Aula 30/09/2020 e Aula 07/10/2020

Tópicos discutidos: usamos vários tópicos da aula anterior, por isso as transparências estão todas juntas.

- ***Statistical learning should not be viewed as a series of black boxes.***

- **No single approach will perform well in all possible applications.**

- **Without understanding all of the cogs (engrenagens ou mecanismos) inside the box, or the interaction between those cogs, it is impossible to select the best box.**

- **Hence, we have attempted to carefully describe the model, intuition, assumptions, and trade-offs behind each of the methods that we consider**.

Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. In this setting, $\hat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

Exemplo 1

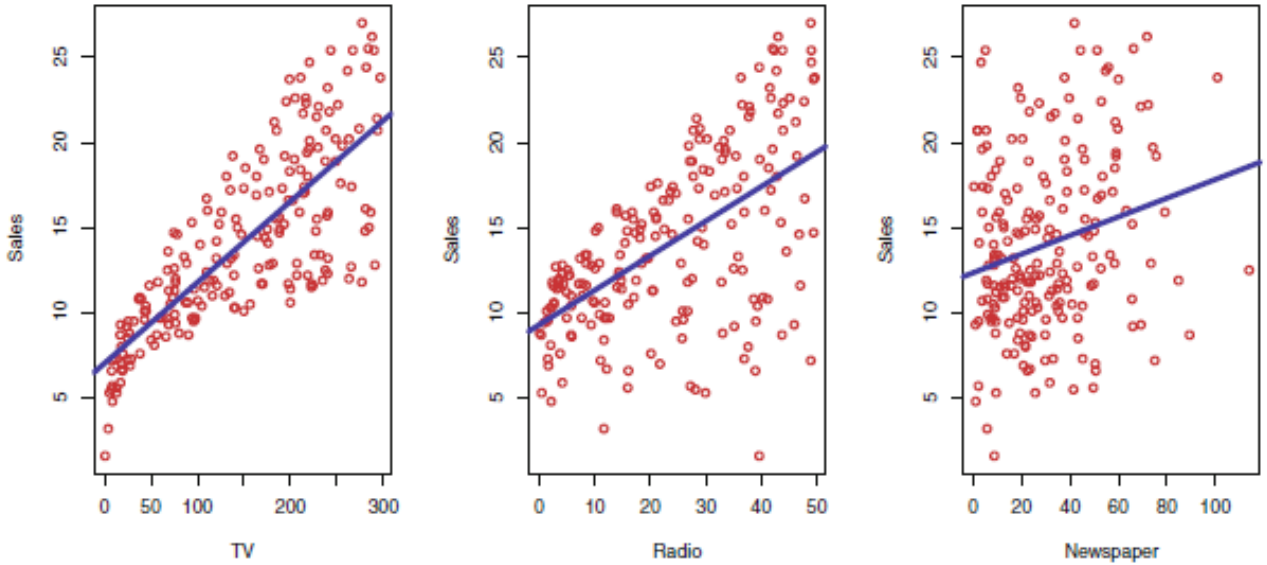


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter* 3*. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

Exemplo 2

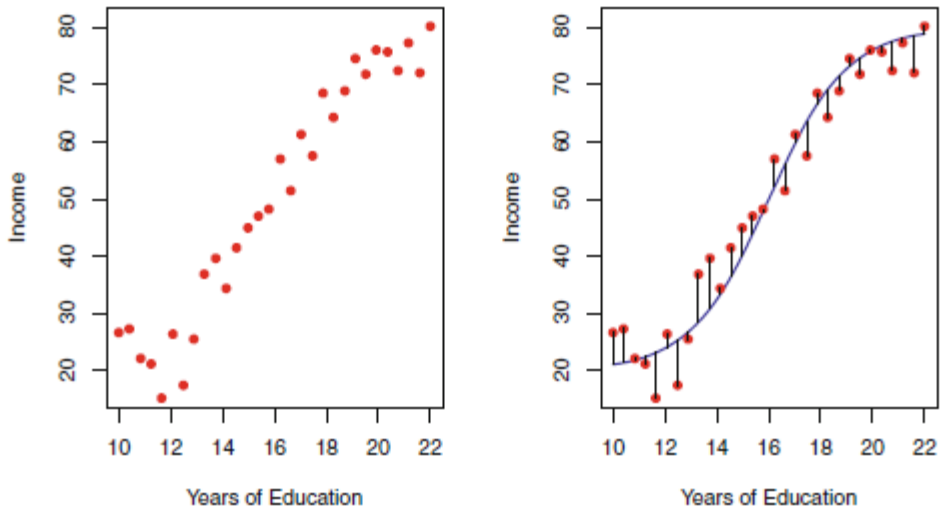


FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

In essence, statistical learning refers to a set of approaches for estimating f.

2.1.1 Why Estimate f?

There are two main reasons that we may wish to estimate f: *prediction* and *inference*. We discuss each in turn.

Predição

Inferência

Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. In this setting, $\hat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

- Temos acesso aos dados de entrada (inputs X) mas é difícil obter os dados de saída (the output Y)
- A forma exata de f não é tão importante, desde que forneça valores de Y com certa precisão.
- No exemplo 2 é mostrada uma f, mas é conhecida por que os dados foram obtidos por uma simulação que usou a própria f

Inference

We are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. We instead want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, \ldots, X_p$. Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in answering the following questions:

- Quais são as variáveis preditoras (input) que interferem de fato no valor de Y (reposta/output) ?
- Como as variáveis preditoras interferem em Y: positiva ou negativamente?
- Qual de fato é a relação das variáveis preditoras e a resposta Y: é uma relação linear?

Exemplo de Inferência

In contrast, consider the `Advertising` data illustrated in Figure 2.1. One may be interested in answering questions such as:

- *Which media contribute to sales?*
- *Which media generate the biggest boost in sales?* or
- *How much increase in sales is associated with a given increase in TV advertising?*

Propaganda (TV, Rádio e Jornal) versus Vendas

Exemplo de Predição

- Company that is interested in conducting a direct-marketing campaign.
- The goal is to identify individuals who will respond positively to a mailing, based on observations of demographic variables measured on each individual.
- In this case, the demographic variables serve as predictors, and response to the marketing campaign (**either positive or negative**) serves as the outcome.
- The company is not interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply wants an accurate model to predict the response using the predictors. This is an example of modelling for prediction.

Exemplo de inferência

This situation falls into the inference paradigm. Another example involves modeling the brand of a product that a customer might purchase based on variables such as price, store location, discount levels, competition price, and so forth. In this situation one might really be most interested in how each of the individual variables affects the probability of purchase. For instance, *what effect will changing the price of a product have on sales?* This is an example of modeling for inference.

- **Problema:** escolha da marca de um **produto** versus **consumidores**
  - Quais as características do produto (de uma marca) que levam um consumidor a comprá-lo? Preço, localização nas gondolas, promoções.
  - Se o preço do produto aumentar, o quanto isso influi no volume de vendas?

Mudar uma variável preditora e saber como a variável resposta se comporta.

Métodos Paramétricos e não-Paramétricos

Regressão Linear versus SPLINE

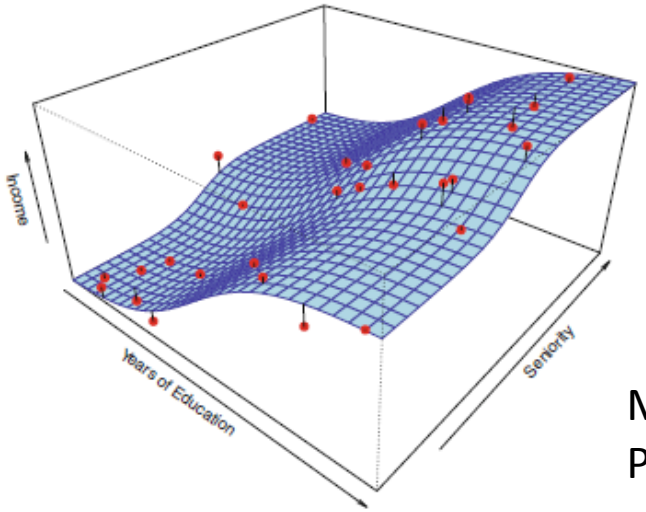


FIGURE 2.3. *The plot displays* `income` *as a function of* `years of education` *and* `seniority` *in the* `Income` *data set. The blue surface represents the true underlying relationship between* `income` *and* `years of education` *and* `seniority`*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

Modelos Paramétricos e não-paramétricos.

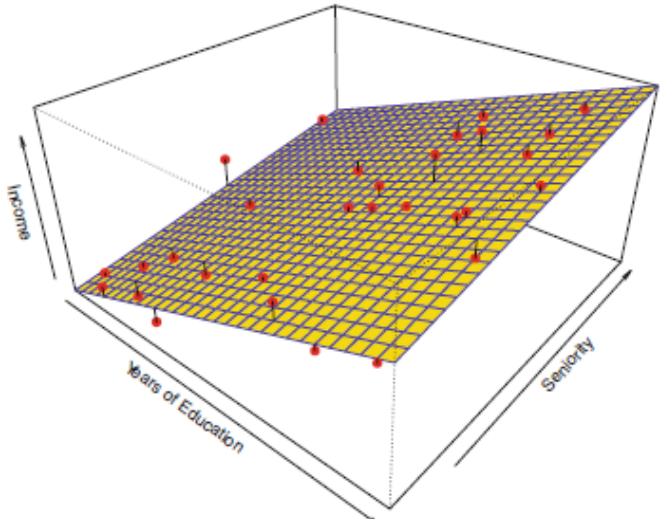


FIGURE 2.4. *A linear model fit by least squares to the* `Income` *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

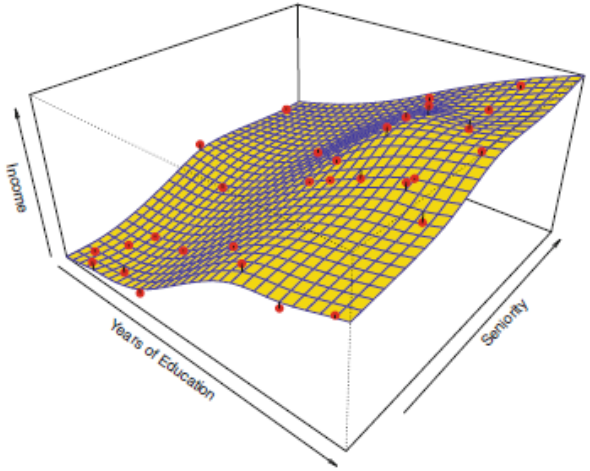


FIGURE 2.5. *A smooth thin-plate spline fit to the* `Income` *data from Figure 2.3 is shown in yellow; the observations are displayed in red. Splines are discussed in Chapter 7.*

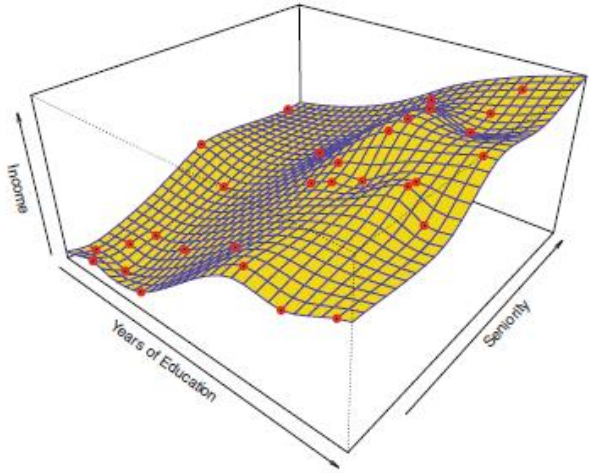


FIGURE 2.6. *A rough thin-plate spline fit to the* Income *data from Figure 2.3. This fit makes zero errors on the training data.*

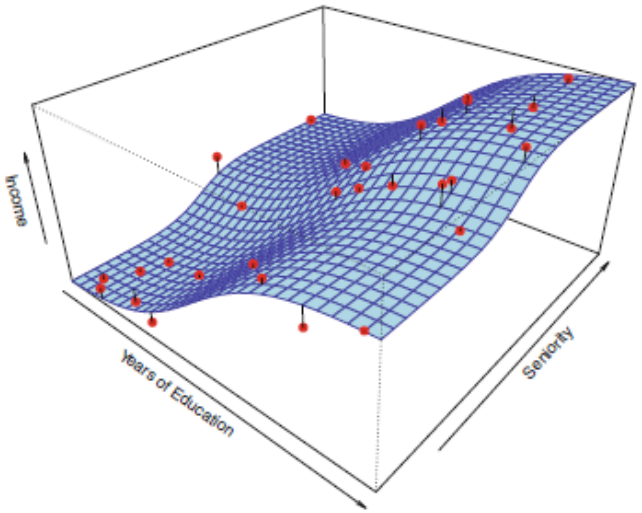


FIGURE 2.3. *The plot displays* income *as a function of* years of education *and* seniority *in the* Income *data set. The blue surface represents the true underlying relationship between* income *and* years of education *and* seniority*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

O spline não consegue representar f de forma que possamos fazer predições /inferencia

2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

- linear regression is a relatively inflexible approach, because it can only generate linear functions. (**Regressão Linear: abordagem pouco flexível**)

- *why would we ever choose to use a more restrictive method instead of a very flexible approach? (**Flexibilidade versus Restrição**)*

- If we are mainly interested in **inference**, then restrictive models are much more **interpretable**.
- **Abordagens Rígidas: Inferência -> Interpretação Facilitada**

- For instance, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between *Y* and *X_1,X_2, . . . , X_p*.

- In contrast, very **flexible approaches**, such as the splines displayed in Figures 2.5 and 2.6, can lead to such complicated estimates of *f* that it is difficult to understand how any individual predictor is associated with the response.

- **Abordagens Mais Flexíveis: Difícil Interpretação dos Resultados -> Não são boas para a Inferência Mas são boas para predição**

In some settings, however, we are only **interested in prediction**, and **the interpretability of the predictive model** is simply not of interest.

- For instance, if we seek to develop an **algorithm to predict the price of a stock**, our sole requirement for the algorithm is that it **predict accurately— interpretability is not a concern.**

- **Surprisingly,** this is not always the case! We will often obtain more accurate predictions using a less flexible method.

- This phenomenon, which may seem **counterintuitive** at first glance, has to do with the **potential for overfitting in highly flexible methods**.

Overfitting (sobreajuste) é um termo usado em estatística para descrever quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados. É comum que a amostra apresente desvios causados por erros de medição ou fatores aleatórios. (Wikipédia)

Supervised Versus Unsupervised Learning

Most statistical learning problems fall into one of two categories:

- **Supervised.**
- **Unsupervised.**

Até agora só falamos de Aprendizado Supervisionado.

Unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1, . . . , n, we observe a vector of measurements x_i but no associated response y_i.

- It is not possible to fit a linear regression model, since there is no response variable to predict.
- In this setting, we are in some sense working blind;
- The situation is referred to as unsupervised because we lack a response variable that can supervise our analysis.

One statistical learning tool that we may use in this setting is **CLUSTER ANALYSIS**, or clustering. The goal of cluster analysis is to ascertain, on the basis of x_1, . . . , x_n, whether the observations fall into analysis relatively distinct groups.

For **example**, in a market segmentation study we might observe multiple characteristics (variables) for potential customers, such as

- **zip code**, **(CEP)**
- **family income**, **(Renda da Família)**
- **shopping habits** **(Hábitos de Compra)**.

We might believe that the **customers fall into different groups**, such as

- **big spenders**, **(gastadores)**
- **low spenders**. **(“econômico”)**

If the information about each customer’s spending patterns **were available**, then a supervised analysis would be possible.

However**, this information is not available**—that is, **we do not kno**w whether each potential customer is a big spender or not.

In this setting, we can try to cluster the customers on the basis of the variables measured, in order to identify distinct groups of potential customers**. (Temos características dos clientes.)**

Identifying such groups can be of interest because it might be that the **groups differ with respect to some property of interest, such as spending habits.**

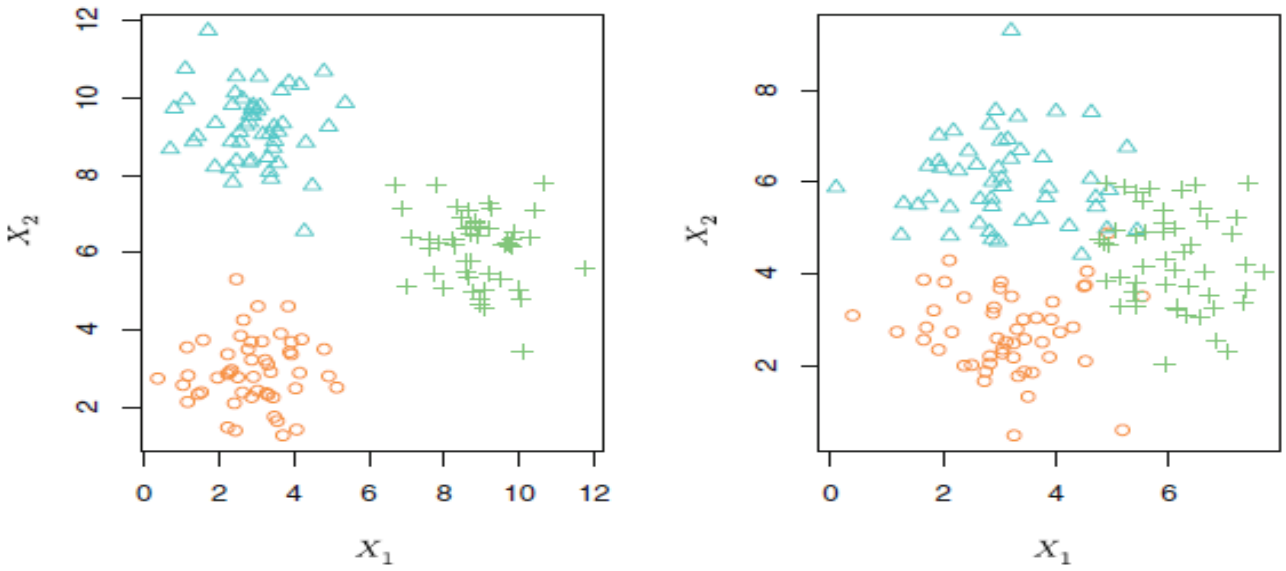
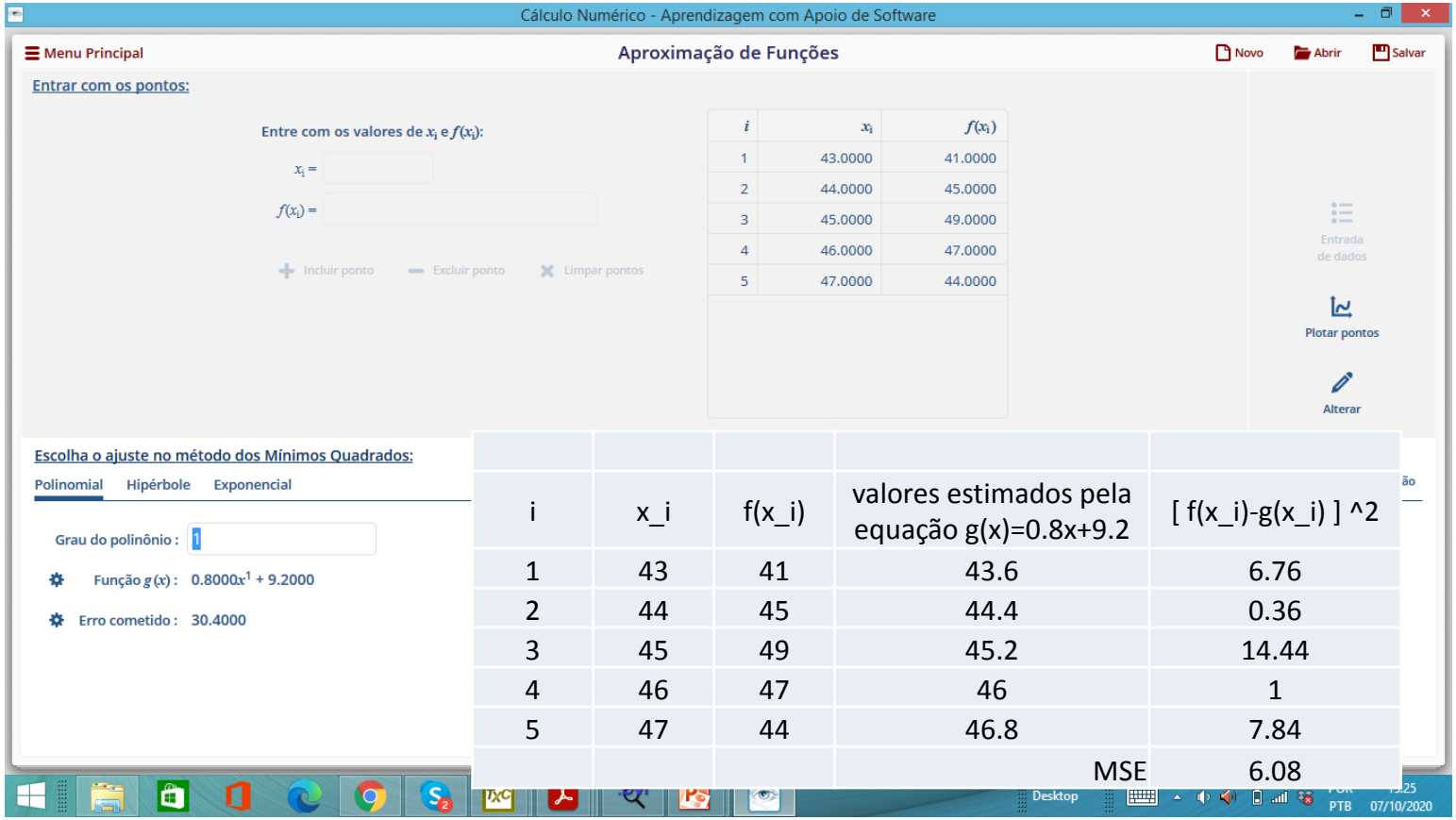


FIGURE 2.8. *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

2.2 Assessing Model Accuracy

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

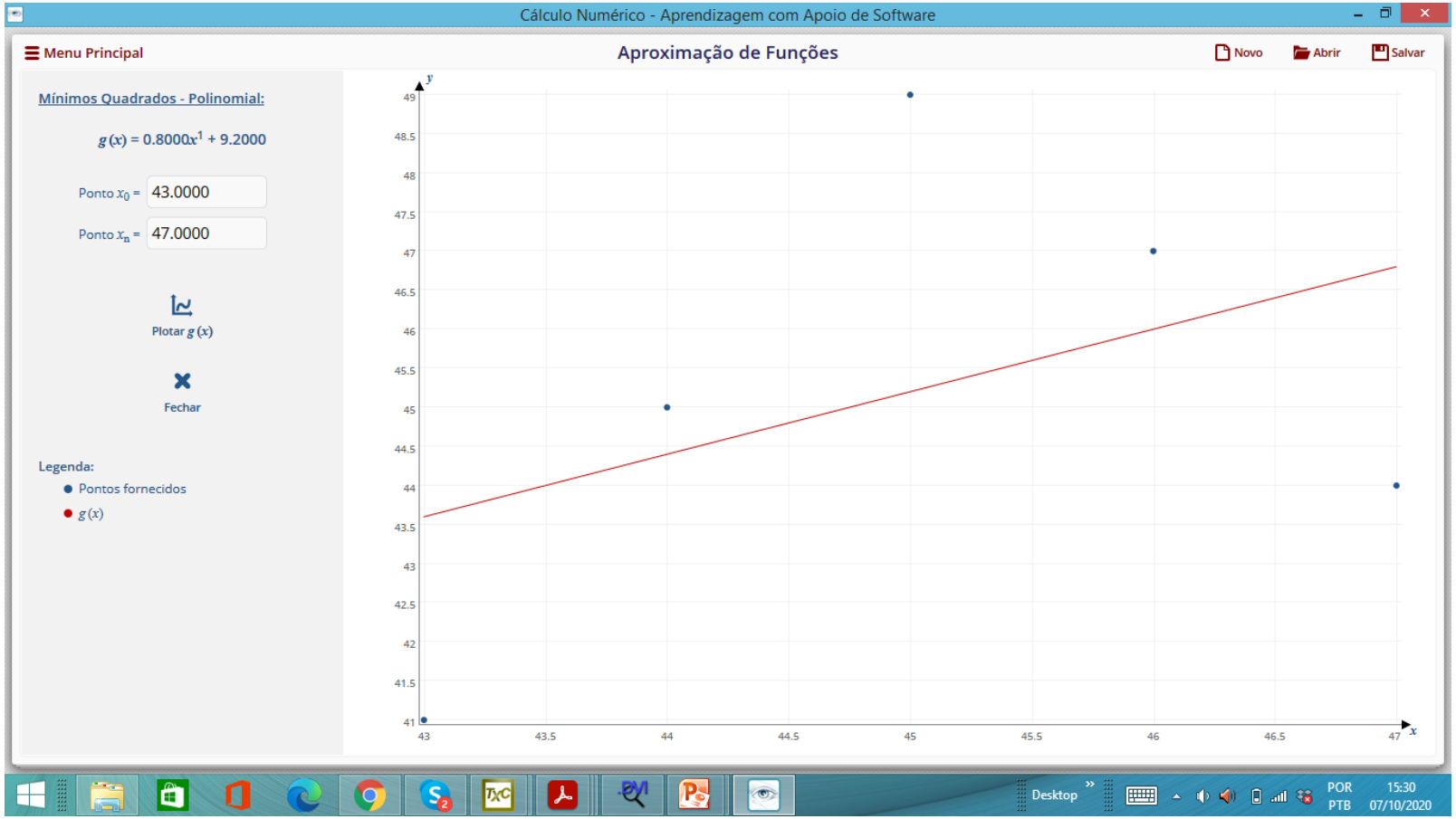


i	x_i	$f(x_i)$
1	43.0000	41.0000
2	44.0000	45.0000
3	45.0000	49.0000
4	46.0000	47.0000
5	47.0000	44.0000

i	x_i	f(x_i)	valores estimados pela equação g(x)=0.8x+9.2	[f(x_i)-g(x_i)] ^2
1	43	41	43.6	6.76
2	44	45	44.4	0.36
3	45	49	45.2	14.44
4	46	47	46	1
5	47	44	46.8	7.84
			MSE	6.08

1. **Encontrar a melhor f que "fita" esses dados.** **Se propusermos uma regressão, podemos usar o método dos mínimos quadrados para determinar os coeficientes.**

2. **Após encontrar os valores aproximados pela f, verificamos o MSE.**



$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

The MSE is computed using the training data that was used to fit the model, and so should more accurately be referred to as the training MSE.

Amostra deve ficar dividida em

- **conjunto de treinamento** $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,
- **o conjunto de teste**.

Do conjunto de teste temos o "set MSE" que deve ser mínimo.

Prediction Error:

Prediction error quantifies one of two things:

- In regression analysis, it's a **measure** of how well the model predicts the response variable.
- In **classification** (machine learning), it's a measure of how well samples are classified to the correct category.

    - Conj. Treinamento
    - Conj. de Teste

The average squared prediction error

Mean Squared Prediction Error (MSPE)

- MSPE summarizes the predictive ability of a model.
- Ideally, this value should be close to zero, which means that your predictor is close to the true value.
- The concept is similar to Mean Squared Error (MSE), which is a measure of the how well an estimator measures a parameter (or how close a regression line is to a set of points).
- MSE is a measure of an estimator's fit,
- MSPE is a measure of a predictor's fit— or how well it predicts the true value.

Exemplo

Suppose that we are interested test data in developing an algorithm to predict a stock's price based on previous stock returns. We can train the method using stock returns from the past 6 months. But we don't really care how well our method predicts last week's stock price. We instead care about how well it will predict tomorrow's price or next month's price.

Exemplo

Suppose that we have clinical measurements (e.g. weight, blood pressure, height, age, family history of disease) for a number of patients, as well as information about whether each patient has diabetes. We can use these patients to train a statistical learning method to predict risk of diabetes based on clinical measurements. In practice, we want this method to accurately predict diabetes risk for *future patients* based on their clinical measurements. We are not very interested in whether or not the method accurately predicts diabetes risk for patients used to train the model, since we already know which of those patients have diabetes.