IT496: Introduction to Data Mining

Lecture 08

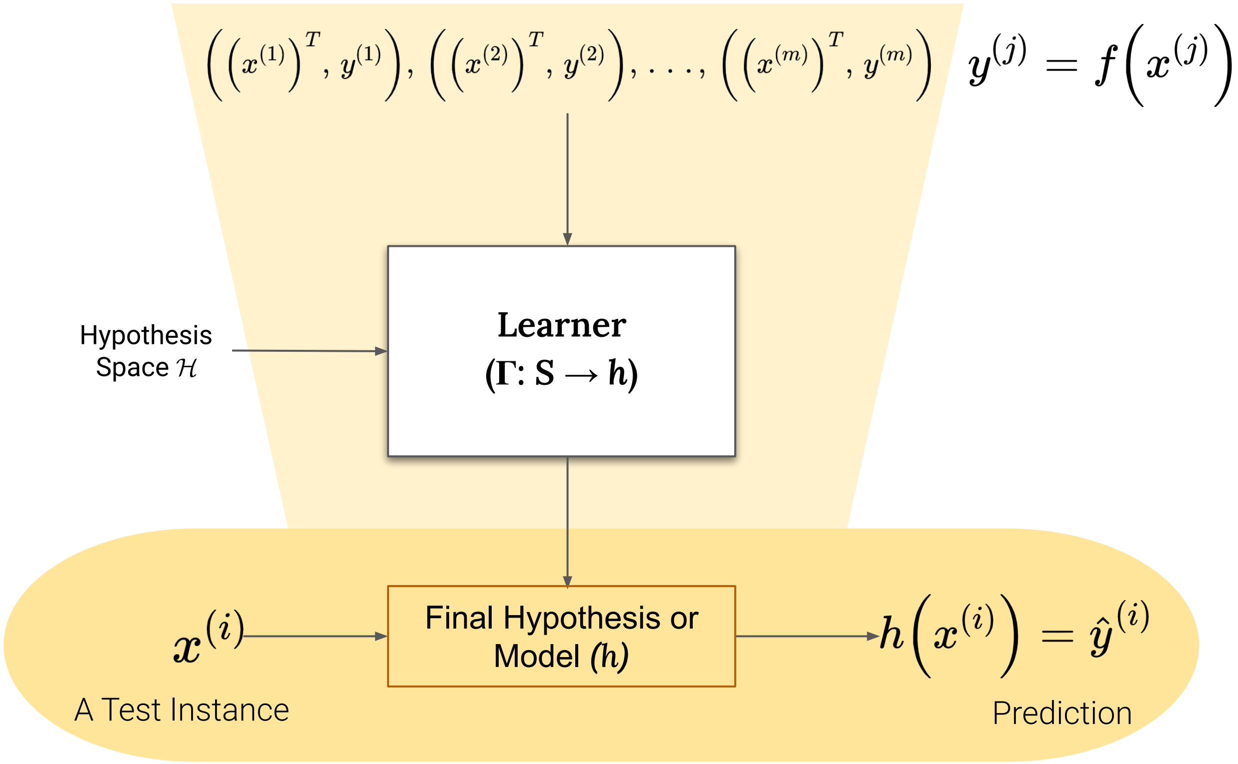
Choosing a Hypothesis Space

[Inductive Bias, Bias-Variance Trade-off, Model Complexity and Expressiveness Trade-off]

Arpit Rana

11th August 2023

Supervised Learning Process



Supervised Learning: Example

Problem: whether to wait for a table at a restaurant.

Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*\$\$\$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*\$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*\$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*\$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*\$\$\$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*\$\$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*\$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*\$\$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*\$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*\$\$\$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*\$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*\$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

- **Alternate**: whether there is a suitable alternative restaurant nearby.
- **Bar**: whether the restaurant has a comfortable bar area to wait in.
- **Fri/Sat**: true on Fridays and Saturdays.
- **Hungry**: whether we are hungry right now.
- **Patrons**: how many people are in the restaurant (values are None, Some, and Full).
- **Price**: the restaurant's price range ($, $$, $$$).
- **Raining**: whether it is raining outside.
- **Reservation**: whether we made a reservation.
- **Type**: the kind of restaurant (French, Italian, Thai, or Burger).
- **WaitEstimate**: host's wait estimate: 0–10, 10–30, 30–60, or >60 minutes.

Supervised Learning: Example

Problem: whether to wait for a table at a restaurant.

Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	$y_1 = Yes$
$\mathbf{x}_2$	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	$y_2 = No$
$\mathbf{x}_3$	No	Yes	No	No	Some	$	No	No	Burger	0–10	$y_3 = Yes$
$\mathbf{x}_4$	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	$y_4 = Yes$
$\mathbf{x}_5$	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	$y_5 = No$
$\mathbf{x}_6$	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	$y_6 = Yes$
$\mathbf{x}_7$	No	Yes	No	No	None	$	Yes	No	Burger	0–10	$y_7 = No$
$\mathbf{x}_8$	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	$y_8 = Yes$
$\mathbf{x}_9$	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	$y_9 = No$
$\mathbf{x}_{10}$	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	$y_{10} = No$
$\mathbf{x}_{11}$	No	No	No	No	None	$	No	No	Thai	0–10	$y_{11} = No$
$\mathbf{x}_{12}$	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	$y_{12} = Yes$

Training Data

$y = f(x)$

Unknown Target function f

Instances

Instance Space (X) 2 x 2 x 2 x 2 x 3 x 3 x 2 x 2 x 4 x 4 = 9216

Size of Hypothesis Space ($|\mathcal{H}|$) of Boolean Functions $\boxed{= 2^{9216}}$

Hypothesis Space vs. Hypothesis

What do we mean by a Hypothesis Space (a.k.a. Model Class) and a hypothesis?

There are three different levels of specificity for using the term *Hypothesis* or *Model*:

- a broad *hypothesis space* (like "polynomials"),
- a *hypothesis space* with *hyperparameters* filled in (like "degree-2 polynomials"), and
- a specific hypothesis with all *parameters* filled in (like $5x^2 + 3x - 2$).

Hypothesis Space vs. Hypothesis

What do we mean by a Hypothesis Space (a.k.a. Model Class) and a hypothesis?

There are three different levels of specificity for using the term *Hypothesis* or *Model*:

$$\boxed{y = f(x) = bx + a}$$
$$= f(x) = e^{-bx}$$
$$= f(x) = \sin(bx)$$
$$\boxed{= f(x) = bx^2}$$
$$\boxed{= f(x) = \sqrt{bx + a}}$$

$\xrightarrow{\text{Polynomials}}$

$$\boxed{y = f(x) = bx + a}$$
$$= f(x) = bx^2$$
$$= f(x) = \sqrt{bx + a}$$

$\xrightarrow{\text{Hyperparameter: degree=1}}$

$$y = f(x) = bx + a$$

Parameters:
a=2, b=3

$$f(x) = 3x + 2$$

Hypothesis Space vs. Hypothesis

How do we choose a good Hypothesis Space or Model Class?

$$\boxed{y = f(x) = bx + a}$$
$$= f(x) = e^{-bx}$$
$$= f(x) = \sin(bx)$$
$$\boxed{= f(x) = bx^2}$$
$$\boxed{= f(x) = \sqrt{bx + a}}$$

Polynomials

$$\boxed{y = f(x) = bx + a}$$
$$= f(x) = bx^2$$
$$= f(x) = \sqrt{bx + a}$$

Hyperparameter:
degree=1

$$y = f(x) = bx + a$$

Parameters:
a=2, b=3

$$f(x) = 3x + 2$$

Hypothesis Space / Representation / Model Class Selection
(popularly known as **Model Selection**)

Optimization
or **Training**

Choosing the Hypothesis Space

Hypothesis Space Selection is Subjective

Most probable hypothesis given the data -

$$h^{\cdot} = \underset{\boldsymbol{h} \in \mathcal{H}}{\arg max}\ P(h \mid S) \quad \equiv \quad h^{\cdot} = \underset{\boldsymbol{h} \in \mathcal{H}}{\arg max}\ P(S \mid h) \boxed{P(h)}$$

- We can say that the prior probability *P*(*h*) is high for a smooth degree-1 or -2 polynomial and lower for a degree-12 polynomial with large, sharp spikes.

Hypothesis Space Selection is Subjective

The observed dataset S alone does not allow us to make conclusions about unseen instances. We need to make some assumptions!

- These assumptions induce the *bias* (a.k.a. *inductive or learning bias*) of a learning algorithm.

- Two ways to induce bias:
    - *Restriction*: Limit the hypothesis space (e.g., degree-2 polynomials)
    - *Preference*: Impose ordering on hypothesis space (e.g., prefer simpler than complex)

Hypothesis Space Selection is not only *subjective* but is *empirical* also.

- Part of hypothesis space selection is <u>*qualitative and subjective*</u>:
  We might select polynomials rather than decision trees based on something that we know about the problem,

  and

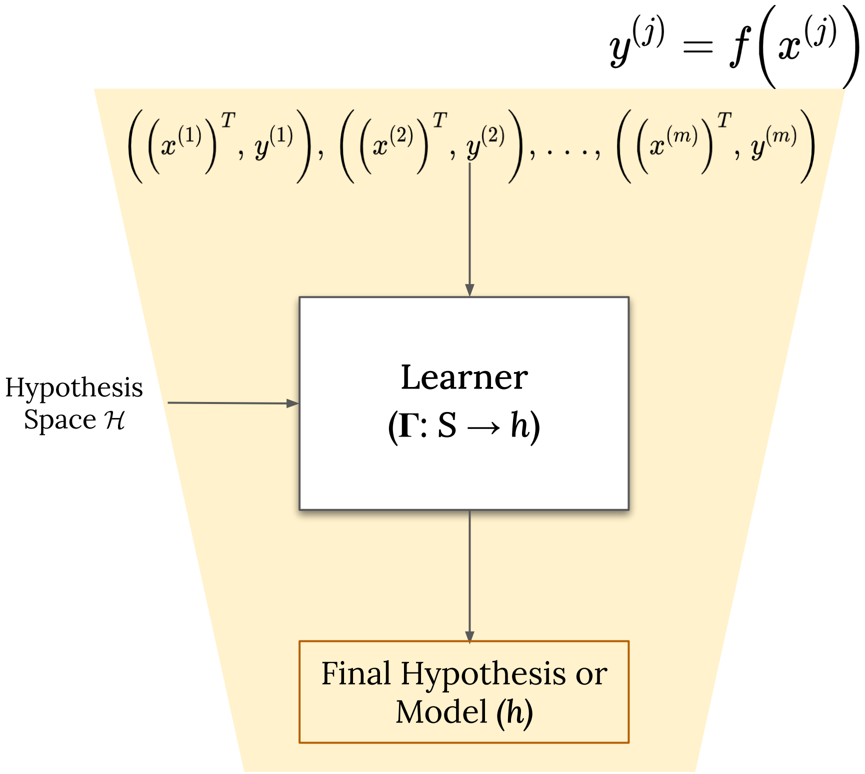
- part is <u>*quantitative and empirical*</u>:
  Within the class of polynomials, we might select Degree = 2, because that value performs best on the validation data set.

Experimental Evaluation of Learning Algorithms

The overall *objective* of the Learning Algorithm is to find a *hypothesis* that -

- is *consistent* (i.e., fits the training data), but more importantly,
- *generalizes well* for previously unseen data.

Experimental Evaluation defines ways to Measure the **Generalizability** of a Learning Algorithm.



Sample Error

The *sample error* of hypothesis *h* with respect to the target function *f* and data sample S is:

$$error_S(h) = \frac{1}{n}\sum_{x \in S} \delta(h(x), f(x))$$

It is *impossible* to asses ***true error***, so we try to estimate it using ***sample error***.

True Error

The *true error* of hypothesis *h* with respect to the target function *f* and the distribution *D* is the probability that *h* will misclassify an instance drawn at random according to *D*:

$$error_D(h) = P_{x \in D}[h(x) \neq f(x)]$$

Generalization Error

Generalization error (a.k.a. *out-of-sample error*) is a measure of how accurately an algorithm is able to predict outcome values for *previously unseen data.*

$$error(h(x),\ f(x)) = \boxed{var(x)} + \boxed{bias(x)^2} + \boxed{\epsilon^2}$$

Variance

Due to the model's sensitivity to small variations in the training data.

It leads to ***overfitting***!

Bias

Due to Wrong Assumptions. Restrictions imposed by -

The Representation Function (i.e., Hypothesis space, such as, linear or quadratic)

The Search Algorithm (e.g., Grid search or Beam search)

It leads to ***underfitting***!

Irreducible Error

Due to the noisiness of the data itself.

The only way to handle it is to clean up the data properly, detect and remove outliers.

Choosing a Hypothesis Space - I

One way to analyze hypothesis spaces is by

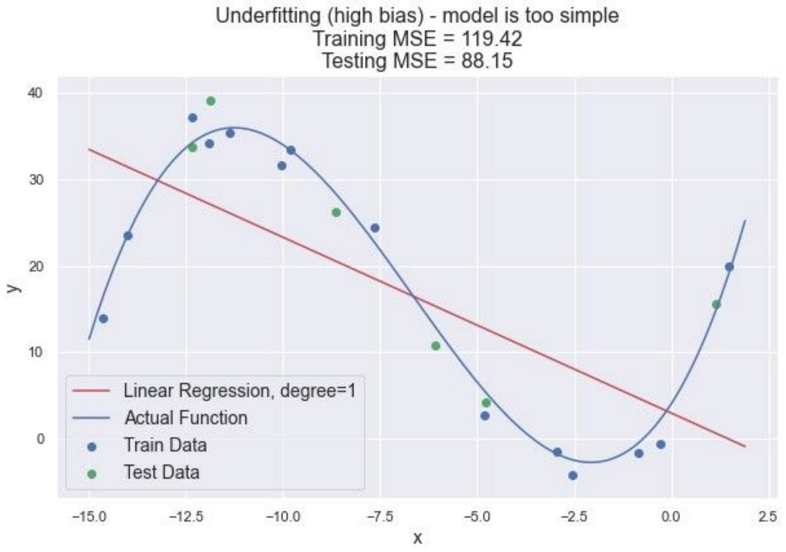
- the <u>bias</u> they impose (regardless of the training data set), and
- the <u>variance</u> they produce (from one training set to another).

Bias

The tendency of a predictive hypothesis to deviate from the expected value when averaged over different training sets.

- Bias often results from restrictions imposed by the hypothesis space.

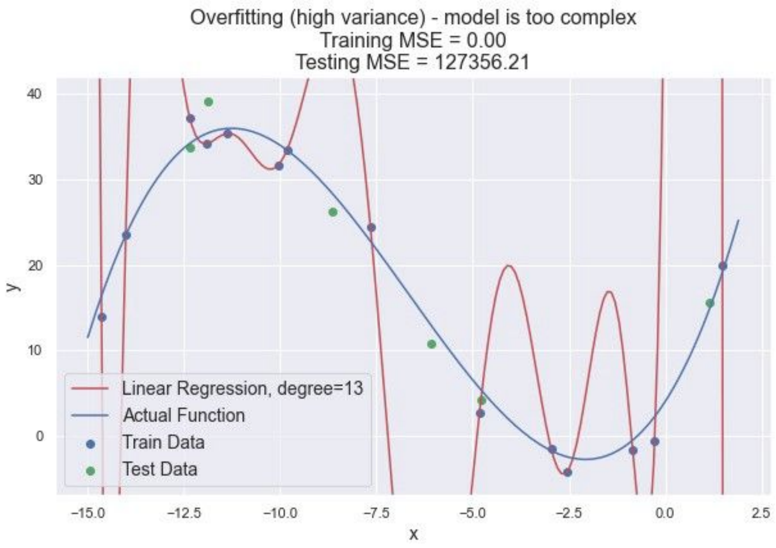
- We say that a hypothesis is ***underfitting*** when it fails to find a pattern in the data.



Variance

The amount of change in the hypothesis due to fluctuation in the training data.

- We say a function is ***overfitting*** the data when it pays too much attention to the particular data set it is trained on.
- It causes the hypothesis to perform poorly on unseen data.



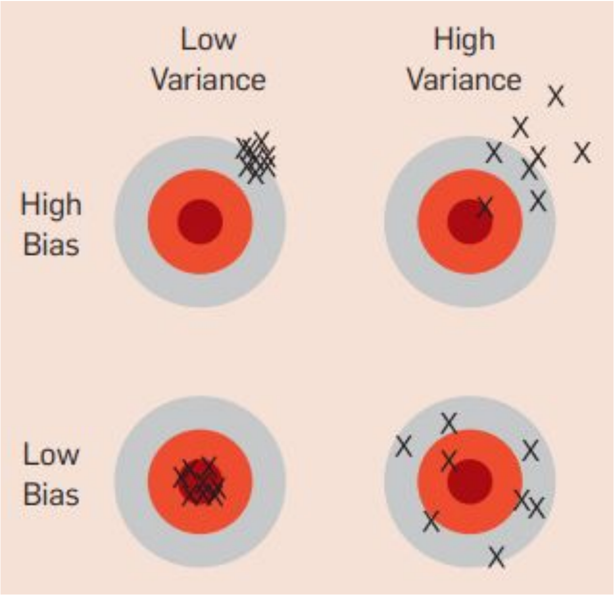
Bias–Variance Trade-off

- **High Variance-High Bias**
  The model is inconsistent and also inaccurate on average

- **Low Variance-High Bias**
  Models are consistent but low on average

- **High Variance-Low Bias**
  Somewhat accurate but inconsistent on average

- **Low Variance-Low Bias**
  Model is consistent and accurate on average



Analogy with throwing darts at a board.

Choosing a Hypothesis Space - II

Another way to analyze hypothesis spaces is by

- the *expressiveness* (i.e., ability of a model to represent a wide variety of functions or patterns) of a hypothesis space, and
    - Can be measured by the size of the hypothesis space
- the *model complexity* (i.e., how intricate the relationships a model can capture) of a hypothesis space.
    - Can be estimated by the number of parameters of a hypothesis

Note-1: Sometimes the term *model capacity* is used to refer to model complexity and expressiveness together.

Note-2: In general, the required amount of training data depends on the model complexity, representativeness of the training sample, and the acceptable error margin.

Choosing a Hypothesis Space - II

There is a tradeoff between the ***expressiveness*** of a hypothesis space and the ***computational complexity*** of finding a good hypothesis within that space.

- Fitting a straight line to data is an easy computation; fitting high-degree polynomials is somewhat harder; and fitting unusual-looking functions may be undecidable.

- After learning h, computing $h(x)$ when h is a linear function is guaranteed to be fast, while computing an arbitrarily complex function may not even guaranteed to terminate.

For example:

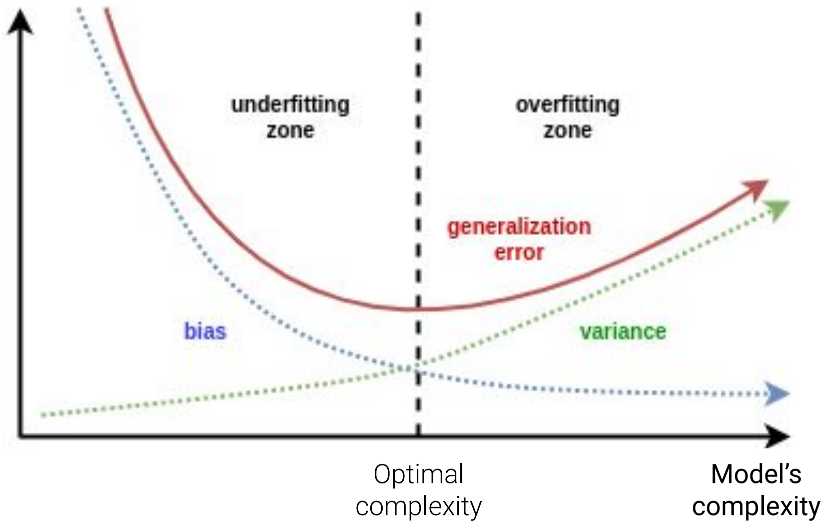
- In Deep Learning, representations are not simple but the $h(x)$ computation still takes only a *bounded number of steps* to compute with appropriate hardware.

Bias-Variance vs. Model's Complexity

The relationship between *bias* and *variance* is closely related to the machine learning concepts of *overfitting*, *underfitting*, and *model's complexity*.

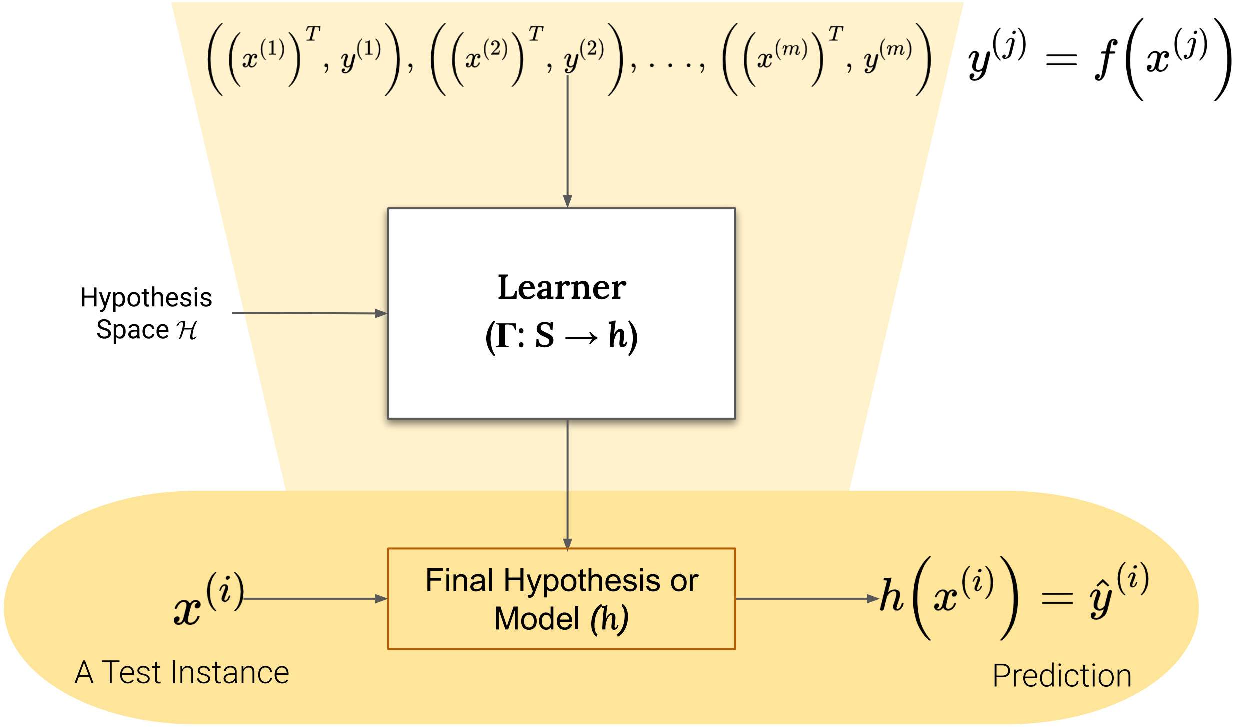
- Increasing a model's complexity typically increases its *variance* and reduces its *bias*.
- Reducing a model's complexity increases its *bias* and reduces its *variance*.

This is why it is called a *tradeoff*.



Learning as a Search

Given a *hypothesis space*, *data*, and a *bias*, the problem of learning can be reduced to one of <u>*search*</u>.



Next lecture

Evaluation

18th August 2023