

# PLN para a Ciência Política e Políticas Públicas

Professora: Lorena Barberia

Semana 10

# Tópicos da Aula

- 1 Motivação
- 2 Reduzindo Viés em PLN
- 3 Reduzindo Variância  
(Single-split & Cross-Validation  
Bootstrap, Ensemble Methods, Bagging e Boosting)
- 4 Bias and Variance Tradeoff



# Reducible Error

- Viés e Variância são os determinantes do erro redutível.
- Hoje, vamos continuar discutindo estratégias para **reduzir** viés e/ou variância.

# Reduzindo Viés em PLN

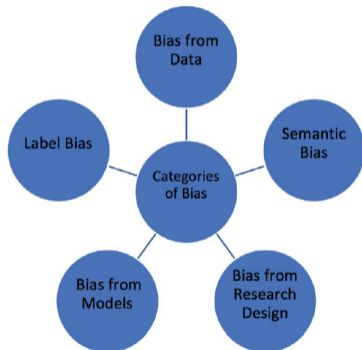


FIGURE 1.1 Different categories of bias. (Hovy & Prabhunoye, 2021.)

## Reduzindo Viés em PLN: Label Bias

- As categorias que adotamos para a mensuração do conceito
- Os anotadores também podem introduzir preconceitos quando estão distraídos, desinteressados ou preguiçosos com a tarefa de anotação.(Hovy Prabhumoye, 2021).

## Reduzindo Viés em PLN: Bias from Data

- A amostra selecionada é representativa da amostra alvo de nossa inferência?

## Reduzindo Viés em PLN: Semantic Bias

- A semântica se refere ao significado que está sendo comunicado.
- "How one's lexicon, grammatical structure, tone, and other elements of a sentence combine to communicate its meaning."
- "Word embeddings pick up on racial and gender biases (Bolukbasi et al., 2016; Manzini et al., 2019)."

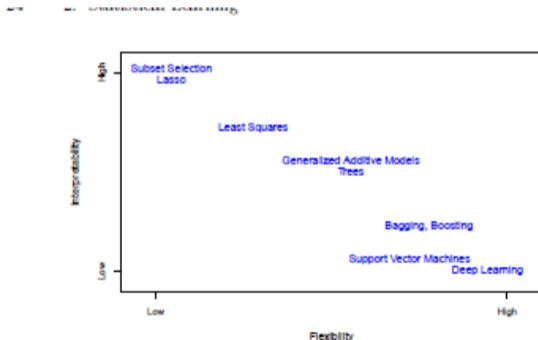
# Reduzindo Viés em PLN: Research Design Bias

- Stochastic Data-Driven Models vs Neural Data-Driven Models



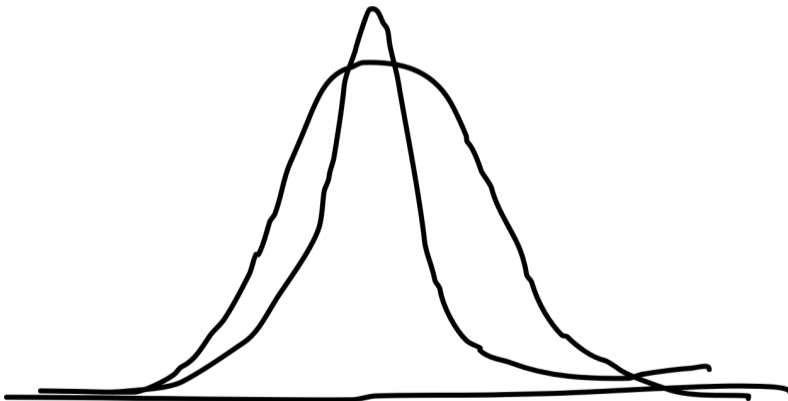
# Reduzindo Viés em PLN: Bias from Models

- Linearidade *Predictors/Features*, Complexidade



**FIGURE 2.7.** A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

# Reduzindo Variância em PLN



# Single-split cross-validation

- 80% da amostra para treinamento e 20% para teste.
- Vantagem: Usamos uma parcela majoritária para avaliar uma parcela minoritária da amostra.
- Desvantagem: Os resultados dependem de como a amostra é dividida.

# 5-fold Cross-validation

Table: K-Folds (K=5)

	Fit on Folds	Test on Fold	MSE
j=1	2,3,4,5	1	$MSE_{j=1}$
j=2	1,3,4,5	2	$MSE_{j=2}$
j=3	1,2,4,5	3	$MSE_{j=3}$
j=4	1,2,3,5	4	$MSE_{j=4}$
j=5	1,2,3,4	5	$MSE_{j=5}$
Average			MSE

Ao dividir os dados em conjuntos de treinamento e teste várias vezes, a validação cruzada pode ajudar a identificar se um modelo está superajustado (overfitting) ou subajustado (underfitting).

# 5-fold Cross-validation

Table: K-Folds (K=5)

	Fit on Folds	Test on Fold	MSE
j=1	2,3,4,5	1	13.98
j=2	1,3,4,5	2	6.49
j=3	1,2,4,5	3	9.36
j=4	1,2,3,5	4	6.41
j=5	1,2,3,4	5	12.23
Average			9.70

## 5-fold Cross-validation vs 10-fold validation

- Depende do tamanho de nossa amostra e o tradeoff entre viés e ineficiência.
- Vantagem: Com 10 *folds*, o tamanho da amostra de treino aumenta, e assim, o viés diminui.
- Desvantagem: A correlação entre amostras de teste aumenta, e isso aumenta a variância.

# Bootstrap

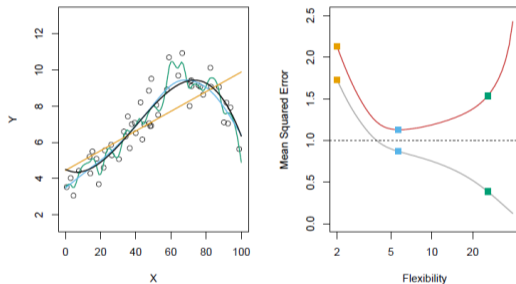
- Método para obter erros padrão (muito útil especialmente em técnicas onde é difícil obter erros padrão)
- Reamostragem da amostra com substituição

# Bias and Variance Tradeoff



## Caso onde Modelo Linear não é o melhor modelo

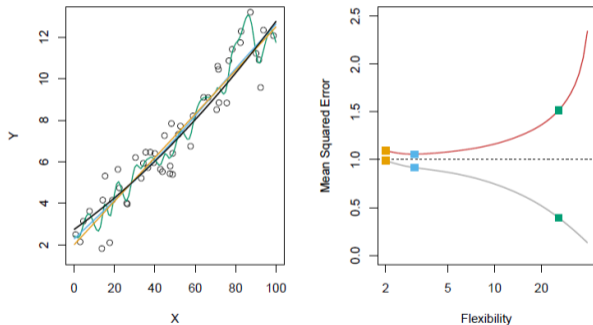
- Fig 2.9 - diminuição do training MSE e U do test MSE.
- Modelo linear (OLS) tem o MSE de treinamento e teste maior.
- Preferimos o modelo com o menor MSE de treino (modelo de quadrados azul).
- Modelo verde = MSE de treinamento menor, mas MSE de teste elevado = **overfitting**.



**FIGURE 2.9.** Left: Data simulated from  $f$ , shown in black. Three estimates of  $f$  are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

## Caso onde Modelo Linear é o melhor modelo

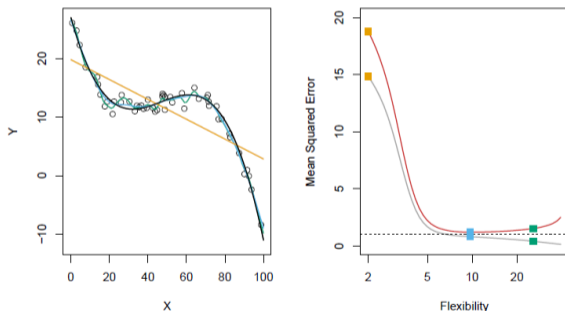
- Fig 2.10 o modelo linear tem o menor MSE no banco de treinamento e teste.



**FIGURE 2.10.** Details are as in Figure 2.9, using a different true  $f$  that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

# Caso onde Modelo Linear não é o melhor modelo

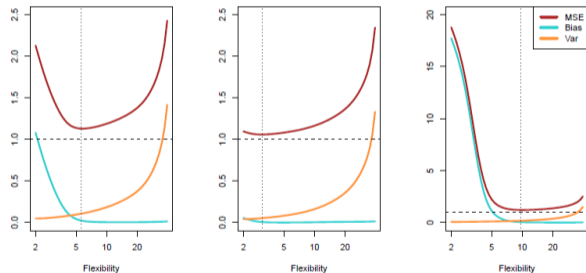
- Fig 2.11 diminui o MSE de treinamento de teste nos modelos alternativos.
- Modelo linear é o pior.



**FIGURE 2.11.** Details are as in Figure 2.9, using a different  $f$  that is far from linear. In this setting, linear regression provides a very poor fit to the data.

# Bias and Variance Tradeoff

- Fig 2.12 demonstra um padrão importante – o MSE de teste sempre cai e depois aumenta. **As bias falls, variance increases. When we choose lowest MSE, we are trying to account for bias-variance tradeoff.**



**FIGURE 2.12.** Squared bias (blue curve), variance (orange curve),  $\text{Var}(\epsilon)$  (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

# Avaliação do Modelo

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <b>Type II Error</b>	<b>Sensitivity</b> $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)	<b>Specificity</b> $\frac{TN}{(TN + FP)}$
		<b>Precision</b> $\frac{TP}{(TP + FP)}$	<b>Negative Predictive Value</b> $\frac{TN}{(TN + FN)}$	<b>Accuracy</b> $\frac{TP + TN}{(TP + TN + FP + FN)}$

Table 1. Confusion matrix with advanced classification metrics

# Laboratório 9