

Avaliando a Qualidade da Predição

Prof. Dr. Leandro Balby Marinho



Aprendizagem de Máquina

Roteiro

1. Erro no Treino/Teste

2. O Trade-Off Bias-Variância

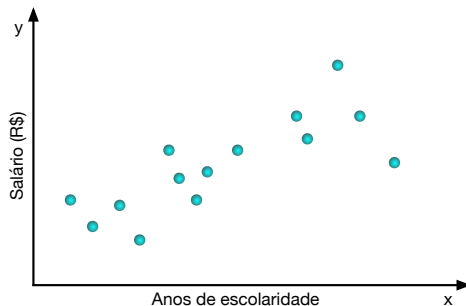
Custo da predição

- ▶ Predição imprecisa da duração de viagens de ônibus.
 - ▶ Muito baixa: perda de compromissos.
 - ▶ Muito alta: perda de tempo.
- ▶ Quanto estou perdendo comparado à predição perfeita?
 - ▶ Predição perfeita: $\text{Custo}=0$
 - ▶ Minha predição: $\text{Custo}=???$

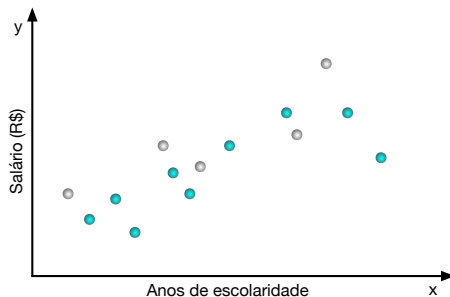
Medindo o custo da predição

- ▶ Uma função de erro/custo $\mathcal{L}(f_{\hat{w}}(\mathbf{x}), y)$ que calcula quão ruim é $f_{\hat{w}}(\mathbf{x})$ se o valor real é y .
- ▶ Há várias opções para \mathcal{L} , e.g., $\mathcal{L} = |f_{\hat{w}}(\mathbf{x}) - y|$
- ▶ Quando $\mathcal{L} = (f_{\hat{w}}(\mathbf{x}) - y)^2$ denominamos \mathcal{L} de *Squared Error (SE)*.

Dados de treino

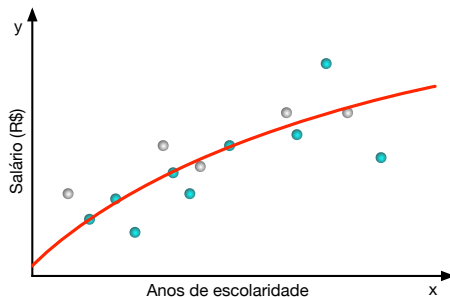


Dados de treino



Se referem à uma amostra da população (pontos em azul).

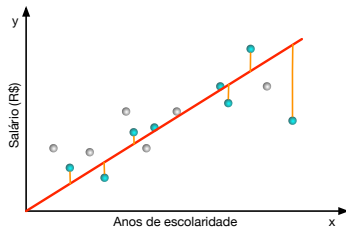
Dados de treino



Construímos nossos modelos usando os dados de treino.

Calculando o erro no treino

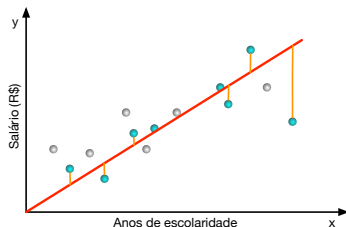
1. Escolha a função de custo $\mathcal{L}(y, f_{\hat{\mathbf{w}}}(\mathbf{x}))$
2. **Erro no treino** = Média sobre todos os erros no treino.



$$\text{err}(\hat{\mathbf{w}}; \mathcal{D}^{\text{train}}) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(f_{\hat{\mathbf{w}}}(\mathbf{x}^{(i)}), y^{(i)})$$

Calculando o erro no treino

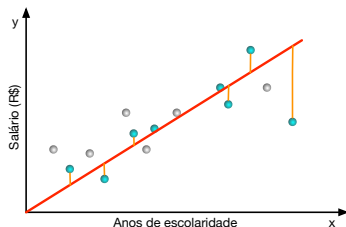
1. Escolha a função de custo $\mathcal{L}(y, f_{\hat{\mathbf{w}}}(\mathbf{x}))$
2. **Erro no treino** = Média sobre todos os erros no treino.



$$\text{MSE}(\hat{\mathbf{w}}; \mathcal{D}^{\text{train}}) = \frac{1}{N} \sum_{i=1}^N \left(f_{\hat{\mathbf{w}}}(\mathbf{x}^{(i)}) - y^{(i)} \right)^2$$

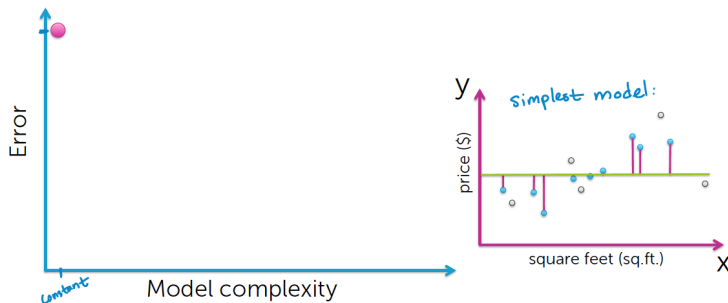
Calculando o erro no treino

1. Escolha a função de custo $\mathcal{L}(y, f_{\hat{w}}(\mathbf{x}))$
2. **Erro no treino** = Média sobre todos os erros no treino.

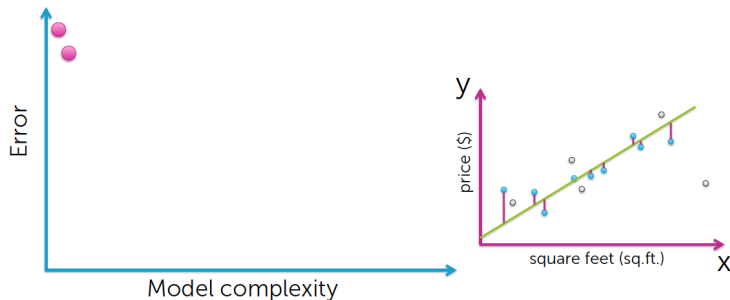


$$\text{RMSE}(\hat{w}; \mathcal{D}^{\text{train}}) = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(f_{\hat{w}}(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

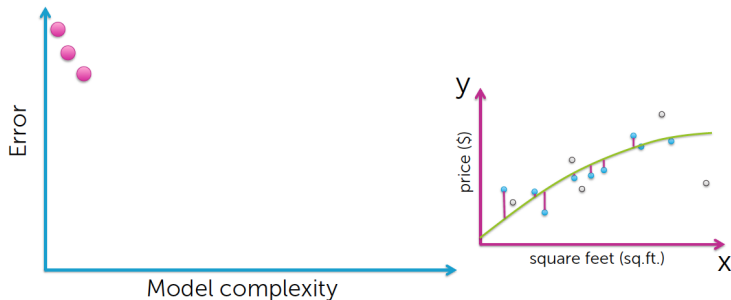
Erro no treino vs Complexidade do Modelo



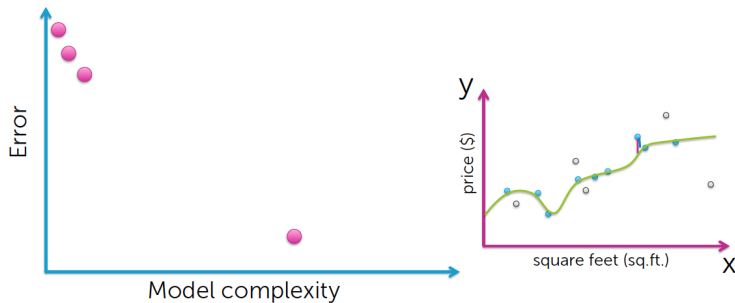
Erro no treino vs Complexidade do Modelo



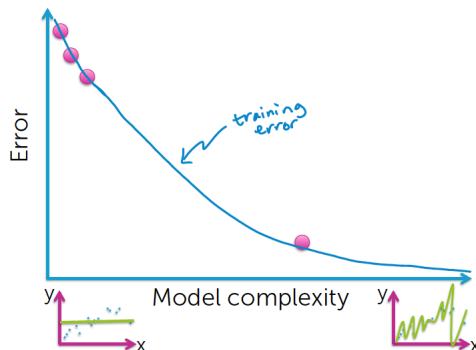
Erro no treino vs Complexidade do Modelo



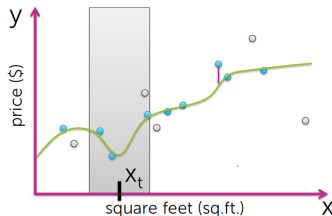
Erro no treino vs Complexidade do Modelo



Erro no treino vs Complexidade do Modelo



Erro no treino é um bom indicador de desempenho?



- ▶ Problema: Erro no treino é muito otimista.
- ▶ Erro pequeno no treino só indica boas previsões se os dados incluírem toda a população.

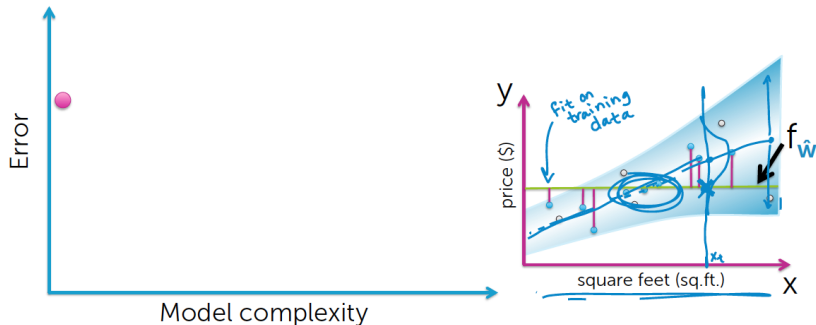
Erro de Generalização (erro real)

- ▶ Estimativa do erro sobre todos os pontos de dados possíveis.
- ▶ Erro de Generalização:

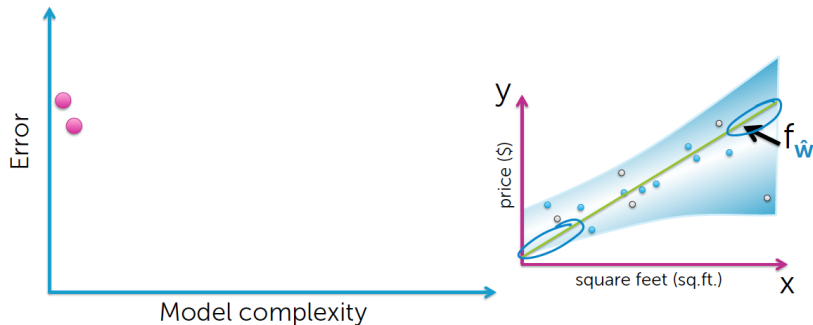
$$E_{\mathbf{x},y}[\mathcal{L}(y, f_{\hat{\mathbf{w}}}(\mathbf{x}))]$$

- ▶ $E_{\mathbf{x},y} \dots$ média sobre os erros de todos os pares (\mathbf{x}, y) ponderada pela probabilidade de cada par.
- ▶ $f_{\hat{\mathbf{w}}}(\mathbf{x}) \dots$ estimado nos dados de treino.

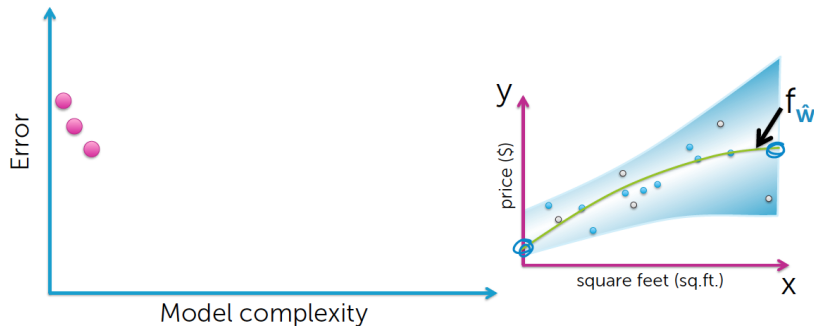
Erro de Generalização vs Complexidade do Modelo



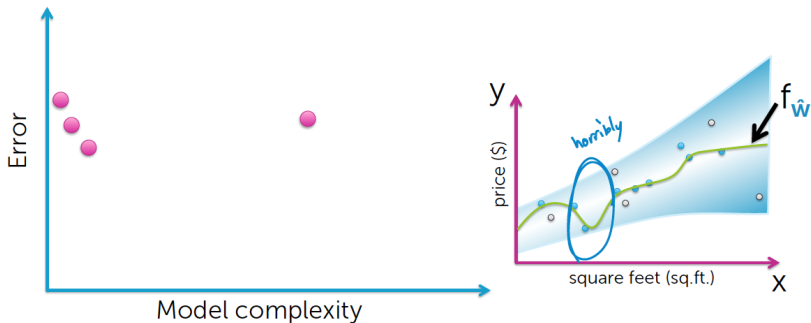
Erro de Generalização vs Complexidade do Modelo



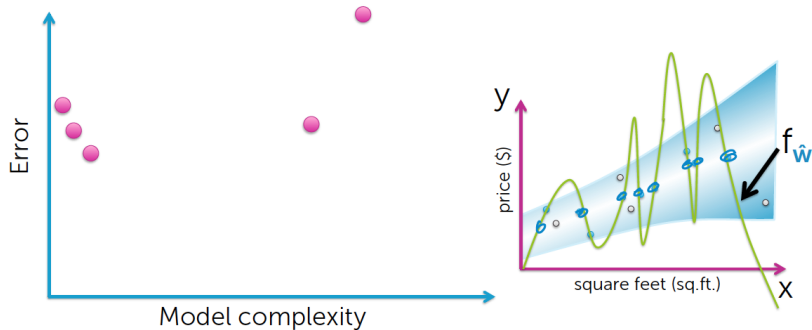
Erro de Generalização vs Complexidade do Modelo



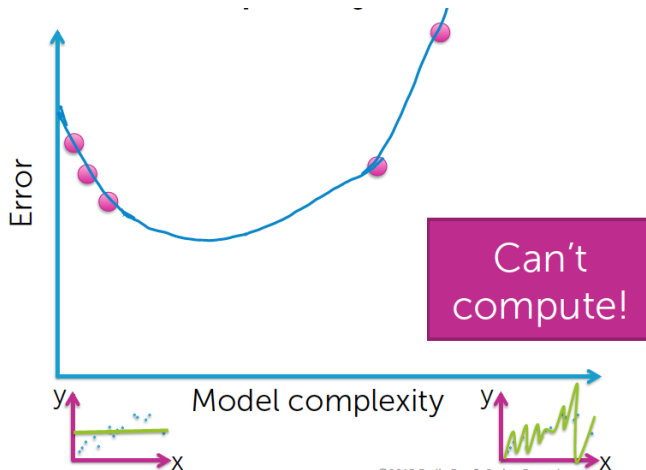
Erro de Generalização vs Complexidade do Modelo



Erro de Generalização vs Complexidade do Modelo



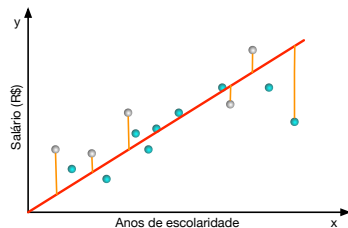
Erro de Generalização vs Complexidade do Modelo



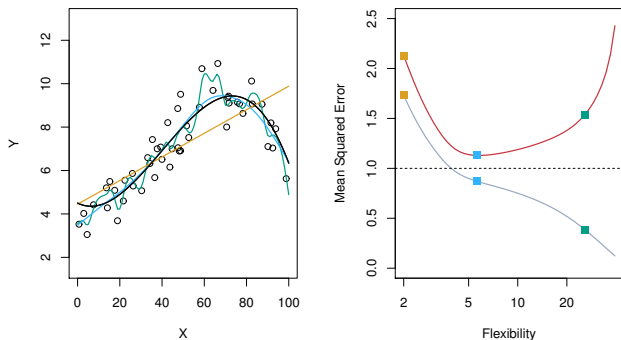
Erro no Teste

- ▶ Estimativa do erro sobre todos os pontos de dados possíveis.
- ▶ Podemos aproximar olhando os dados não presentes no treino.
- ▶ Erro no teste:

$$\frac{1}{|\mathcal{D}^{\text{test}}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}^{\text{test}}} (f_{\hat{\mathbf{w}}}(\mathbf{x}) - y)^2$$

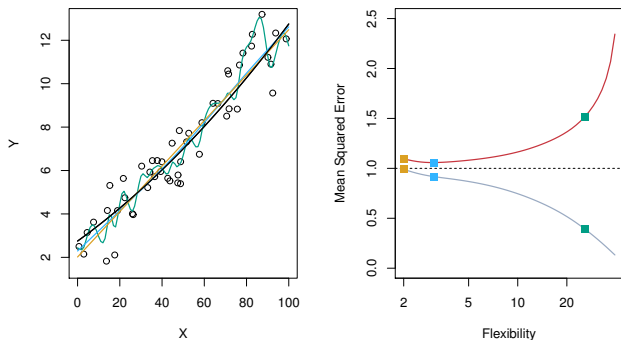


Erro no Treino vs. Erro no Teste



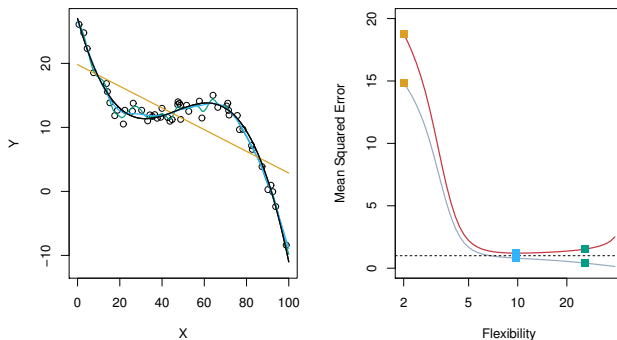
Problema: não há nenhuma garantia que o método com o menor erro no treino também terá o menor erro no teste!

Erro no Treino vs. Erro no Teste



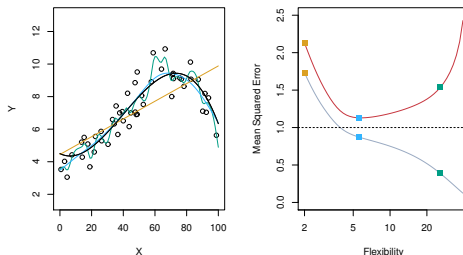
Problema: não há nenhuma garantia que o método com o menor erro no treino também terá o menor erro no teste!

Erro no Treino vs. Erro no Teste



Problema: não há nenhuma garantia que o método com o menor erro no treino também terá o menor erro no teste!

Overfitting



O overfitting acontece se existir um modelo com parâmetros estimados $\hat{\mathbf{w}}$ tal que:

1. o erro no treino para $\hat{\mathbf{w}}$ < erro no treino para \mathbf{w}'
2. o erro real para $\hat{\mathbf{w}}$ > erro real para \mathbf{w}'

(Re-)Formalizando o Problema de Regressão

- ▶ Dado um conjunto de treino $\mathcal{D}^{\text{train}}$,
- ▶ Queremos encontrar $\hat{\mathbf{w}}$ (estimado no treino) tal que para um conjunto de teste $\mathcal{D}^{\text{test}}$ (desconhecido durante o treino), o erro no teste

$$\text{err}(\hat{\mathbf{w}}; \mathcal{D}^{\text{test}}) = \frac{1}{|\mathcal{D}^{\text{test}}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}^{\text{test}}} \mathcal{L}(f_{\hat{\mathbf{w}}}(\mathbf{x}), y)$$

seja mínimo.

Divisão Treino/Teste



- ▶ Dados suficientes no **treino** para uma boa estimativa do modelo.
- ▶ Dados suficientes no **teste** para formar uma boa estimativa do erro real.
- ▶ Normalmente proporções do tipo 70/30 para treino e teste são usadas.

Método Holdout

- ▶ Os dados são particionados aleatoriamente em **treino** e **teste**.
- ▶ O modelo é induzido no treino e avaliado no teste.
- ▶ O método pode ser repetido várias vezes para melhorar a confiabilidade das predições (**random subsampling**).
- ▶ Nesse caso, o MSE é dado por:

$$\text{MSE}(n) = \frac{1}{n} \sum_{i=1}^n \text{MSE}_i$$

onde n é o número de partições treino-teste geradas e MSE_i o MSE na partição i .

Roteiro

1. Erro no Treino/Teste

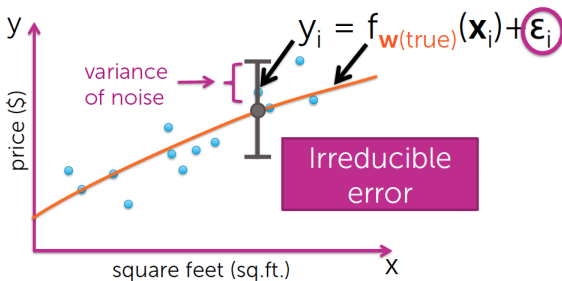
2. O Trade-Off Bias-Variância

Três fontes de erro

Na formação de predições há três fontes de erro:

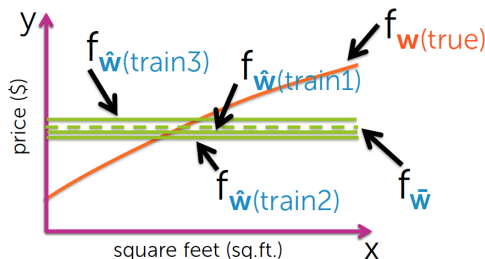
- ▶ Ruído
- ▶ Bias
- ▶ Variância

Dados são naturalmente ruidosos



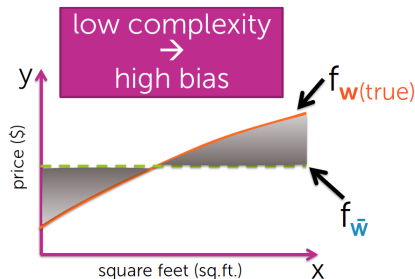
Bias

Em média, como será o meu ajuste para todos os conjuntos de treino de tamanho N possíveis?



Bias

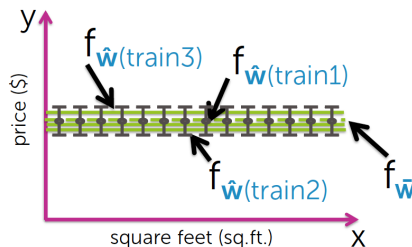
$$\text{Bias}(\mathbf{x}) = f_{\mathbf{w}(\text{true})}(\mathbf{x}) - f_{\bar{\mathbf{w}}}(\mathbf{x})$$



O nosso modelo é flexível o suficiente para capturar $f_{\mathbf{w}(\text{true})}$? Senão, erros nas predições.

Variância de Modelos Simples

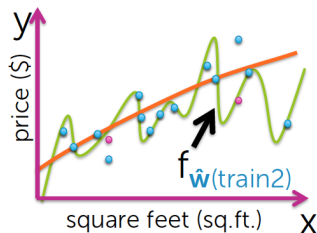
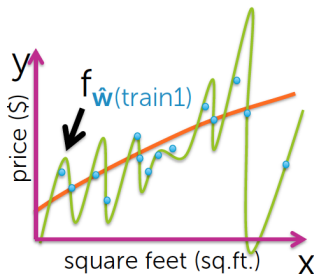
Como ajustes específicos variam em relação ao ajuste médio?



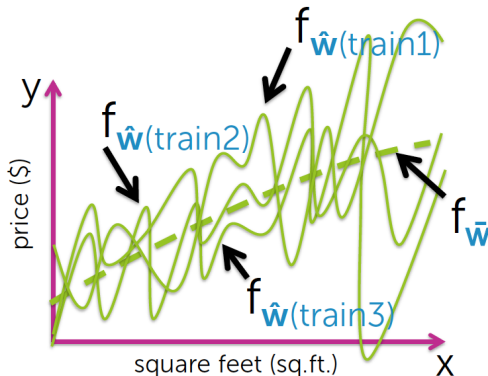
- ▶ Baixa complexidade \Rightarrow baixa variância.
- ▶ Alta variância \Rightarrow previsões erráticas.

Variância de Modelos Complexo

Considerando um ajuste polinomial.

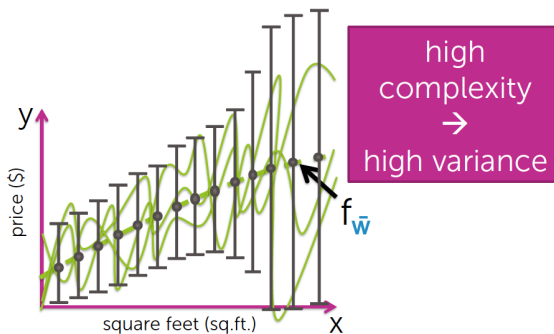


Variância de Modelos Complexo

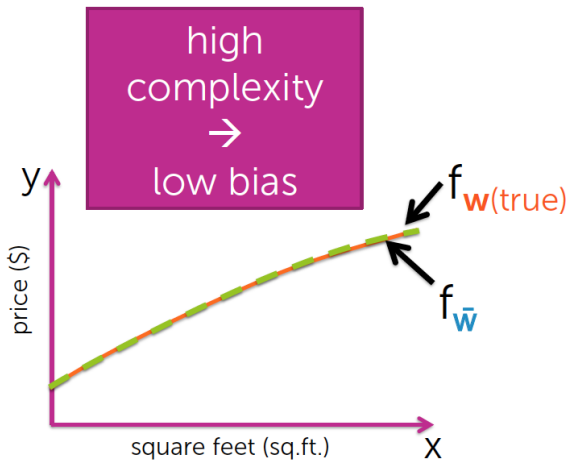


A variabilidade entre os modelos é grande mas a média é uma curva bem comportada.

Variância de Modelos Complexo



Variância de Modelos Complexo



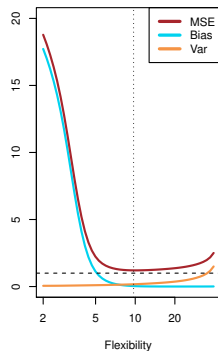
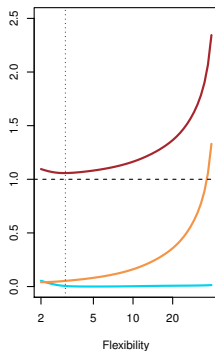
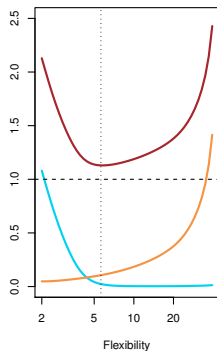
O Trade-Off *Bias-Variância*

- ▶ **Variância** se refere à quantidade de mudança em $f_{\hat{\mathbf{w}}}$ caso ele fosse estimado em um conjunto de treino diferente.
- ▶ **Bias** se refere ao erro associado ao grau de simplificação do modelo em relação ao problema que pode ser muito mais complexo.
- ▶ Para um dado \mathbf{x}_t no teste o MSE pode ser decomposto em três termos:

$$\underbrace{E(y - f_{\hat{\mathbf{w}}}(\mathbf{x}_t))^2}_{\text{MSE}} = \underbrace{\text{Var}(f_{\hat{\mathbf{w}}}(\mathbf{x}_t))}_{\text{Variância}} + \underbrace{[\text{Bias}(f_{\hat{\mathbf{w}}}(\mathbf{x}_t))]^2}_{\text{Bias}} + \underbrace{\text{Var}(\epsilon)}_{\text{Erro irreduzível}}$$

- ▶ $E(y - f_{\hat{\mathbf{w}}}(\mathbf{x}_t))^2 \dots$ se refere à média do MSE considerando estimar \mathbf{w} repetidamente usando um grande conjunto de dados de treino.

O Trade-Off *Bias-Variância*



Workflow da Regressão

- ▶ **Seleção de Modelos:** Normalmente, escolhe-se um parâmetro de ajuste λ relacionado à complexidade do modelo (e.g. ordem do polinômio).
- ▶ **Avaliação do Modelo:** Selecionado o modelo, avaliar o erro de generalização.

Instanciando o Workflow: Forma Ingênu



Conjunto de Treino

Conjunto de Teste

1. **Seleção de Modelos:** Para cada complexidade λ
 - i. Estimar \hat{w}_λ nos **dados de treino**.
 - ii. Avaliar o desempenho de \hat{w}_λ nos **dados de teste**.
 - iii. Escolher λ^* para ser o λ com **menor erro no teste**.
2. **Avaliação do Modelo:** Calcular o erro no teste de \hat{w}_{λ^*}

Instanciando o Workflow: Forma Ingênu



Conjunto de Treino

Conjunto de Teste

1. **Seleção de Modelos:** Para cada complexidade λ
 - i. Estimar \hat{w}_λ nos **dados de treino**.
 - ii. Avaliar o desempenho de \hat{w}_λ nos **dados de teste**.
 - iii. Escolher λ^* para ser o λ com **menor erro no teste**.
2. **Avaliação do Modelo:** Calcular o erro no teste de \hat{w}_{λ^*}




Problema: λ foi selecionado nos dados de teste!

Instanciando o Workflow: Forma correta



- ▶ Selecionar λ^* que minimiza \hat{w}_{λ}^* nos **dados de validação**.
- ▶ Avaliar o desempenho de \hat{w}_{λ}^* nos **dados de teste**.
- ▶ Que proporção usar para treino/validação/teste?
 - ▶ 80/10/10
 - ▶ 50/25/25
 - ▶ ...

Referências

-  Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2013.
-  Yaser S. Abu-Mostafa, Malik Magdon-Ismail. Learning from Data. AMLBook, 2012.
-  Emily Fox and Carlos Guestrin. Machine Learning Specialization. Curso online disponível em <https://www.coursera.org/specializations/machine-learning>. Último acesso: 11/09/2017.