Statistical learning & machine learning

Introduction

Prof. Dr. Jan Kirenz
HdM Stuttgart

HOCHSCHULE DER MEDIEN

What is statistical learning?

Suppose we want to **predict sales** of a product

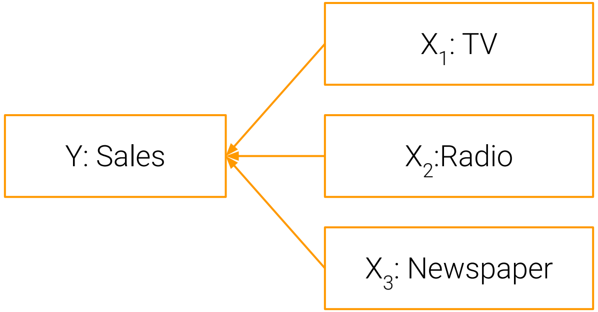
This is our **data**

- Variables:
  - Sales
  - TV advertising spendings
  - Radio advertising spendings
  - Newspaper advertising spendings

Source: James et al (2021)

We first need to define the outcome and predictor variables

- **Outcome** variable (response, dependent variable):
  - Sales (Y)
- **Input** variable (predictors, independent variables, features)
  - TV (X_1)
  - Radio (X_2)
  - Newspaper (X_3)



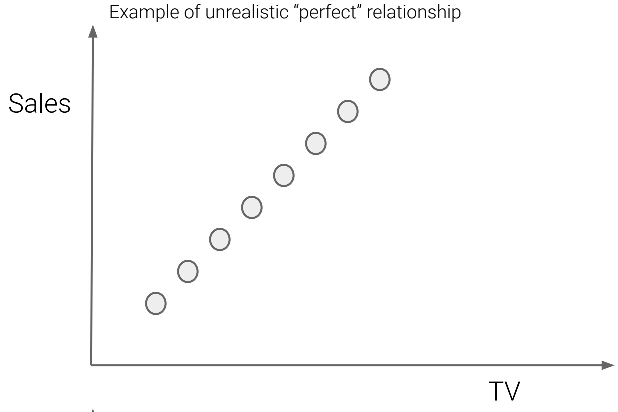
Source: James et al (2021)

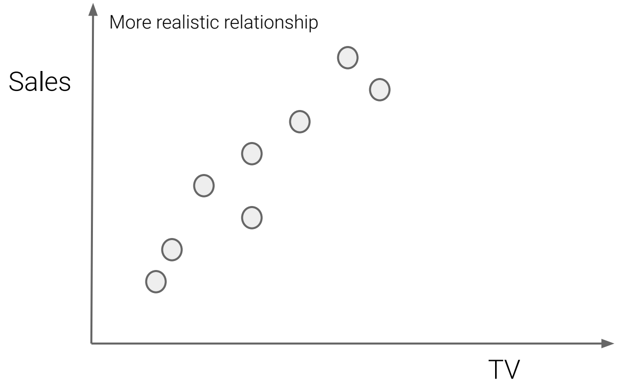
Statistical learning refers to a set of approaches for **estimating** a **function** (f) between Y and X.

Y: Sales = **function**(X_1: TV X_2:Radio X_3: Newspaper)

Example for sales and TV

- **Outcome** variable (response, dependent variable):
  - **Sales** (Y)
- **Input** variable (predictors, independent variables, features)
  - **TV** (X_1)
  - Radio (X_2)
  - Newspaper (X_3)





Source: James et al (2021)

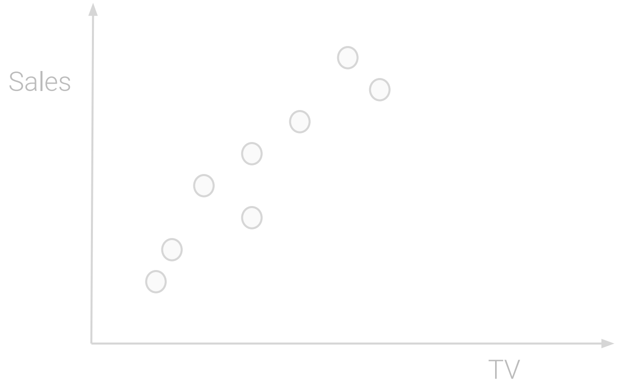
We don't know the true association between our outcome and the predictors

The "true" but unknown relationship

$$Y = f(X) + \epsilon.$$

Y: Sales = function(X_1: TV | X_2:Radio | X_3: Newspaper) + noise

- f is some fixed but unknown function of $X_1, \ldots, X_p$,
- $X = (X_1, X_2, \ldots, X_p)$
- ϵ is a random error term (noise), which is independent of X and has mean zero



Source: James et al (2021)

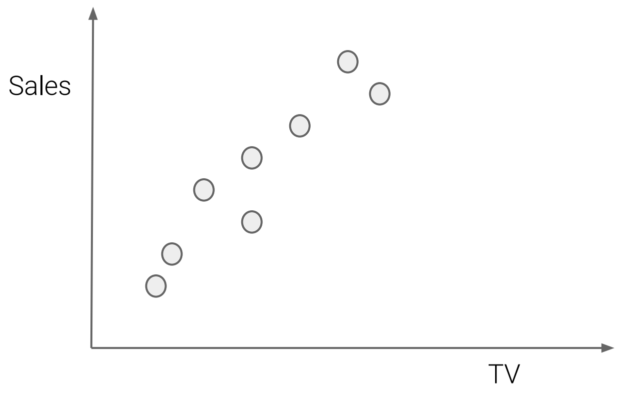
We don't know the true association between our outcome and the predictors

The "true" but unknown relationship

$$Y = f(X) + \epsilon.$$

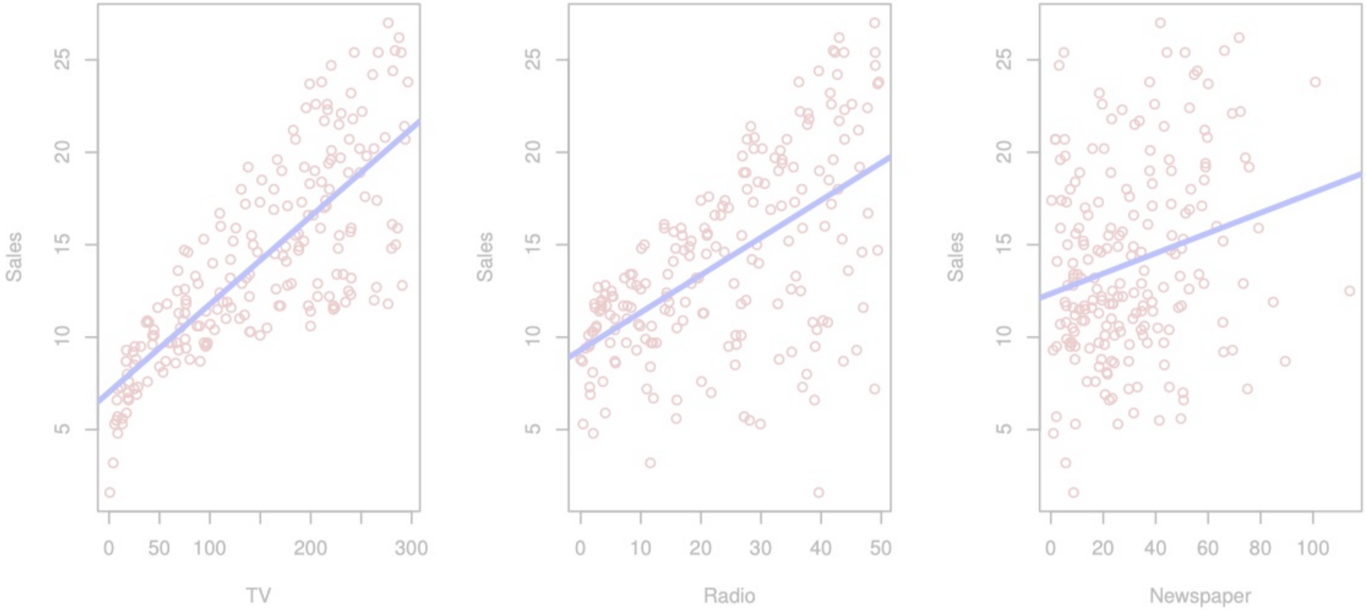
$\boxed{Y\text{: Sales}}$ = function($\boxed{X_1\text{: TV}}$ $\boxed{X_2\text{:Radio}}$ $\boxed{X_3\text{: Newspaper}}$) + noise

- f is some fixed but unknown function of $X_1, \ldots, X_p$,
- $X = (X_1, X_2, \ldots, X_p)$
- ϵ is a random error term (noise), which is independent of X and has mean zero



Source: James et al (2021)

We want to make predictions ($\hat{Y}$) for given values of X by using a function (model)



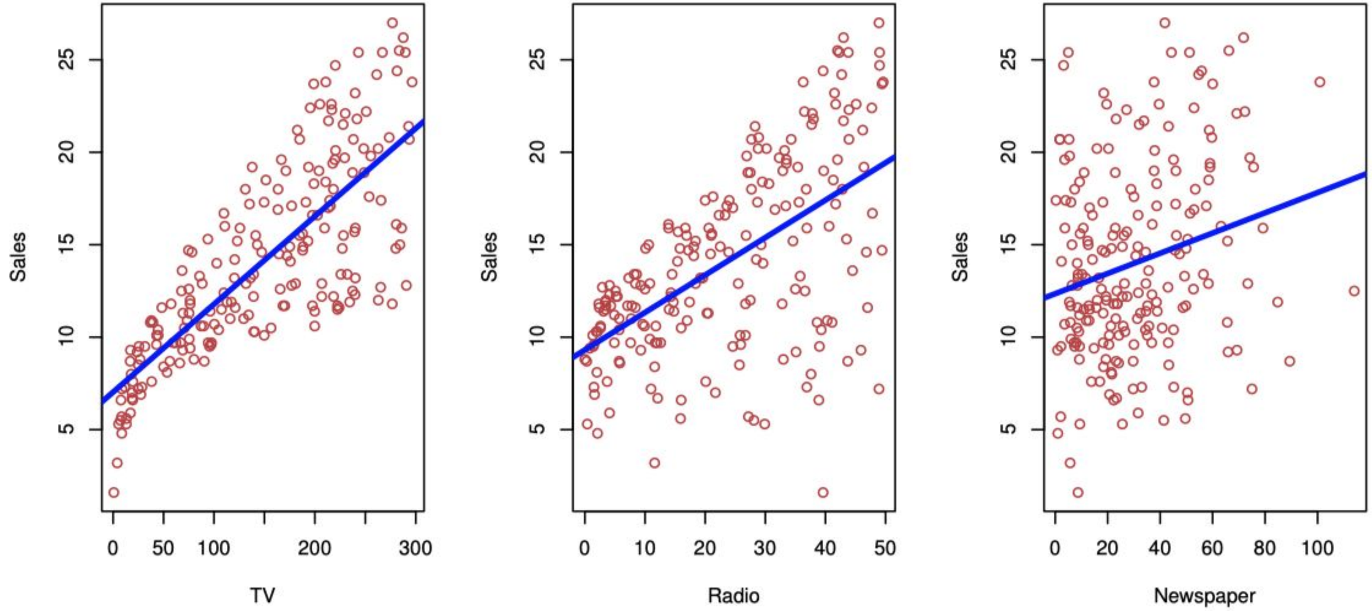
Prediction

$$\hat{Y} = \hat{f}(X)$$

- $\hat{f}$ represents our estimate for f
- $\hat{Y}$ represents the resulting prediction
- $X = (X_1, X_2, \ldots, X_p)$
- *ϵ averages to zero, so we don't need to state it here*

Source: James et al (2021)

We want to make predictions (Ŷ) for given values of X by using a function (model)



Prediction

$$\hat{Y} = \hat{f}(X)$$

- $\hat{f}$ represents our estimate for f
- Ŷ represents the resulting prediction
- $X = (X_1, X_2, \ldots, X_p)$
- *ϵ averages to zero, so we don't need to state it here*

Source: James et al (2021)

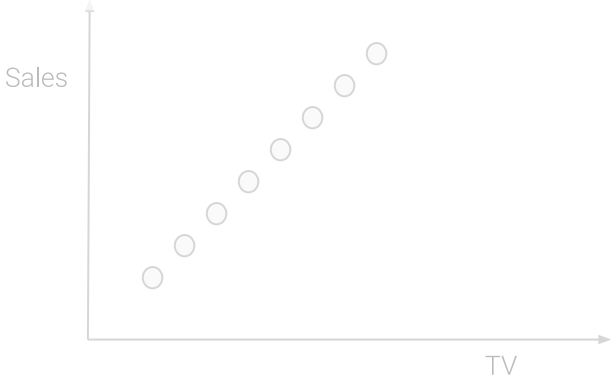
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \boxed{\underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

“natural” noise

$E(Y-\hat{Y})^2$= Expected value of the squared difference between the actual and predicted value

Example of unrealistic “perfect” relationship with **no noise**



Source: James et al (2021)

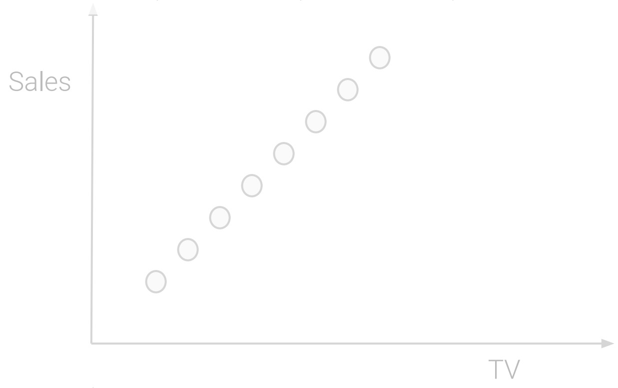
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\
&= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

"natural" noise

$E(Y-\hat{Y})^2$ = Expected value of the squared difference between the actual and predicted value

Example of unrealistic "perfect" relationship with **no noise**



More realistic relationship with **natural noise**

Source: James et al (2021)

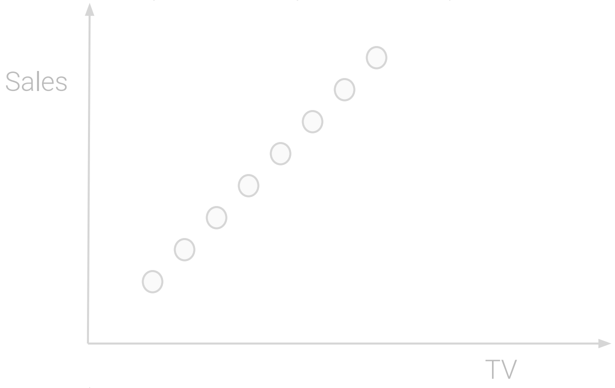
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \boxed{\underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}}} + \underbrace{\operatorname{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

"natural" noise

$E(Y-\hat{Y})^2$ = Expected value of the squared difference between the actual and predicted value

Example of unrealistic "perfect" relationship with **no noise**



More realistic relationship with **natural noise**

Source: James et al (2021)

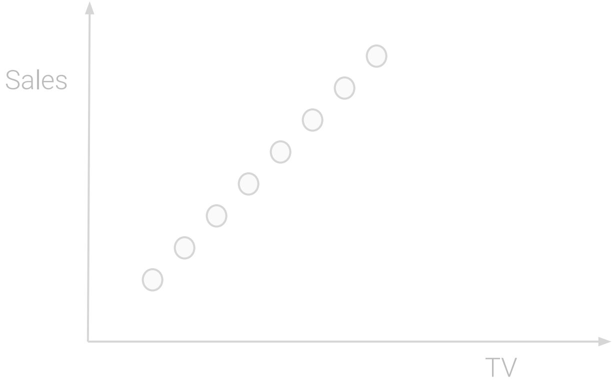
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$\mathrm{E}(Y-\hat{Y})^2 = \mathrm{E}[f(X)+\epsilon-\hat{f}(X)]^2$$
$$= \boxed{\underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}$$

"natural" noise

$\mathrm{E}\,(Y-\hat{Y})^2$= Expected value of the squared difference between the actual and predicted value

Example of unrealistic "perfect" relationship with **no noise**



More realistic relationship with **natural noise**

Source: James et al (2021)

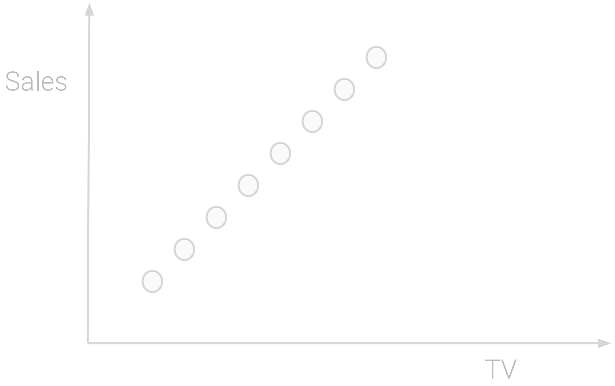
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$E(Y-\hat{Y})^2 = E[f(X)+\epsilon-\hat{f}(X)]^2$$
$$= \boxed{\underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}$$

"natural" noise

$E(Y-\hat{Y})^2$ = Expected value of the squared difference between the actual and predicted value

Example of unrealistic "perfect" relationship with **no noise**



More realistic relationship with **natural noise**

Source: James et al (2021)

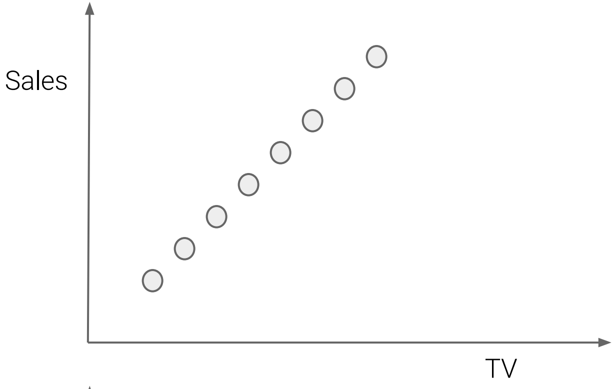
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \boxed{\underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

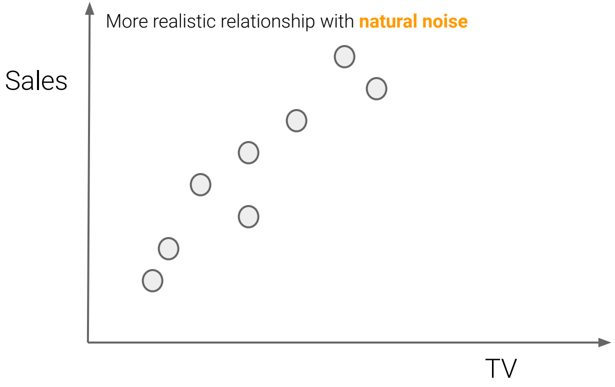
“natural” noise

$E(Y-\hat{Y})^2$ = Expected value of the squared difference between the actual and predicted value

Example of unrealistic “perfect” relationship with **no noise**



More realistic relationship with **natural noise**



Source: James et al (2021)

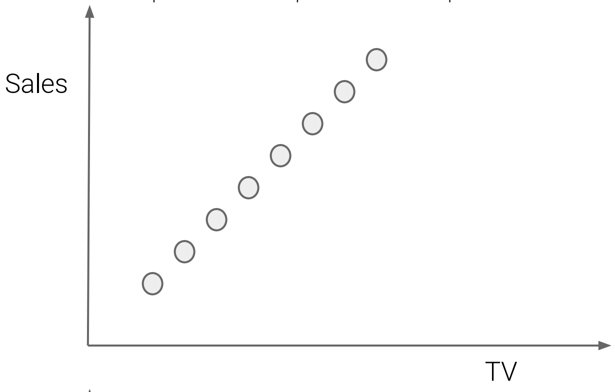
We want to **predict Ŷ** with the aim to **minimize** the **reducible error**

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{\boxed{[f(X)-\hat{f}(X)]^2}}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

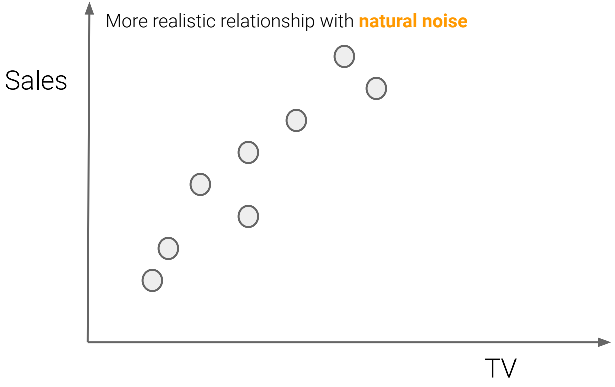
“natural” noise

$E(Y-\hat{Y})^2$ = Expected value of the squared difference between the actual and predicted value

Example of unrealistic “perfect” relationship with **no noise**



More realistic relationship with **natural noise**



Source: James et al (2021)

Statistical learning is not just about **predictions** but also about **inference**

We want to figure out the association between our outcome and input variables.

Typical questions:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Source: James et al (2021)

Statistical learning is not just about **predictions** but also about **inference**

We want to figure out the association between our outcome and input variables.

Typical questions:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Source: James et al (2021)

Example of direct marketing campaign

Goal: identify individuals who are likely to respond positively to a mailing

- **Outcome**: response to the marketing campaign (either positive or negative)
- **Predictors**: demographic variables (age, gender, address,...)
- 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 to accurately predict the response using the predictors.
- This is an example of modeling for **prediction**

Source: James et al (2021)

In our **sales** example, we would be interested in answering questions like ...

- Which media are associated with sales?
- Which media generate the biggest boost in sales?
- How large of an increase in sales is associated with a given increase in TV advertising?
- This is an example of modeling for **inference**

Source: James et al (2021)

Difference between statistical learning and machine learning

Prediction & inverence

Prediction wants to accurately predict a response using some predictors

Focus of machine learning

Despite convincing prediction results, the lack of an explicit model can make ML solutions difficult to interpret and directly relate to existing theoretical knowledge.

Inference is about understanding the relationship between the response and predictors

Focus of statistical learning

E.g., compute a quantitative measure of confidence that a discovered relationship describes a 'true' effect that is unlikely to result from noise.

ML vs statistics

ML requires us to choose a predictive algorithm by relying on its empirical capabilities.

Statistics requires us to choose a model that incorporates our knowledge of the system.

Source: Bzdok,, Altman & Krzywinski (2018)

nature methods

This Month | Published: 03 April 2018

Points of Significance

Statistics versus machine learning

Danilo Bzdok, Naomi Altman & Martin Krzywinski

Nature Methods **15**, 233–234 (2018) | Download Citation

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.

Two major goals in the study of biological systems are inference and prediction. Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future behavior, such as whether a mouse with a given gene expression pattern has a disease. Prediction makes it possible to identify best courses of action (e.g., treatment choice) without requiring understanding of the underlying mechanisms. In a typical research project, both inference and prediction can be of value—we want to know how biological processes work and what will happen

Source: Bzdok,, Altman & Krzywinski (2018)

Literature

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of Significance: Statistics versus machine learning. Nature Methods, 15(4), 233-234. DOI: 10.1038/nmeth.4642

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Statistical learning. In An introduction to statistical learning. Springer, New York, NY.