C., *Pattern Recognition and Machine Learning, Springer*, 2009.

Quando o super-ajuste acontece?

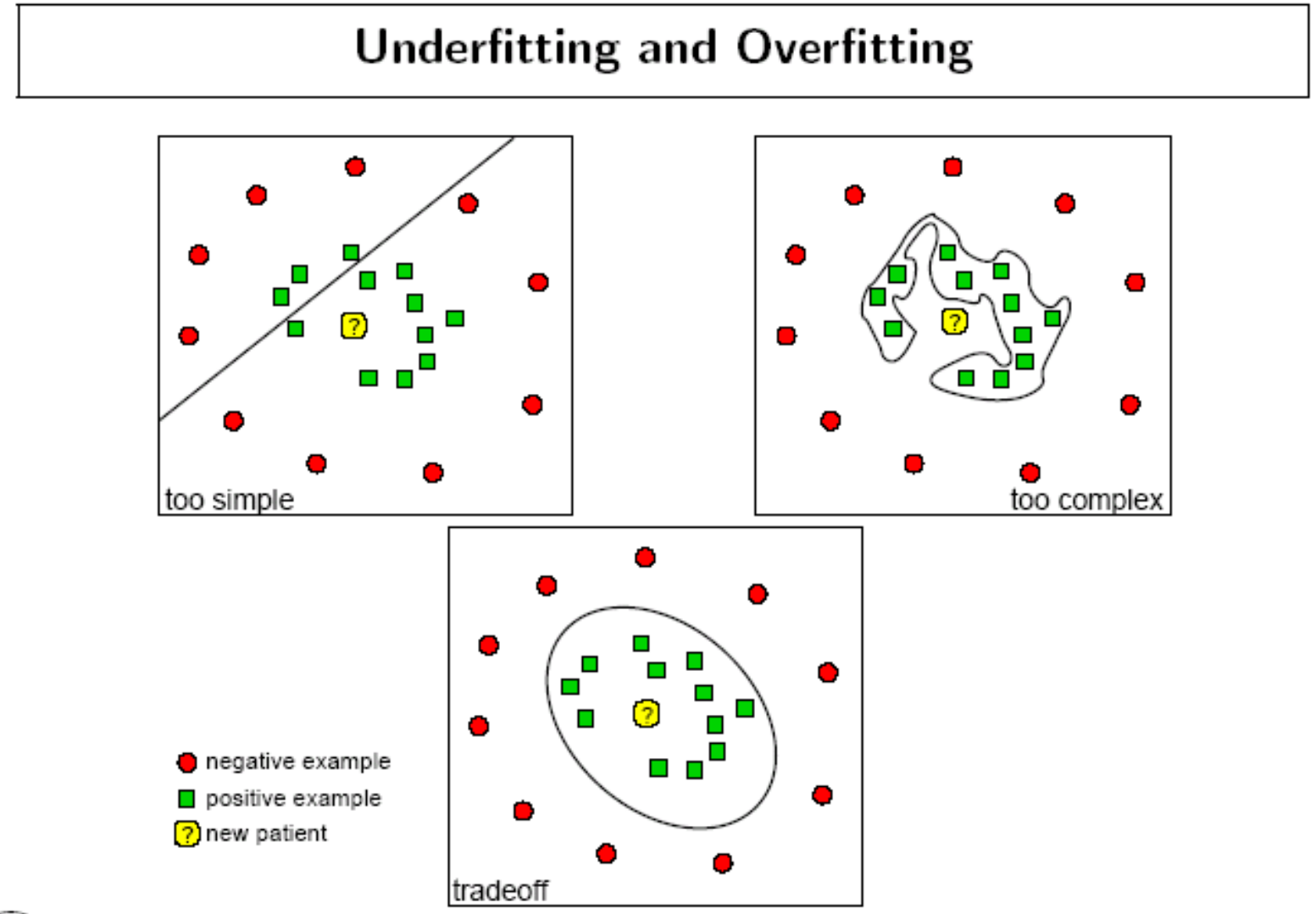
- Quando o modelo descreve o ruído, em vez do *sinal* de interesse;
- Como diferenciar um modelo ruim de uma situação em que ocorre super-ajuste?
- Dilema *bias* x variância

Bias X Variância

- *Bias* (*underfit*):
  - Erro no conjunto de treinamento alto;
  - Erro na validação cruzada alto;

- Variância (*overfit*):
  - Erro no conjunto de treinamento baixo;
  - Erro na validação cruzada muito maior do que no conjunto de treinamento.

Overfitting/Underfitting em classificação:



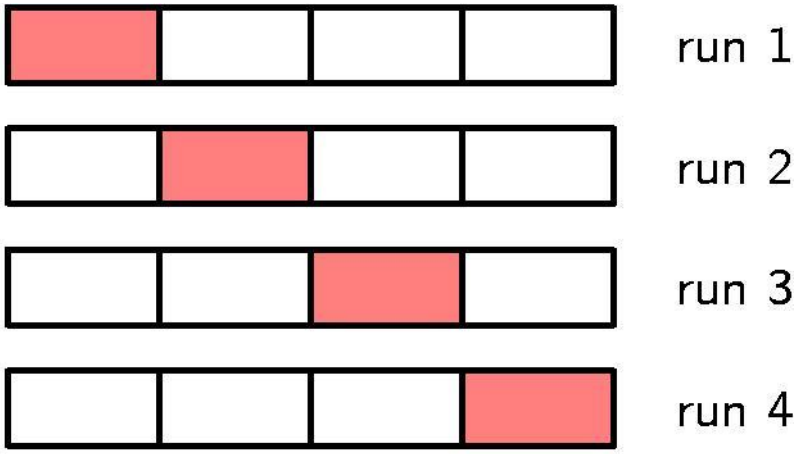
Indícios de Super-ajuste

- Estabilidade
  - Modelo bem ajustado é estável para diferentes amostras de treinamento;
  - Modelo super-ajustado possui desempenho *inconsistente*.
- Desempenho
  - Modelo bem ajustado possui baixo erro de teste;
  - Modelo ruim possui alto erro de teste.

  - Como estimar a qualidade do modelo com alguma confiança ?

Escolhendo modelos via Validação Cruzada:

- Ilustrando o conceito (4 pastas):
  - Dividir o conjunto completo de dados em 4 sub-conjuntos;
  - Usar 75% dos dados pra treinar o modelo (incluindo ajuste de parâmetros) e 25% dos dados pra testar o modelo:



  - Computar medida de interesse (e.g., acurácia) e escolher o melhor modelo. Finalmente, (re)construir o *modelo final*.

Agenda:

- Conceitos de Classificação
- Técnicas de Classificação
  - *One Rule* (1R)
  - Naive Bayes (com seleção de atributos)
  - Árvores de Decisão
  - K-Vizinhos Mais Próximos (K-NN)
- Super-ajuste (*Overfitting*)
- Validação cruzada
- Combinação de Modelos