Key Question: How do we "learn" models from data, and make predictions?

Agenda

- Least Squares Review
- Applications of Least Squares

Least Squares

In real-world models, we often have noise that prevents us from being able to fit a model to our data perfectly.

For example, if we have a matrix $A = \begin{bmatrix} | & | & & | \\ \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \\ | & | & & | \end{bmatrix}$ and a vector $\vec{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$, and we want to find a vector $\vec{x}$ such that transforming $\vec{x}$ by A gives us the vector $\vec{b}$, there may not be an actual solution because of inconsistent equations.

But some noise shouldn't prevent us from finding a model to represent our data!

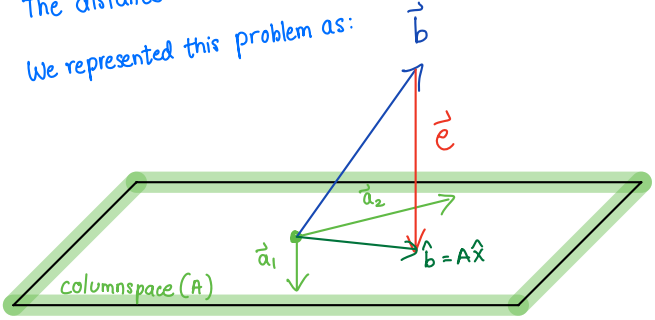
We can try to find the closest model instead, that represents our data as much as possible.

We wanted to find $\vec{x}$ such that $A\vec{x} = \vec{b}$, but it wasn't possible.

Instead, we try to find $\hat{x}$ which makes $A\hat{x}$ as close to $\vec{b}$ as possible. In other words, the length of the difference between $A\hat{x}$ and $\vec{b}$, or the distance between $A\hat{x}$ and $\vec{b}$ should be minimized.

The distance between $A\hat{x}$ and $\vec{b}$ is the length of the error vector, $\vec{e}$.

We represented this problem as:



Goal: $\underset{\hat{x}}{\operatorname{argmin}} \|\vec{e}\| = \|A\hat{x} - \vec{b}\|$

By using the idea that $\vec{e}$ must be orthogonal to the columnspace of A, we solved for $\hat{x}$ and found that

$$\hat{x} = (A^TA)^{-1}A^T\vec{b}$$

Note that this formula works even when $\vec{b}$ does exist in the columnspace of A.

But, this depended on A having full column rank, or linearly independent columns, so that $\text{Null}(A)$ would be trivial, and since $\text{Null}(A) = \text{Null}(A^TA)$, $\text{Null}(A^TA)$ would also be trivial. When a matrix has a trivial nullspace (the only vector in the nullspace is $\vec{0}$), it has linearly independent columns, so A^TA would also have linearly independent columns. We already know A^TA is square, so then A^TA would be invertible.

What happens if A^TA is not invertible?

$A^TA\hat{x} = A^T\vec{b}$

$\hat{x}$ would have infinite solutions with the same $\vec{e} = A\hat{x} - \vec{b}$. There would be no unique solution.

What was the projection, then?

The projection is as close as we can get to the desired vector in a given direction.

Using the least squares formula, we can find an $\hat{x}$ that allows $A\hat{x}$ to get as close to $\vec{b}$ as possible. But what is the vector that is as close to $\vec{b}$ as possible? $\hat{b}$, the projection of $\vec{b}$ onto the columnspace of A.

Thus, the projection of $\vec{b}$ onto the columnspace of A is:

$$\hat{b} = A\hat{x} = A(A^TA)^{-1}A\vec{b}$$

In order for this to exist, again, the columns of A have to be linearly independent.

We also saw this formula in a 2D case, which was just a result of substituting vectors in the above form.

The formula for the projection of one vector $\vec{y}$ onto another vector $\vec{x}$ is:

$$\text{proj}_{\vec{x}}\vec{y} = \frac{\vec{x}^T\vec{y}}{\|\vec{x}\|^2}\vec{x}$$

Examples and Applications

Example 1: $\begin{bmatrix}2\\1\end{bmatrix}[x] = \begin{bmatrix}1\\1\end{bmatrix}$

Find x.

Let's try solving this algebraically.

$2x = 1 \rightarrow x = 0.5$
$1x = 1 \rightarrow x = 1$ } Inconsistent!

What about with Gaussian Elimination?

$$\left[\begin{array}{c|c}2&1\\1&1\end{array}\right] \xrightarrow{R_2 - \frac{1}{2}R_1 \to R_2} \left[\begin{array}{c|c}2&1\\0&\frac{1}{2}\end{array}\right]$$

No solution! Inconsistent equations

This is an overdetermined system, so let's try using the least squares formula.

$$\hat{x} = (A^TA)^{-1}A^T\vec{b}$$
$$= ([2\ 1]\begin{bmatrix}2\\1\end{bmatrix})^{-1}[2\ 1]\begin{bmatrix}1\\1\end{bmatrix}$$
$$= 5^{-1}(3) = \frac{3}{5}$$

$$\hat{b} = A\hat{x}$$
$$= \begin{bmatrix}2\\1\end{bmatrix}\cdot\frac{3}{5} = \begin{bmatrix}6/5\\3/5\end{bmatrix}$$

$$\vec{e} = \vec{b} - \hat{b}$$
$$= \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}6/5\\3/5\end{bmatrix} = \begin{bmatrix}1/5\\-2/5\end{bmatrix}$$

$$\|\vec{e}\| = \sqrt{\frac{1}{25} + \frac{4}{25}} = \sqrt{\frac{1}{5}}$$

If our possible x values were 0.5 and 1 from solving the system, how come our x wasn't the average $\frac{0.5+1}{2} = 0.75$?

If $\hat{x}' = \frac{3}{4}$,

$\hat{b}' = A\hat{x}' = \begin{bmatrix}2\\1\end{bmatrix}\cdot\frac{3}{4} = \begin{bmatrix}6/4\\3/4\end{bmatrix}$

$\vec{e}' = \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}6/4\\3/4\end{bmatrix} = \begin{bmatrix}1/2\\-1/4\end{bmatrix}$

$\|\vec{e}'\| = \sqrt{\frac{1}{4} + \frac{1}{16}} = \sqrt{\frac{5}{16}}$ ← This error is larger!

Example 2: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

Find x_1 and x_2.

This matrix-vector equation tells us that $x_1 = 1$, $x_2 = 2$, and $x_2 = 3$, which is inconsistent!

Let's try Gaussian Elimination:

$$\left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 3 \end{array}\right] \xrightarrow{R_3 - R_2 \to R_3} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{array}\right]$$

No solution! Inconsistent equations.

Let's try using the least squares formula.

$$\hat{x} = (A^TA)^{-1}A^T\vec{b}$$

$$= \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}\right)^{-1}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right)^{-1}\begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

invert diagonal →

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.5 \end{bmatrix}$$

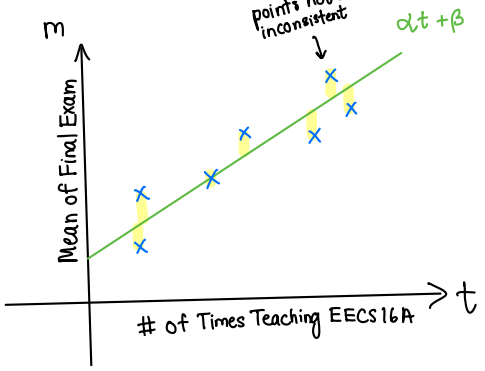
$$\hat{b} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2.5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.5 \\ 2.5 \end{bmatrix}$$

$$\vec{e} = \vec{b} - \hat{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2.5 \\ 2.5 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.5 \\ 0.5 \end{bmatrix}$$

$$\|\vec{e}\| = \sqrt{(-0.5)^2 + (0.5)^2}$$

$$= \sqrt{0.25}$$

Example 3: Linear Regression



represents $\vec{e} = A\hat{x} - \vec{b}$

unknown model parameters to find.

Model: $m = \alpha t + \beta$

Data: Waller: (t_1, m_1)
Sahai: (t_2, m_2)
Alon: (t_3, m_3)
Ranade: (t_4, m_4)
Courtade: (t_5, m_5)
Arias: (t_6, m_6)
Lustig: (t_7, m_7)

t_i: # times taught; m_i: mean of final exam

$$\begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ t_3 & 1 \\ t_4 & 1 \\ t_5 & 1 \\ t_6 & 1 \\ t_7 & 1 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \end{bmatrix} \longrightarrow \hat{x} = (A^TA)^{-1}A^T\vec{b}$$

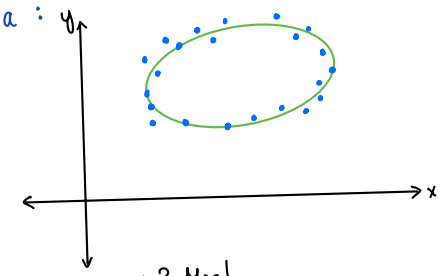
A $\quad \vec{x}$ $\quad \vec{b}$

Example 4: Regression

Gauss found the planet Ceres by using Kepler's laws. Gauss used this technique to predict where certain planets would be in orbit.

Model: $ax^2 + by^2 + cxy + dx + ey = 1$

Data:



22 measured data points
a,b,c,d,e define the best-fit ellipse.
We can now guess the missing measurements!

Is this a linear fit? Yes!

Known values: $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$

Unknown values: $\vec{p} = [a\ b\ c\ d\ e]^T$

$$\begin{bmatrix} x_1^2 & y_1^2 & x_1y_1 & x_1 & y_1 \\ x_2^2 & y_2^2 & x_2y_2 & x_2 & y_2 \\ x_3^2 & y_3^2 & x_3y_3 & x_3 & y_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^2 & y_n^2 & x_ny_n & x_n & y_n \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

22×5 (A) $\quad 5 \times 1$ ($\vec{p}$) $\quad 22 \times 1$ ($\vec{b}$)

22 equations
5 unknowns
"Overdetermined"

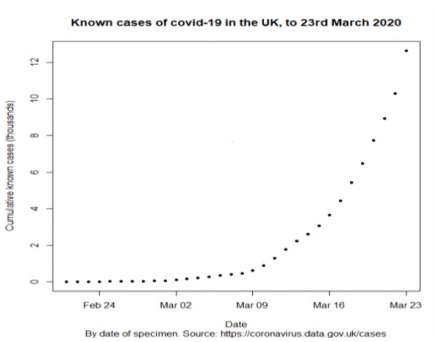
in machine learning, these columns are called features.

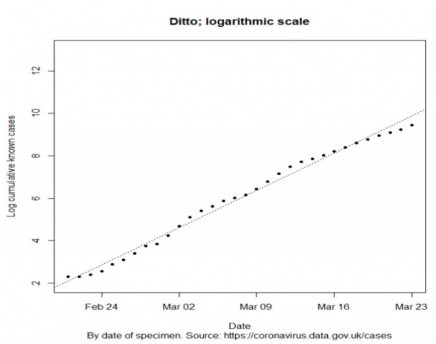
$$\hat{p} = (A^T A)^{-1} A^T \vec{b}$$

We can use a computer to calculate this!

Example 5: Exponential Regression

Model: $y = ce^{ax}$





Is this a linear fit? No, but it can be made linear!

New Model: $\log(y) = \log(c) + \log(e^{ax})$

$= b + ax$

Known values: $(x_1, \log(y_1)), (x_2, \log(y_2)), \dots (x_n, \log(y_n))$

Unknown values: $\vec{p} = [a\ b]^T$

$$\underset{A}{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}} \underset{\vec{p}}{\begin{bmatrix} a \\ b \end{bmatrix}} = \underset{\vec{y}}{\begin{bmatrix} \log y_1 \\ \log y_2 \\ \vdots \\ \log y_n \end{bmatrix}} \longrightarrow \hat{p} = (A^T A)^{-1} A^T \vec{y}$$

$\hat{c} = e^{\hat{b}}$ (if $y = ce^{ax} = e^{ax}e^k$ where $e^k = c$)

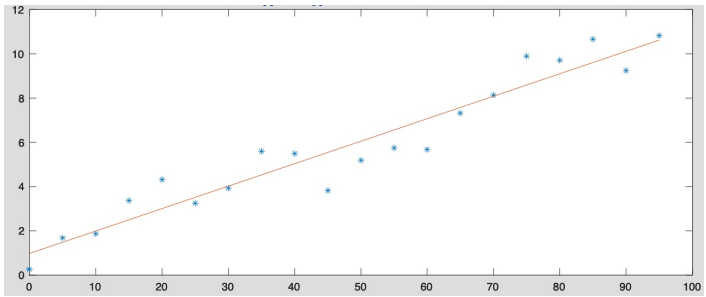
Example 6 : Overfitting

Consider noisy measurements of $y = 0.1x + 1$

Model : $y = ax + b$

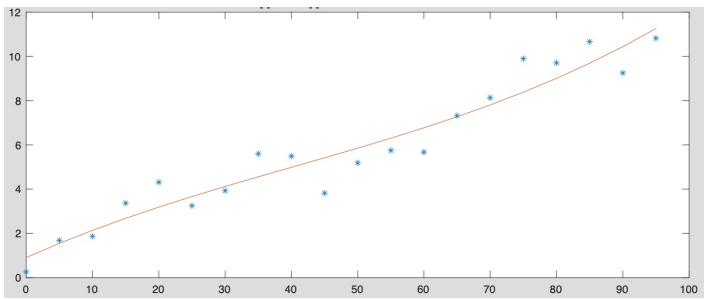
$\hat{p} = [0.1015 \quad 0.9757]^T$

$\|\vec{e}\| = 3.85$



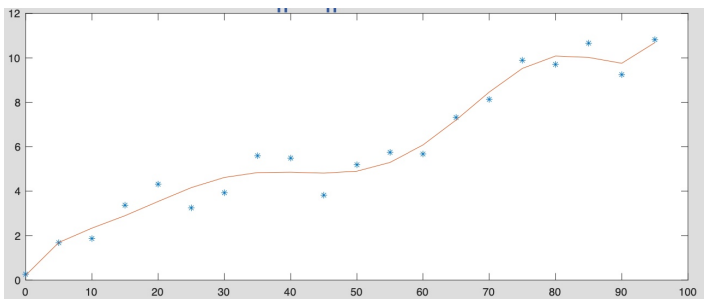
Model : $y = ax^3 + bx^2 + cx + d$

$\|\vec{e}\| = 3.71$



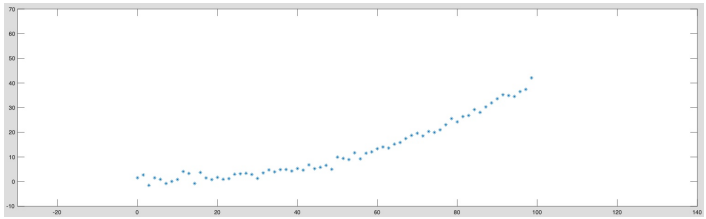
Model : $y = ax^7 + bx^6 + cx^5 + dx^4 + ex^3 + fx^2 + gx + h$

$\|\vec{e}\| = 2.42$



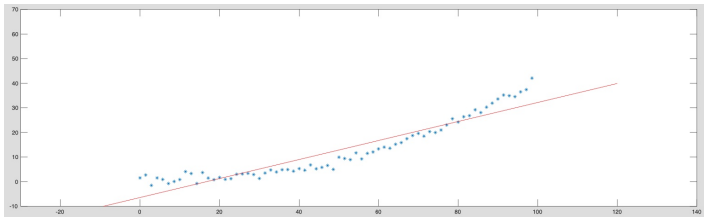
Example 7 : Model Order Selection

We have some data :

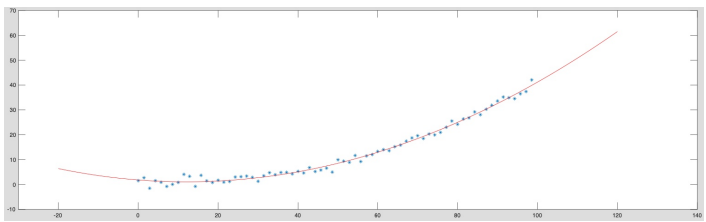


What kind of model do we use?

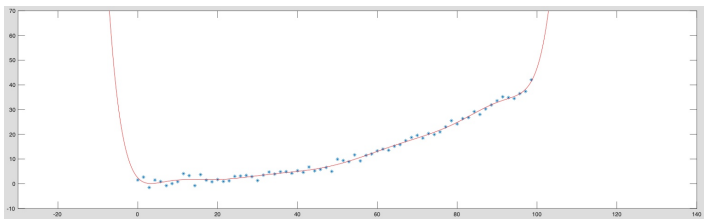
Model : $y = ax + b$



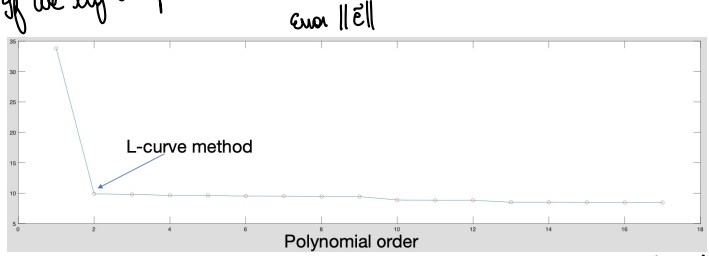
Model : $y = ax^2 + bx + c$



Model : $y = ax^{10} + bx^9 + cx^8 + dx^7 + ex^6 + fx^5 + gx^4 + hx^3 + ix^2 + jx + k$



If we try to pick the model that provides us with the least error :



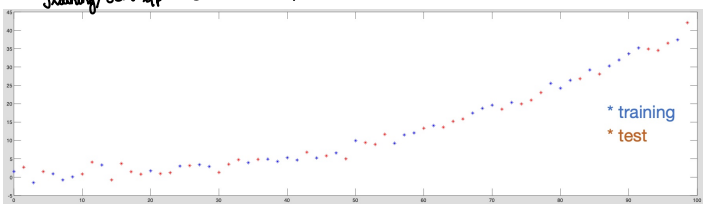
L-curve method : we choose parameters based on the maximal curvature of the L-curve.

To make sure our model can accurately represent new data too, we can use only a portion of our data to fit the best model. Then, we can evaluate the model on the withheld data.

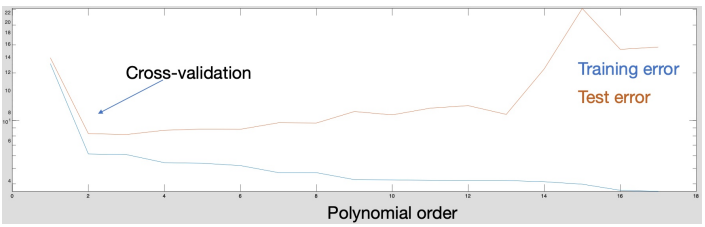
So, we:

1. Split data into training set and test set
2. Fit model on training set
3. Evaluate error on test set.

Training/Test Split (random sample)

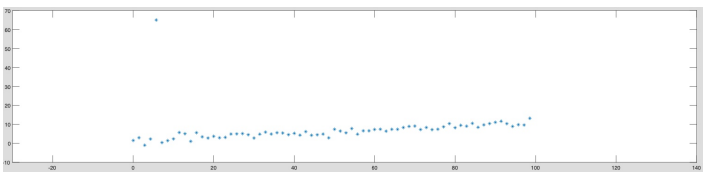


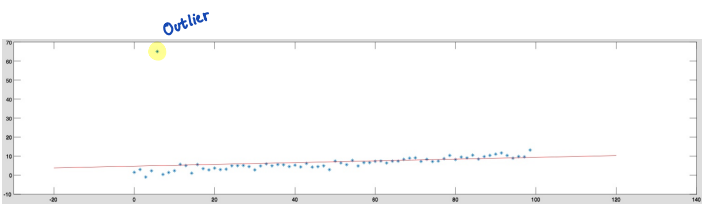
Error ||$\vec{e}$||

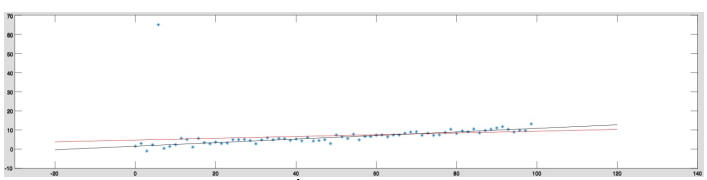


Example 8: Outlier

Model: $Y = ax + b$







One outlier can change the model!

There are different approaches to dealing with outliers based on the situation and problem at hand.

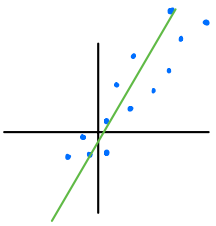
Least Squares Examples

1. Collect a lot of useful data.
   The more data, the better!
2. Choose a model to fit to the data.
   Existing laws or principles or the method of data collection may help inspire this.
3. Use least squares to determine parameters of model.
4. Use model to make predictions!

What kinds of models can we use?

- Linear: $y = mx + b$
- Polynomial: $y = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0$
- Exponential: $y = ce^{ax}$

What are considerations we should make when choosing a model?



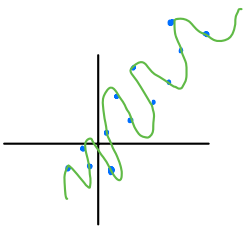
Simple Model:

Pros:

- Easy to fit
- Low variance
  (changing the dataset doesn't affect our parameters as much)

Cons:

- High bias
  (unable to capture the complexity of the data completely)
- High error



Complex Model:

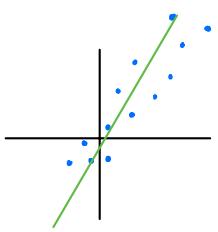
Pros:

- Low error
- Low bias
  (able to capture the complexity of the data)

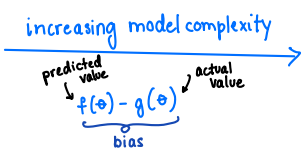
Cons:

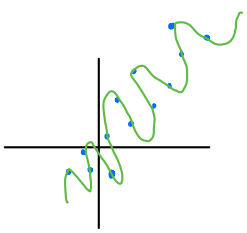
- More challenging to fit
- High variance
  (changing data really affects parameters)
  (model not able to fit data it hasn't seen before)

Bias-Variance Tradeoff



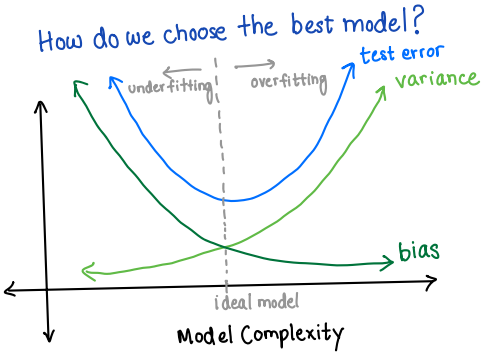
High bias
Low variance
High training error
"Underfitting"





Low bias
High variance
Low training error
"Overfitting"

How do we choose the best model?



Least squares is a technique for supervised machine learning!

Use labeled datasets to train algorithms to classify data or predict outcomes

In this course, you learned how to approach something unfamiliar and systematically build understanding.

We started with foundational linear algebra concepts that we then used to create, analyze, and learn models to solve problems.

Our goal is for you to develop a problem-solving approach that you can apply in your future as an engineer!

We know you may not remember every concept from this class in a few years, but we hope you learned a new way of approaching the world and thinking about design.

After all, this is Designing Information Devices and Systems I!

If you take EECS 16B, the next course in this two course sequence, you will learn:

Module 4: Advanced circuit design/analysis

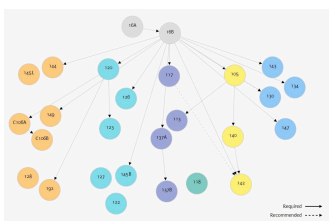
Module 5: Introduction to control and robotics

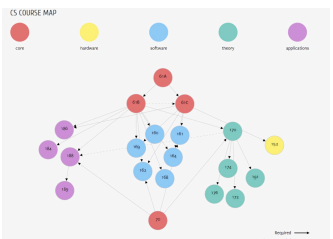
Module 6: Introduction to data analysis and signal processing

There are so many questions we can ask about the world, but also about the topics we learned in this class!

Berkeley offers many more interesting courses to help answer them!

EECS Courses at UC Berkeley





hkn.eecs.berkeley.edu/courseguides

And beyond!

This isn't the end yet!

See you on Monday for a review lecture!