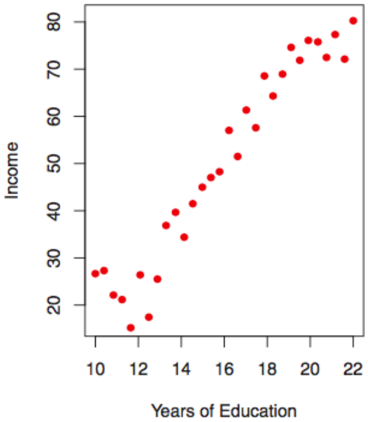
Capítulo 2 - Fundamentos do Aprendizado de Máquina

Cristiano Leite de Castro

Sumário

Exemplo: Income dataset



Problema de Aprendizagem de Máquina (Supervisionado)

- **Objetivo**: estimar uma dependência funcional $\hat{f}(X)$ a partir de um conjunto de n observações:

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} .$$

- assume-se que valores observados (y_i) da variável de saída são gerados de acordo com a seguinte expressão

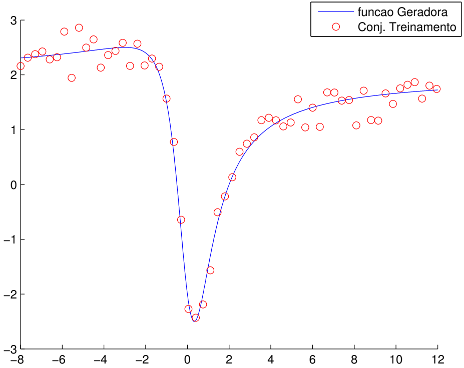
$$Y = f(X) + \epsilon \tag{1}$$

onde

  - $f(X)$ representa o relacionamento real (desconhecido) entre a entrada X e a saída Y e,
  - ϵ é uma v.a. com média 0 e variância σ^2, independente de X.

- Exemplo:

$$Y = \frac{(X-2)(2X+1)}{1+X^2} + \sim N(0, 0.5)$$



Erro Redutível x Erro Irredutível

- A acurácia de $\hat{f}(x_i)$ como um preditor para y_i depende de 2 termos: erro redutível e erro irredutível.

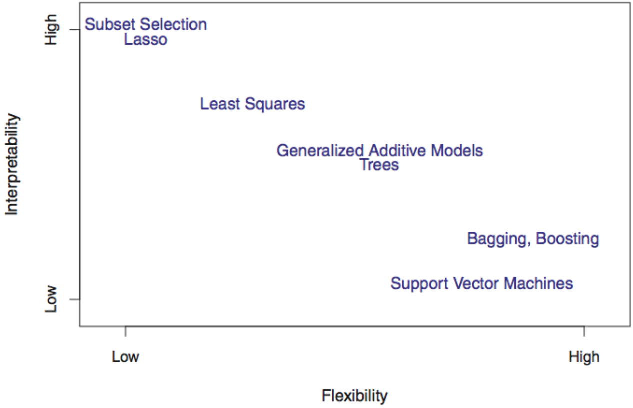
$$
\begin{aligned}
E\left[\left(Y-\hat{f}(X)\right)^2 | X = x_i\right] &= E\left[\left(Y - f(X) + f(X) - \hat{f}(X)\right)^2\right] \\
&= E\left[(Y - f(X))^2\right] + E\left[\left(f(X)-\hat{f}(X)\right)^2\right] \\
&+ \underbrace{2E\left[(Y - f(X))\left(f(X)-\hat{f}(X)\right)\right]}_{0} \\
&= E\left[(\epsilon)^2\right] + E\left[\left(f(X)-\hat{f}(X)\right)^2\right] \\
&= \underbrace{E\left[\left(f(X)-\hat{f}(X)\right)^2\right]}_{\text{Redutivel}} + \underbrace{VAR(\epsilon)}_{\text{Irredutivel}}
\end{aligned}
$$

- se v.a. $Z \sim \mathcal{N}(\mu, \sigma^2)$ então $E\left[(Z)^2\right] = VAR(Z) + E\left[Z\right]^2$.

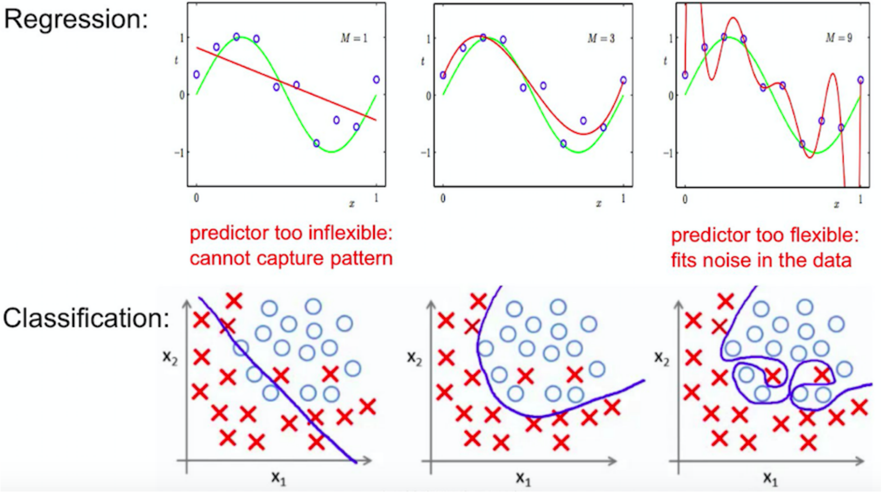
- **Erro Redutível**: $E\left[\left(f(X) - \hat{f}(X)\right)^2 | X = x_i\right]$
  - surge porque $\hat{f}(X)$ não será uma estimativa perfeita para $f(X)$.
  - pode ser reduzido pois, há sempre a possibilidade de se melhorar a acurácia de $\hat{f}(X)$ usando um método de aprendizagem de máquina mais apropriado.
- **Erro Irredutível**: porque ele existe?
  - pode conter variáveis não medidas que seriam úteis na predição de Y.
  - pode conter variações não mensuráveis nos dados, uma vez que para um dado $X = x$ pode existir uma infinidade de possíveis valores de y.
- o **Erro Irredutível** impõe um limite inferior para o erro de $\hat{f}(X)$.

- **Foco deste curso**:
  - estudo de métodos para estimar $f(X)$ com o objetivo de se reduzir o **Erro Redutíve**l.
- **Alguns Trade-offs**:
  - Acurácia versus Interpretabilidade;
  - modelo bem ajustado versus modelo mal ajustado (*underfitting and overfitting*).

Acurácia x Interpretabilidade



Modelo bem ajustado x modelo mal ajustado



- Como avaliar o desempenho de um modelo $\hat{f}(X)$ sobre um conjunto de dados particular?
- Existe um modelo cujo desempenho é dominante sobre todos os outros?
- Como saber se estamos próximos ou distantes da função ideal $f(X)$ (desconhecida)?
- É importante saber decidir, para um dado problema em mãos (conjunto de dados), qual modelo produz melhores resultados.

- Em problemas típicos de Regressão a v.a. Y assume valores no domínio dos números reais.
- conjunto de treinamento contendo n observações

$$T = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \text{ com } y_i \in \mathbb{R}.$$

- Erro quadratico médio (MSE) calculado sobre o conjunto de treinamento T.

$$MSE_{train} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \tag{2}$$

onde $\hat{f}(x_i)$ é a predição que $\hat{f}$ fornece sobre a i-ésima observação.

- Erro quadratico médio (MSE) calculado sobre um conjunto muito grande de observações de teste (desconhecidas)

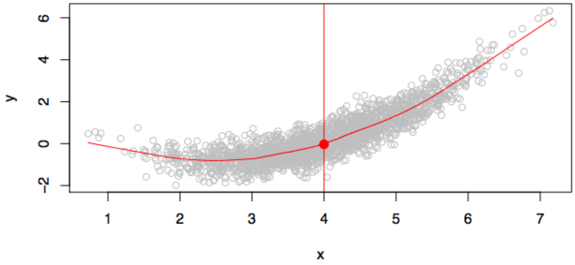
$$MSE_{test} \approx \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{f}(x_i))^2, \text{ para } m >> n \tag{3}$$

- Escolher o modelo usando somente MSE_{train} pode levar a *overfitting*.
- Não há garantias de que o algoritmo de aprendizagem que fornece o menor MSE_{train} também fornece o menor MSE_{test}.
- Se fosse possível calcular MSE_{test}, o melhor modelo seria aquele para o qual MSE_{test} é mínimo.
    - MSE_{test} poderia ser calculado se a **função geradora** dos dados fosse conhecida porém, usualmente este não é o caso.

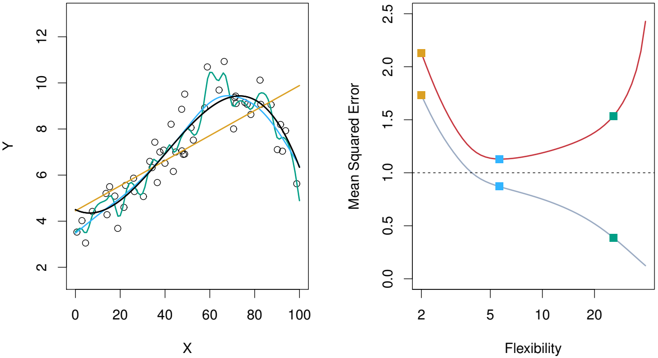
- Função Ideal ou função de Regressão

$$f(x) = E\left[Y|X = x\right] \tag{4}$$

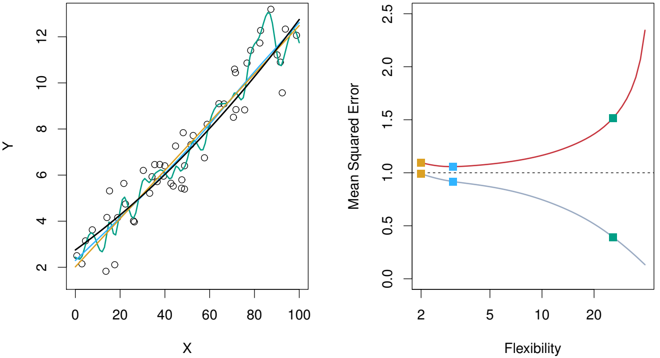
é a função que minimiza $E\left[(Y - g(X))^2 \,|X = x\right]$ sobre todas as funções g em todos os pontos $X = x$.



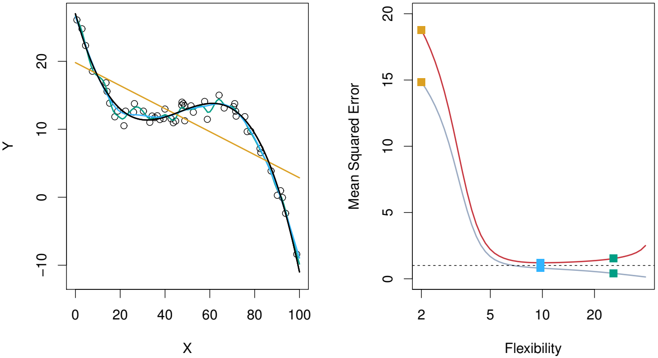
Relacionamento entre MSE_{train} e MSE_{test} (1)



Relacionamento entre MSE_{train} e MSE_{test} (2)



Relacionamento entre MSE_{train} e MSE_{test} (3)



Conclusões

- MSE_{train} decresce monotonicamente com o nível de flexibilidade enquanto MSE_{test} apresenta curva "U-shape"
- Esta é uma propriedade fundamental que é válida independentemente do problema em mãos e do algoritmo de aprendizagem de máquina
- O nível ótimo de flexibilidade pode variar consideralmente entre problemas.
  - a pergunta é: como alcançar este ponto mínimo?

- Em problemas de Classificação a v.a. Y é qualitativa, podendo assumir valores dentro de um conjunto enumerável: $\{1, 2, \ldots, K\}$, contendo k rótulos (ou classes).

- conjunto de treinamento contendo n observações

$$T = \{(x_1, y_1), \ldots, (x_n, y_n)\}, \text{ com } y_i \in Y = \{1, 2, \ldots, K\}$$

- Taxa de Erro sobre o conjunto de treinamento T.

$$E_{train} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i)) \tag{5}$$

onde $\hat{f}(x_i)$ é a classe predita para a i-ésima observação e

$$I(y_i \neq \hat{f}(x_i)) = \begin{cases} 1 & \text{se } y_i \neq \hat{f}(x_i), \\ 0 & \text{caso contrário.} \end{cases}$$

- Taxa de Erro calculado sobre um conjunto de observações de teste (desconhecidas)

$$E_{test} = \frac{1}{m} \sum_{i=1}^{m} I(y_i \neq \hat{f}(x_i)) \tag{6}$$

$$I(y_i \neq \hat{f}(x_i)) = \begin{cases} 1 & \text{se } y_i \neq \hat{f}(x_i), \\ 0 & \text{caso contrário.} \end{cases}$$

- Classificador ideal ou ótimo

$$f(x) = k \text{ se } P(Y = k|X = x) = \max\{P(Y = 1|X = x), \dots, P(Y = K|X = x)\} \tag{7}$$

onde $P(Y = k|X = x)$ é a probabilidade (condicional) que $Y = k$, dado que x foi observado. Isto significa que o classificador ótimo atribui cada observação à classe mais provável.

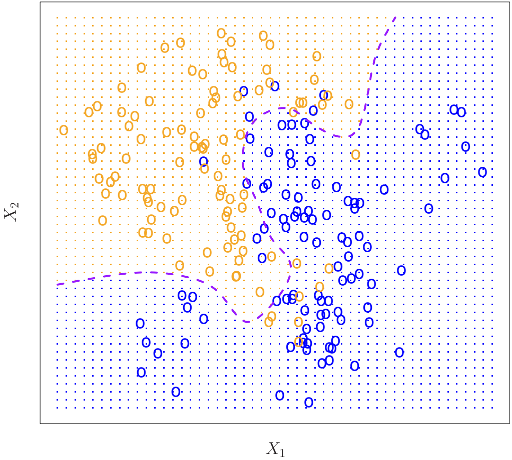
- Ao se considerar o caso particular de 2 classes, onde $Y = \{1, 2\}$, o classificador ótimo corresponde a

$$f(x) = \begin{cases} 1 & \text{se } P(Y = 1|X = x) \geq 0.5, \\ 2 & \text{caso contrário.} \end{cases} \tag{8}$$

- O classificador ótimo produz a menor taxa de Erro possível, conhecida como Taxa de Erro de *Bayes*.

- Como o classificador ótimo sempre escolhe a classe para a qual $P(Y|X)$ é máxima, a taxa de Erro em $X = x$ é $1 - \max_k P(Y = k|X = x)$. Dessa forma, o Erro Global de *Bayes* é dado por
$$1 - E[\max_k P(Y = k|X)]$$
onde o operador esperança é aplicado sobre todos os valores possíveis de X.

Exemplo de Classificador Ótimo



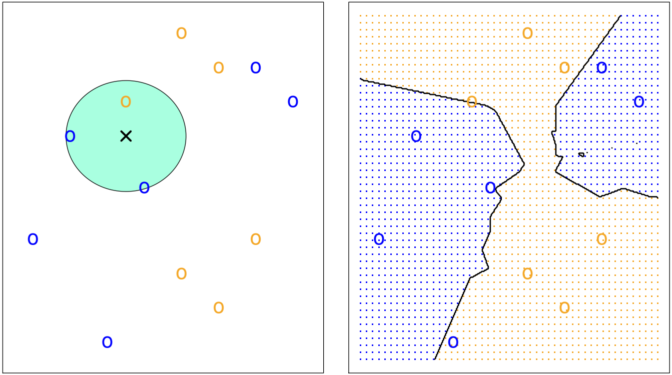
- Dado um inteiro positivo K e uma observação de teste x_0, o classificador *KNN* identifica, primeiramente, os K pontos do conjunto de treinamento que são mais próximos de x_0, representado pelo conjunto $\mathcal{N}_0$. Em seguida, *KNN* estima a probabilidade condicional para a classe k como a fração de pontos em $\mathcal{N}_0$ cujo valor de saída (y_i) é igual a k

$$P(Y = k|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = k)$$

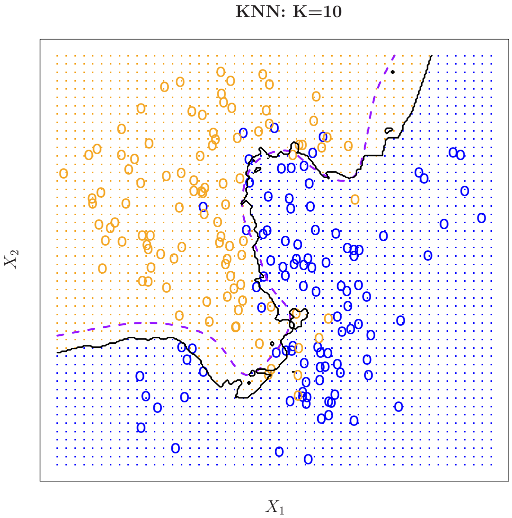
Finalmente, *KNN* aplica a regra de Bayes e classifica a observação x_0 segundo a regra de decisão

$$f(x) = j \text{ se } P(Y = j|X = x) = \max\{P(Y = 1|X = x), \ldots, P(Y = K|X = x)\}$$

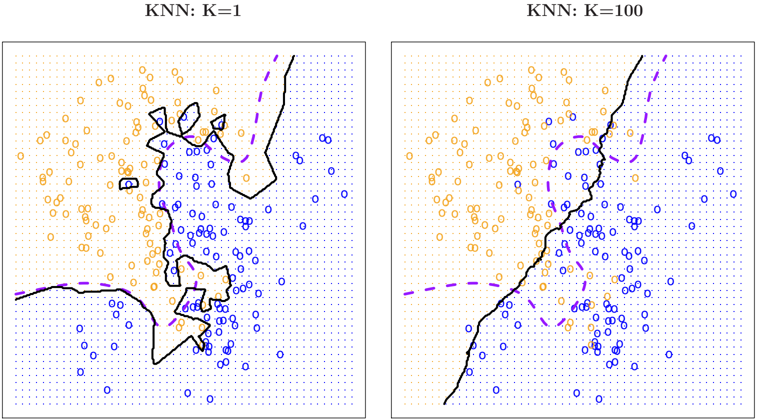
Exemplo - KNN



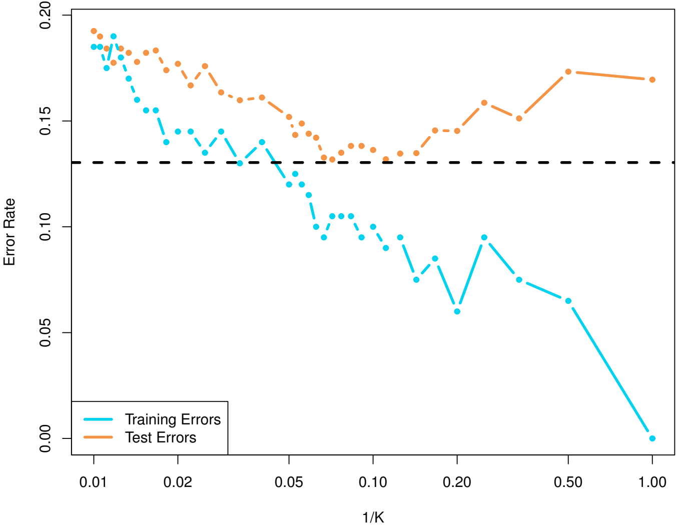
KNN com K = 10



KNN com K = 1 e K = 100



KNN - Relacionamento entre MSE_{train} e MSE_{test}



- (x_0, y_0) é uma observação arbitrária (de teste) extraída da população;
- $Var(\cdot)$ e $E[\cdot]$ correspondem, respectivamente, à variância e esperança de uma variável aleatória;
- O valor esperado do Erro (MSE_{Test}), para uma dada observação x_0, pode ser decomposta na soma de três termos fundamentais, ou seja

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \qquad (9)$$

A notação $E(y_0 - \hat{f}(x_0))^2$ refere-se ao Erro médio que seria obtido se $\hat{f}$ fosse repetidamente estimado usando um grande número de conjuntos de treinamento e, cada estimativa testada sobre a observação x_0.

Bias x Variance Tradeoff - Dedução (1)

Foi mostrado anteriomente que

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = E\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + VAR(\epsilon)$$

Desenvolvendo o termo referente ao Erro redutível, tem-se

$$\begin{aligned}
E\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] &= E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right] + E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right] \\
&= E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right] + E\left[\left(E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right] \\
&\quad + \underbrace{2E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)\left(E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)\right]}_{0} \\
&= \underbrace{E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right]}_{Bias^2(\hat{f})} + \underbrace{E\left[\left(E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right]}_{VAR(\hat{f})}
\end{aligned}$$

Bias x Variance Tradeoff - Dedução (2)

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \underbrace{E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right]}_{Bias^2(\hat{f})} + \underbrace{E\left[\left(E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right]}_{VAR(\hat{f})} + VAR(\epsilon)$$

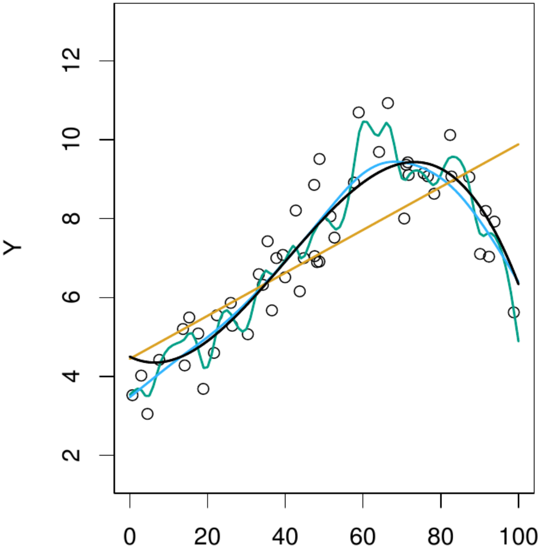
Finalmente, pode se calcular o MSE_{Test} (global) tomando-se a média sobre todas as observações (x_0, y_0) pertencentes ao conjunto de teste.

$$MSE_{test} = Ave\left(E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]\right) \ \forall (x_0, y_0).$$

$$Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \qquad (10)$$

- o termo Variância se refere à quantidade pela qual $\hat{f}$ mudaria se ele fosse estimado usando um conjunto de treinamento diferente. Em geral, métodos mais flexíveis tendem a ter maior variância.
$$Var(\hat{f}(x_0)) = E[(E[\hat{f}(x_0)] - \hat{f}(x_0))^2]$$
- Como zerar o termo de variância?

Exemplo - Variância



Bias

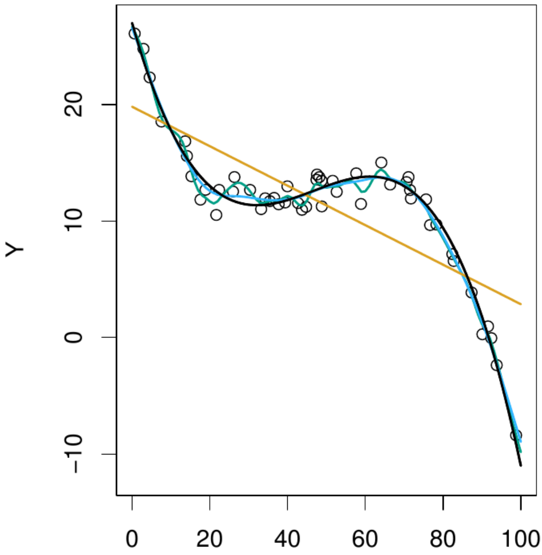
$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \quad (11)$$

- O termo Bias se refere ao Erro que é introduzido por se tentar aproximar um problema real, que pode ser extremamente complexo, por um modelo mais simples. Em geral, métodos mais flexíveis resultam em menos *Bias*.

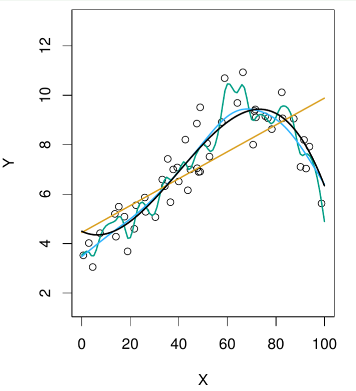
$$[Bias(\hat{f}(x_0))]^2 = E[(f(x_0) - E[\hat{f}(x_0)])^2]$$

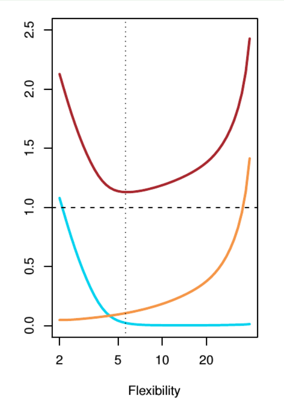
- Como zerar o termo de bias?

Exemplo - Bias

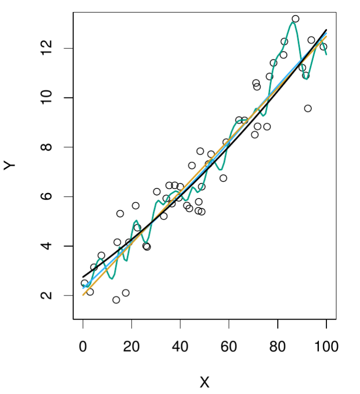


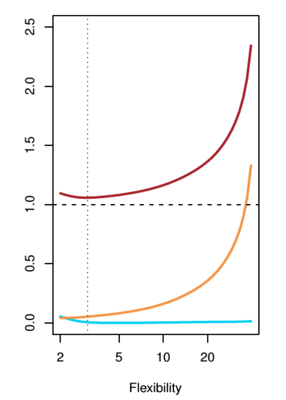
Bias x Variance Tradeoff (1)



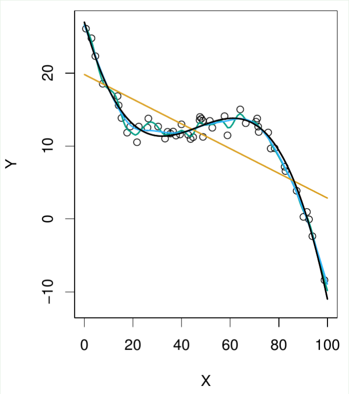


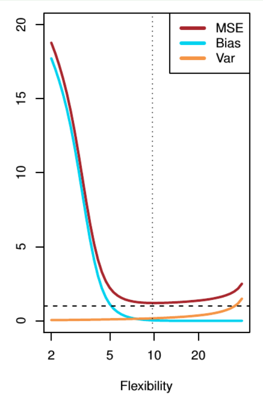
Bias x Variance Tradeoff (2)





Bias x Variance Tradeoff (3)





- Com o objetivo de minimizar MSE_{Test}, deve-se selecionar um modelo que ao mesmo tempo obtém
  - valores reduzidos de **Bias** e **Variancia**.
- Em geral, quando a flexibilidade de $\hat{f}$ aumenta, sua **Variancia** aumenta enquanto o seu **Bias** reduz.
- O nível ótimo de flexibilidade, correspondente ao MSE_{Test} mínimo, pode variar consideravelmente de problema para problema.

Referências

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning*. 2013.
- Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction."The Mathematical Intelligence, 2001
- Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
- Bishop, C. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc, 2006.