

CS-GY 6923: Lecture 2

Multiple Linear Regression + Feature Transformations + Model Selection

NYU Tandon School of Engineering, Prof. Christopher Musco

- Lab 1 due **Monday, by midnight.**
- Lab 2 will be released today, due in 10 days.
- First written assignment will be released early next, due in 10 days.

10% bonus on the first written assignment if you typeset your solutions in Latex or Markdown. More information on course website.

Training Dataset:

- Given input pairs $(x_1, y_1), \dots, (x_n, y_n)$.
- Each x_i is an input data vector (the predictor).
- Each y_i is an output variable (the target).

Objective:

- Have the computer automatically find some function $f(x)$ such that $f(x_i)$ is close to y_i for the input data.

Standard approach: Convert the supervised learning problem to a multi-variable optimization problem.

SUPERVISED LEARNING DEFINITIONS

What are the three components needed to setup a supervised learning problem?

- **Model** $f_{\theta}(x)$: Class of equations or programs which map input x to predicted output. We want $f_{\theta}(x_i) \approx y_i$ for training inputs.
- **Model Parameters** $\underline{\theta}$: Vector of numbers. These are numerical nobs which parameterize our class of models.
- **Loss Function** $L(\theta)$: Measure of how well a model fits our data. Typically some function of $\underbrace{f_{\theta}(x_1) - y_1, \dots, f_{\theta}(x_n) - y_n}_{\text{ }}$

Empirical Risk Minimization: Choose parameters θ^* which minimize the Loss Function:

$$\theta^* = \arg \min_{\theta} L(\theta)$$

f_{θ^*}

$$\sum_{i=1}^n (f_{\theta}(x_i) - y_i)^2 = L(\theta).$$

Simple Linear Regression

- Model: $f_{\beta_0, \beta_1}(x) = \beta_0 + \beta_1 \cdot x$
- Model Parameters: β_0, β_1
- Loss Function: $L(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - f_{\beta_0, \beta_1}(x_i))^2$

Goal: Choose β_0, β_1 to minimize

$$L(\beta_0, \beta_1) = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|^2.$$

Simple closed form solution: $\beta_1 = \sigma_{xy}/\sigma_x^2$, $\beta_0 = \bar{y} - \beta_1 \bar{x}$. **How did we solve for this solution?**

MULTIPLE LINEAR REGRESSION

Multiple Linear Regression Model:

Predict

$$y_i \approx \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_d x_{id} : f_{\beta_1, \dots, \beta_d}(\bar{x}_i)$$

Data matrix:

$$\text{data example}$$

$$\begin{pmatrix} \bar{X} = & \left[\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ x_{31} & x_{32} & \dots & x_{3d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{array} \right] & = & \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{array} \right] \begin{array}{cccc} x_{12} & \dots & x_{1d} \\ x_{22} & \dots & x_{2d} \\ x_{32} & \dots & x_{3d} \\ \vdots & & \vdots \\ x_{n2} & \dots & x_{nd} \end{array} \right] & \left[\begin{array}{c} y_1 \\ \vdots \\ y_n \end{array} \right] \end{pmatrix}$$

Linear algebraic form:

$$\beta = (\beta_1, \dots, \beta_d)$$

$$\boxed{\quad}$$

$$\begin{matrix} \text{predicts} \\ y_i \sim \langle x_i; \beta \rangle = f_{\beta}(x_i) \\ y \sim X\beta \end{matrix}$$

$$\stackrel{f}{\downarrow} n \times (d+1) \times 1 \rightarrow n \times 1$$

MULTIPLE LINEAR REGRESSION

Linear Least-Squares Regression.

- Model Parameters:

$$\beta = [\beta_1, \beta_2, \dots, \beta_d]^T \parallel \begin{bmatrix} y \\ - \\ \vdots \end{bmatrix} \parallel,$$

- Model:

$$f_{\beta}(x) = \langle x, \beta \rangle$$

- Loss Function:

$$\begin{aligned}\underline{L(\beta)} &= \sum_{i=1}^n |y_i - \langle x_i, \beta \rangle|^2 \\ &= \|y - X\beta\|_2^2\end{aligned}$$

LOSS MINIMIZATION

Goal: minimize the loss function $L(\boldsymbol{\beta}) : \mathbb{R}^d \rightarrow \mathbb{R}$.

Find possible optima by determining for which $\boldsymbol{\beta} = [\beta_1, \dots, \beta_d]$ all partial derivatives equal $\mathbf{0}$. I.e., when do we have:

$$\underline{\nabla L(\boldsymbol{\beta})} = \begin{bmatrix} \frac{\partial L}{\partial \beta_1} \\ \frac{\partial L}{\partial \beta_2} \\ \vdots \\ \frac{\partial L}{\partial \beta_d} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}$$

The list of partial derivatives is called the **gradient** of L at $\boldsymbol{\beta}$, denoted by $\nabla L(\boldsymbol{\beta})$.¹

¹Sanity check: For a model with d parameters, gradient always has length d .

GRADIENT

Claim: The gradient of the multivariate linear regression least squares loss function, $L(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2$, is: $(d \times n)(n \times 1) = d \times 1$

$$\nabla L(\beta) = \underbrace{-2 \cdot \mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta)}$$

$$\nabla L(\beta) : -2 \mathbf{X}^T \mathbf{y} + 2 \mathbf{X}^T \mathbf{X} \beta = 0$$

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \quad \boxed{\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}}$$

Can check that this is equal to 0 only when $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.
There are no other options, so this must be the minimum.

$$\begin{array}{ccc} (d \times n)(n \times d) & & (\mathbf{X}^T \mathbf{X})^{-1} \rightarrow d \times 1 \\ \mathbf{X}^T \mathbf{X} & \mapsto & (d \times d) \\ & & (d \times d)(d \times n)(n \times 1) = d \times 1 \end{array}$$

SINGLE VARIABLE WARMUP

What is the derivative of: $f(x) = x^2$?

$$f'(x) = 2x$$

$$f'(x) = \lim_{\Delta \rightarrow 0} \frac{f(x+\Delta) - f(x)}{\Delta} = \lim_{\Delta \rightarrow 0} \frac{(x+\Delta)^2 - x^2}{\Delta}$$

$$= \lim_{\Delta \rightarrow 0} \frac{x^2 + 2x\Delta + \Delta^2 - x^2}{\Delta} = \lim_{\Delta \rightarrow 0} 2x + \Delta = \boxed{2x}.$$

GRADIENT

Loss function: $L(\beta) = \|y - X\beta\|_2^2$.

$$\nabla L(\beta) = \begin{bmatrix} \frac{\partial L}{\partial \beta_1} \\ \vdots \\ \frac{\partial L}{\partial \beta_d} \end{bmatrix}$$

$$\frac{\partial L}{\partial \beta_i} = \lim_{\Delta \rightarrow 0} \frac{L(\beta + \Delta e_i) - L(\beta)}{\Delta} \quad \xrightarrow{\{000, 1000\}} \Delta X e_i = \Delta x^{(i)}$$

$$= \lim_{\Delta \rightarrow 0} \frac{\|y - X\beta - X\Delta e_i\|_2^2 - \|y - X\beta\|_2^2}{\Delta} = \lim_{\Delta \rightarrow 0} \frac{\|X\beta - X\Delta e_i - X\beta\|_2^2 - \|y - X\beta\|_2^2}{\Delta}$$

$$\|a - b\|_2^2 = (a - b)^T(a - b) = a^T a - 2b^T a + b^T b = \|a\|_2^2 + \|b\|_2^2 - 2b^T a$$

$$= \lim_{\Delta \rightarrow 0} \left(\|y - X\beta\|_2^2 + \|\Delta x^{(i)}\|_2^2 - 2\Delta x^{(i)\top} (y - X\beta) - \|y - X\beta\|_2^2 \right) / \Delta$$

$$\lim_{\Delta \rightarrow 0} \frac{\|\Delta x^{(i)}\|_2^2 - 2\Delta x^{(i)\top} (y - X\beta)}{\Delta} = -2x^{(i)\top} (y - X\beta) + \Delta \|x^{(i)}\|_2^2$$

$$= \boxed{-2x^{(i)\top} (y - X\beta)}$$

$$\Delta^2 \|x^{(i)}\|_2^2$$

GRADIENT

Loss function: $L(\beta) = \|y - X\beta\|_2^2$.

$$\nabla L(\beta) = \begin{bmatrix} \frac{\partial L}{\partial \beta_1} \\ \vdots \\ \frac{\partial L}{\partial \beta_d} \end{bmatrix}$$

$$\nabla L(\beta) = \begin{bmatrix} -2X^{(1)\top}(y - X\beta) \\ -2X^{(2)\top}(y - X\beta) \\ \vdots \\ -2X^{(d)\top}(y - X\beta) \end{bmatrix}$$

$$= -2 \begin{bmatrix} X^{(1)\top} \\ \vdots \\ X^{(d)\top} \end{bmatrix} \begin{bmatrix} | \\ | \\ | \\ | \end{bmatrix} = -2X^\top(y - X\beta)$$

MULTIPLE LINEAR REGRESSION SOLUTION

Take away: simple form for the gradient means that multiple linear regression models are easy and efficient to train.

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Exactly how efficient?

$(\mathbf{X}^T \mathbf{X})$ is a $d \times d$ $(\mathbf{X}^T \mathbf{X})^{-1}$ is $\Theta(d^3)$ time.

$$\frac{\mathbf{X}^T \mathbf{X}}{d \times n} \quad \Theta(d^2 n) \quad \Theta(n d^2)$$

$$\sim \Theta(nd)$$

MULTIPLE LINEAR REGRESSION SOLUTION

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- $\boldsymbol{\beta}^*$ can be computed directly in $O(\underline{\underline{nd^2}})$ time for an $n \times d$ data matrix \mathbf{X} .
- There are iterative approximation methods (fancy versions of gradient descent) that run in roughly $O(nd)$ time. We will use one called LSQR for Lab 2, since d is large.

Take away: simple form for the gradient means that multiple linear regression models are easy and efficient to train.

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- Often the “go to” first regression method. Throw your data in a matrix and see what happens.
- Serve as the basis for much richer classes of models.

ENCODING DATA AS A NUMERICAL MATRIX

It is not always immediately clear how to do this! One of the first issue we run into is categorical data:

$$x_1 = [42, 4, 104, \text{hybrid}, \text{ford}]$$
$$x_2 = [18, 8, 307, \text{gas}, \text{bmw}]$$
$$x_3 = [31, 4, 150, \text{gas}, \text{honda}]$$
$$\vdots$$

ENCODING DATA AS A NUMERICAL MATRIX

Binary data is easy to deal with – pick one category to be 0, one to be 1. The choice doesn't matter – it will not impact the overall loss of the model

$$\mathbf{x}_1 = [42, 4, 104, \underbrace{\text{hybrid}}_{1}, \text{ford}]$$

$$\mathbf{x}_2 = [18, 8, 307, \underbrace{\text{gas}}_{0}, \text{bmw}]$$

$$\mathbf{x}_2 = [31, 4, 150, \underbrace{\text{gas}}_{0}, \text{honda}]$$

⋮

$$\mathbf{x}_1 = [42, 4, 104, 1, \text{ford}]$$

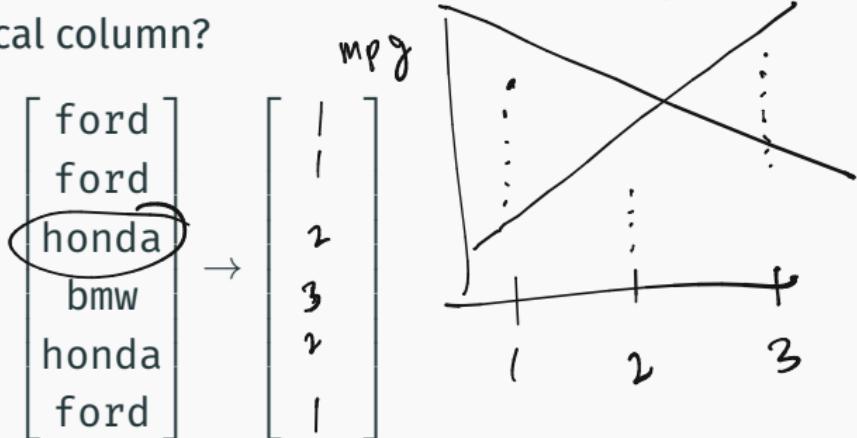
$$\mathbf{x}_2 = [18, 8, 307, 0, \text{bmw}]$$

$$\mathbf{x}_2 = [31, 4, 150, 0, \text{honda}]$$

⋮

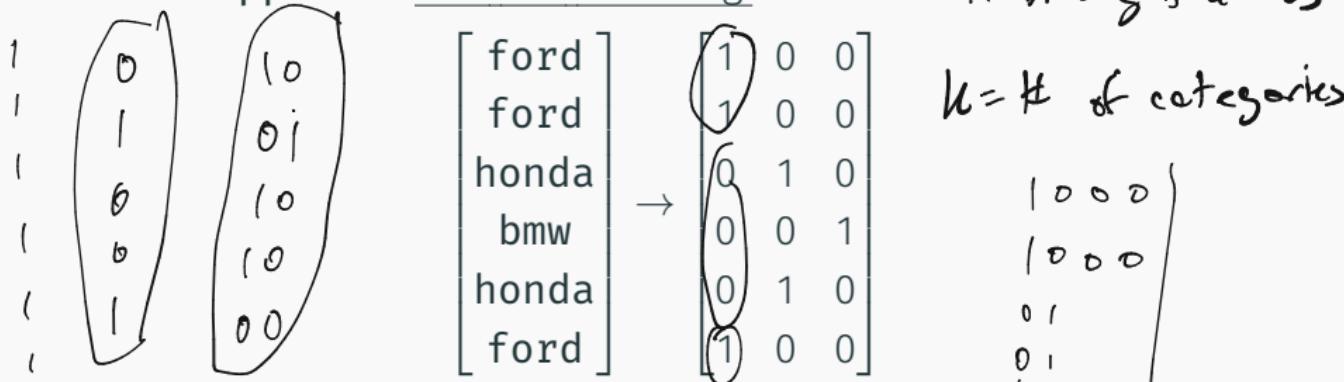
DEALING WITH CATEGORICAL VARIABLES

What about a categorical predictor variable for car make with more than 2 options: e.g. Ford, BMW, Honda. **How would you encode as a numerical column?**



ONE HOT ENCODING

Better approach: One Hot Encoding.



- Create a separate feature for every category, which is 1 when the variable is in that category, zero otherwise.
- Not too hard to do by hand, but you can also use library functions like `sklearn.preprocessing.OneHotEncoder`.

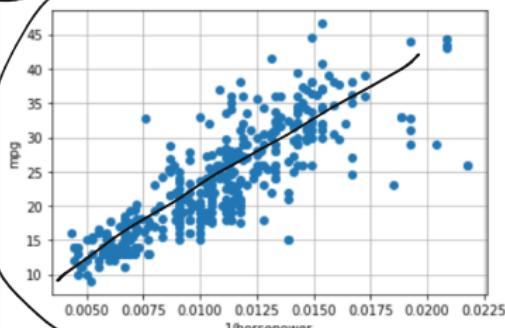
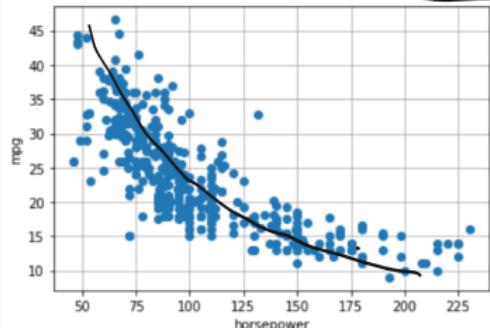
Avoids adding inadvertent linear relationships.

TRANSFORMED LINEAR MODELS

EXAMPLE FROM LAST TIME

$$y_j - \bar{y}(\beta_0 + \beta_1 \cdot x_j)^{1/2}$$

Instead of fitting the model $\text{mpg} \approx \beta_0 + \beta_1 \cdot \text{horsepower}$, fit the model $\text{mpg} \approx \beta_0 + \beta_1 \cdot 1/\text{horsepower}$.



How would you know to make such a transformation?

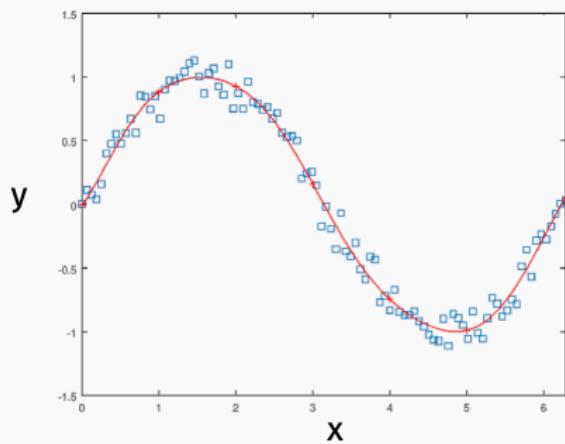
Better approach: Choose a more flexible non-linear model class. What is would be an example of a non-linear curve you could fit?

TRANSFORMED LINEAR MODELS

Suppose we have singular variate data examples (x, y) . We could fit the non-linear polynomial model:

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

$\beta + \Delta e$:



$$\begin{matrix} \beta_1 & 0 \\ \beta_2 & 0 \\ \vdots & + \Delta \\ \beta_3 & 1 \\ \vdots & 0 \\ \beta_0 & 0 \end{matrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \\ \vdots \\ \beta_0 \end{bmatrix}$$



Claim: This can be done using an algorithm for multivariate regression! No need to compute another gradient or write good to optimize β_0, \dots, β_3 .

TRANSFORMED LINEAR MODELS

β_0

Transform into a multiple linear regression problem:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\Rightarrow X\beta = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \underbrace{\begin{bmatrix} \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 \\ \vdots \\ \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \beta_3 x_n^3 \end{bmatrix}}_{y}$$

What is the output of the model $X\beta$ with parameters

$$\beta = [\beta_0, \dots, \beta_3]?$$

$$\beta^* = (X^\top X)^{-1} X^\top y$$

TRANSFORMED LINEAR MODELS

More generally, each column j can be generated by a different basis function $\phi_j(x)$. Could have:

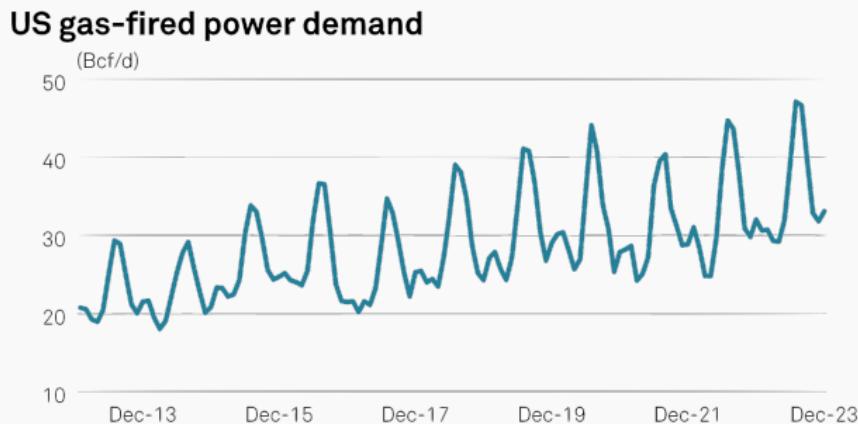
- $\phi_j(x) = x^q$
- $\phi_j(x) = \underline{\sin(x)}$
- $\phi_j(x) = \underline{\cos(10x)}$
- $\phi_j(x) = \underline{1/x}$

When might you want to include sins and cosines?

TRANSFORMED LINEAR MODELS

When might you want to include sines and cosines?

Time series data:



Source: S&P Global Commodity Insights

MULTIVARIATE MODELS

Transformations can also be for multivariate data.

Example: Multivariate polynomial model.

- Given a dataset with target y and predictors x, z .
- For inputs $(x_1, z_1), \dots, (x_n, z_n)$ construct the data matrix:

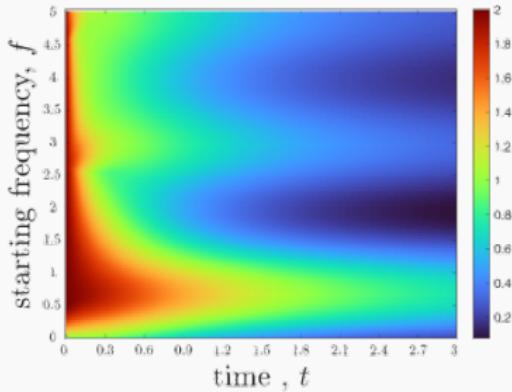
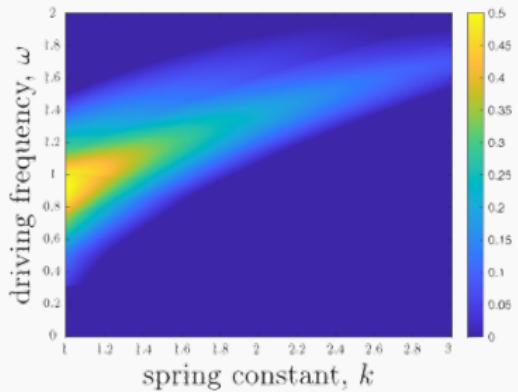
$$\begin{pmatrix} 1 & x_1 & z_1 \\ 1 & \vdots & \vdots \\ 1 & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}$$

$$\begin{bmatrix} 1 & \underline{x_1} & \underline{x_1^2} & \underline{z_1} & \underline{z_1^2} & \textcircled{x_1 z_1} \\ 1 & x_2 & x_2^2 & z_2 & z_2^2 & x_2 z_2 \\ \vdots & \vdots & & \vdots & & \\ 1 & x_n & x_n^2 & z_n & z_n^2 & x_n z_n \end{bmatrix}$$

- Captures non-linear interaction between x and z .

MULTINOMIAL MODEL

We use multivariate polynomials a lot in my work to fit models for physical phenomenon over low-dimensional surfaces:



Return at 3:24.

Feature transformation is an extremely powerful tool that can improve models substantially. However, as will see in the remainder of the lecture, it must be used with care.

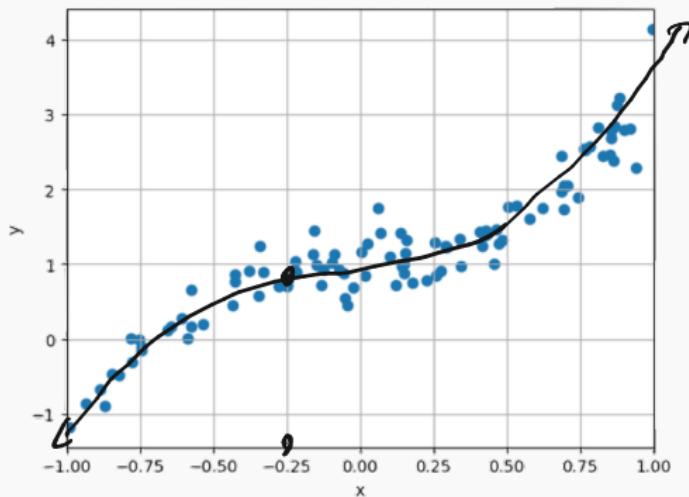
Remainder of lecture: Through a simple example, learn about the **overfitting problem** and how it can be addressed with model selection tools like the **test/train paradigm** and **cross-validation**

We will post a Python demo working through this example.

FITTING A POLYNOMIAL

Simple experiment:

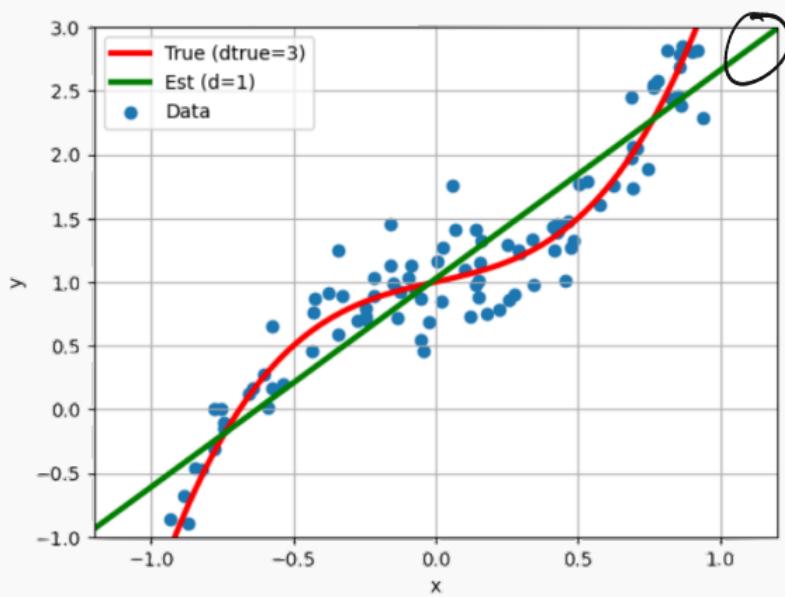
- Randomly select data points $x_1, \dots, x_n \in [-1, 1]$.
- Choose a degree 3 polynomial $p(x)$.
- Create some fake data: $y_i = p(x_i) + \eta$ where η is a random number (e.g., random Gaussian).



FITTING A POLYNOMIAL

Simple experiment:

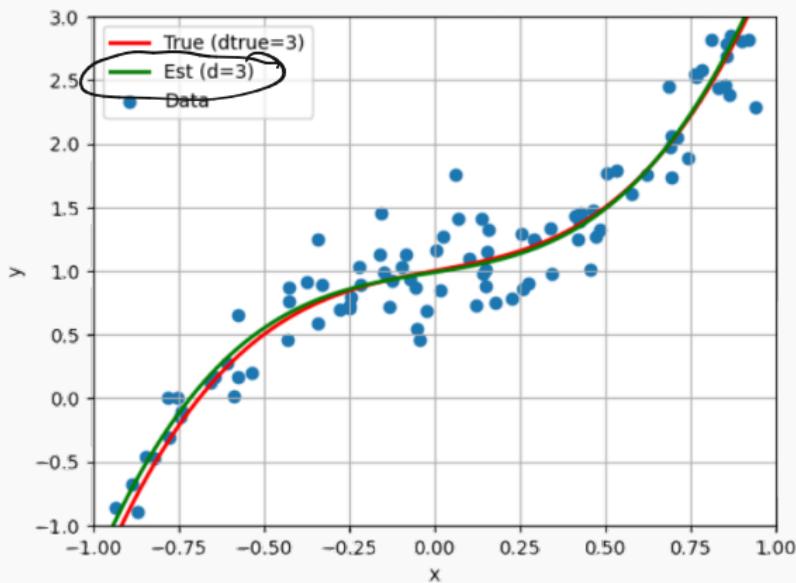
- Use multiple linear regression to fit a line (degree 1 polynomial). This model seems underfit.



FITTING A POLYNOMIAL

Simple experiment:

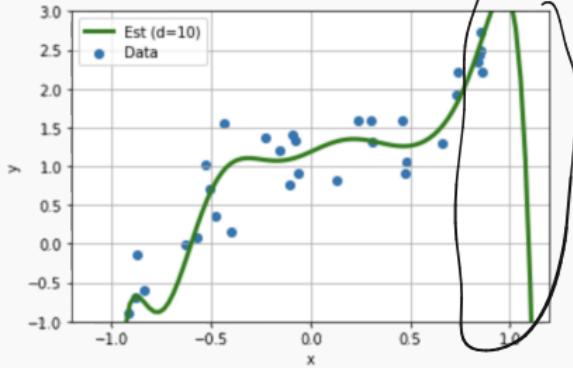
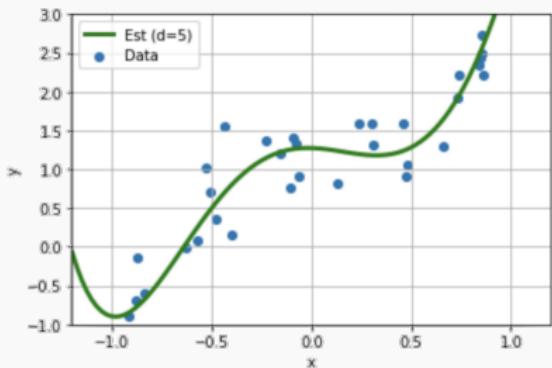
- Use multiple linear regression to fit a degree 3 polynomial.
Almost perfectly captures the true function!



FITTING A POLYNOMIAL

What if we fit a higher degree polynomial?

- Fit degree 5 polynomial under squared loss.
- Fit degree 10 polynomial under squared loss.

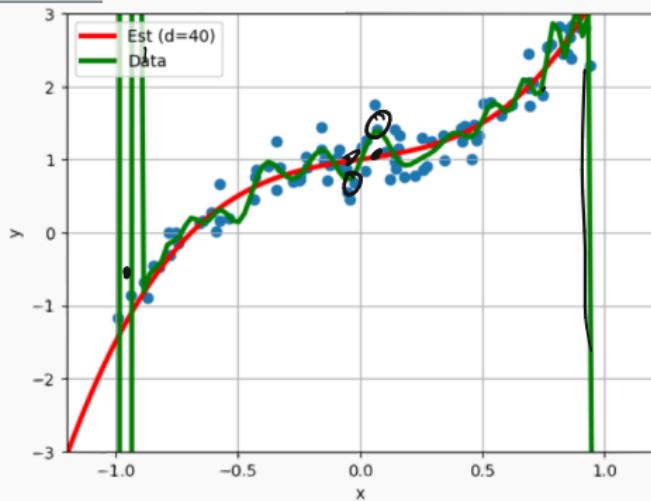


FITTING A POLYNOMIAL

Even higher?

0

- Fit degree 40 polynomial under squared loss. This model seems overfit.



The model “overreacts” to minor variations in the data, which can lead to some bad behavior..

QUICK ASIDE ON NUMERICAL ISSUES

In the demo we have you use

numpy.polynomial.polynomial. However, as we discussed early, we can use multiple linear regression instead by constructing the data matrix:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^1 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$$

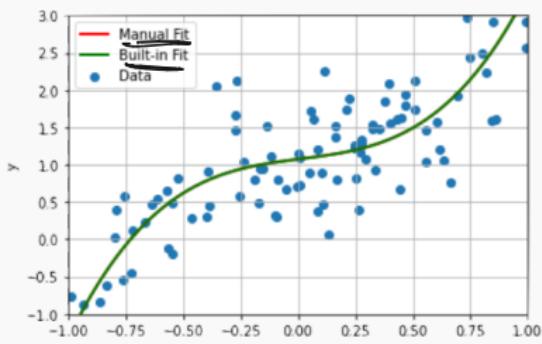
Then find polynomial coefficients as $\beta = (X^T X)^{-1} X^T y$.

QUICK ASIDE ON NUMERICAL ISSUES

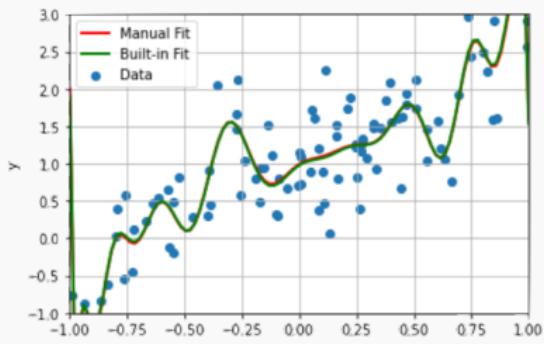
```
# built-in function  
beta_hat = poly.polyfit(xdat,ydat,d)
```

```
# manual fit using naive multivariate regression
```

```
X = np.zeros([len(xdat),d+1])  
for i in range(d+1):  
    X[:,i] = xdat**i  
my_beta = np.linalg.inv(np.transpose(X)@X)@np.transpose(X)@ydat
```



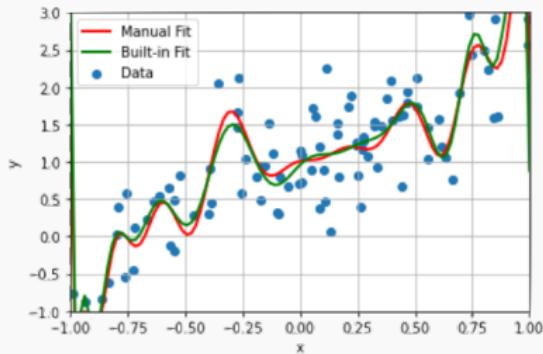
Degree 3



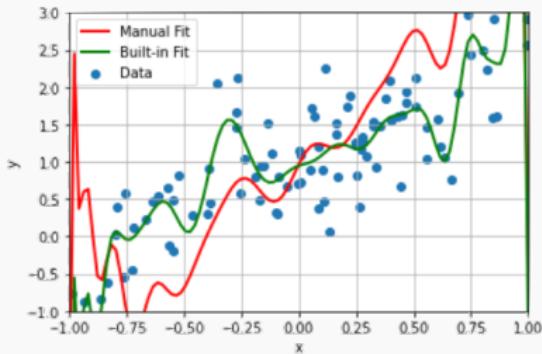
Degree 22

QUICK ASIDE ON NUMERICAL ISSUES

```
# built in function  
beta_hat = poly.polyfit(xdat,ydat,d)  
  
# manual fit using naive multivariate regression|  
X = np.zeros([len(xdat),d+1])  
for i in range(d+1):  
    X[:,i] = xdat**i  
my_beta = np.linalg.inv(np.transpose(X)@X)@np.transpose(X)@ydat
```



Degree 23



Degree 30

Has to due with numerical roundoff error. (Scipy still uses linear regression, but with extra “tricks” to avoid numerical issues)

QUICK ASIDE ON NUMERICAL ISSUES

- Your computer can easily deal with both very large and very small numbers. Underflow and overflow are extremely unlikely to be issues in floating point arithmetic.
- The issue is when you compute using numbers of very differing magnitude.

```
print(.3*10**-34 + 10**-36 - 10**-36)
```

```
3e-35
```

```
print(.3*10**-34 + 10 - 10)
```

```
0.0
```

QUICK ASIDE ON NUMERICAL ISSUES

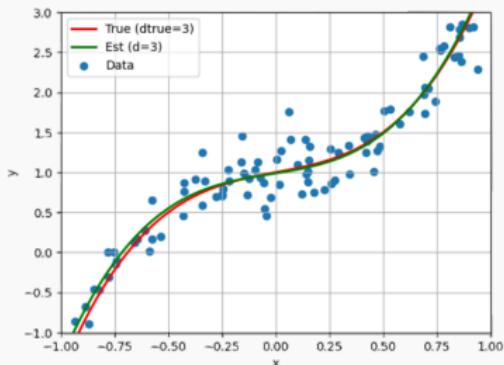
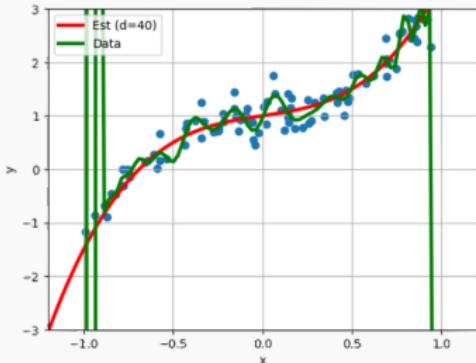
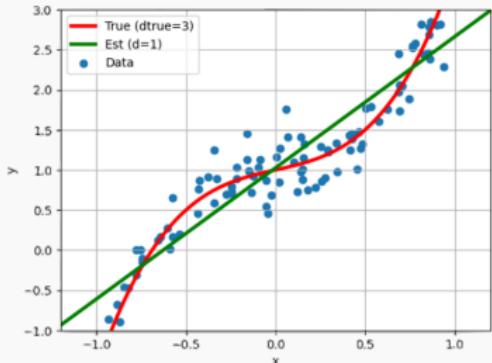
$$x_1 = 1 \quad x_3 = 1$$

Recall that we chose each $x_i \in [-1, 1]$ uniformly at random.

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^1 & x_2^3 \\ 1 & x_3 & x_3^2 & \textcircled{x}_3^3 \\ \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \quad | \quad | \quad | \quad |$$

| .1 .1² .1³1³⁰

BACK TO THE PROBLEM AT HAND



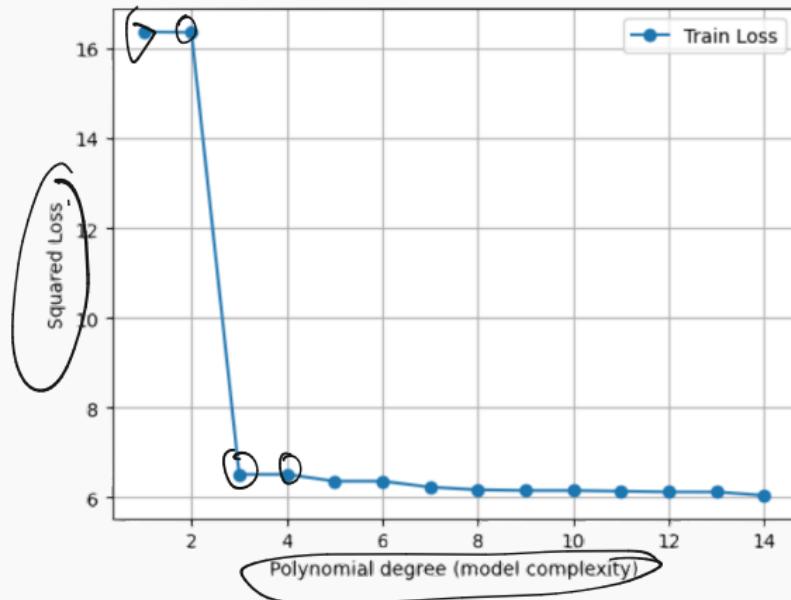
Underfit, overfit, just right.

MODEL ORDER SELECTION

For high-dimensional data, we cannot produce such easy to read plots. How can we automatically detect when we have “underfit” or “overfit” to choose the right model?

MODEL COMPLEXITY VS. LOSS

Typically, the more **complex** our model, the better our loss:



For transformed linear models, this is formally true: more feature transformations leads to lower loss.

MODEL SELECTION

Consider $\underline{X} \in \mathbb{R}^{n \times d}$ and $\bar{X} = [\underline{X} \ z] \in \mathbb{R}^{n \times d+1}$ with one additional column appended on.

Claim:

$$\min_{\beta \in \mathbb{R}^{d+1}} \|\bar{X}\bar{\beta} - y\|_2^2 \leq \min_{\beta \in \mathbb{R}^d} \|X\beta - y\|_2^2$$

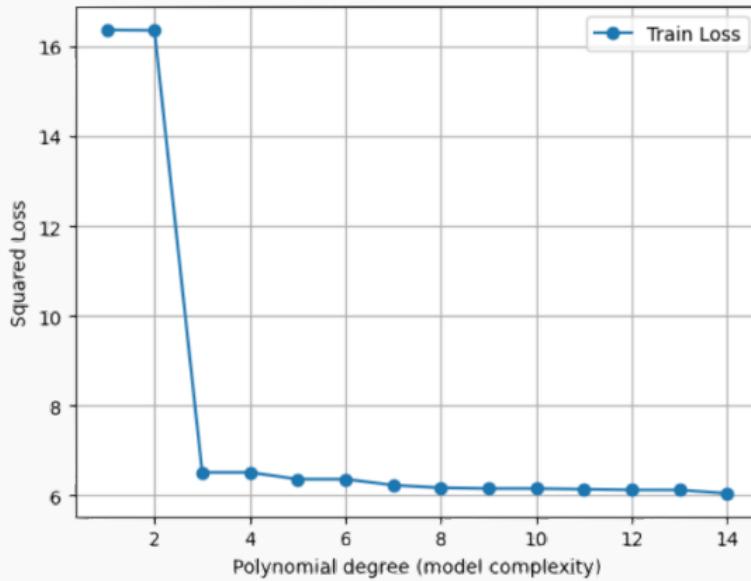
$= \min \|X\beta - y\|_2^2$

$$\beta^* := \arg \min_{\beta} \|X\beta - y\|_2^2$$

$$\hat{\beta}^* = \left[\begin{matrix} \beta^* \\ 0 \end{matrix} \right] \quad \left. \right\}^{d+1}$$

MODEL SELECTION

The more **complex** our model class the better our loss:

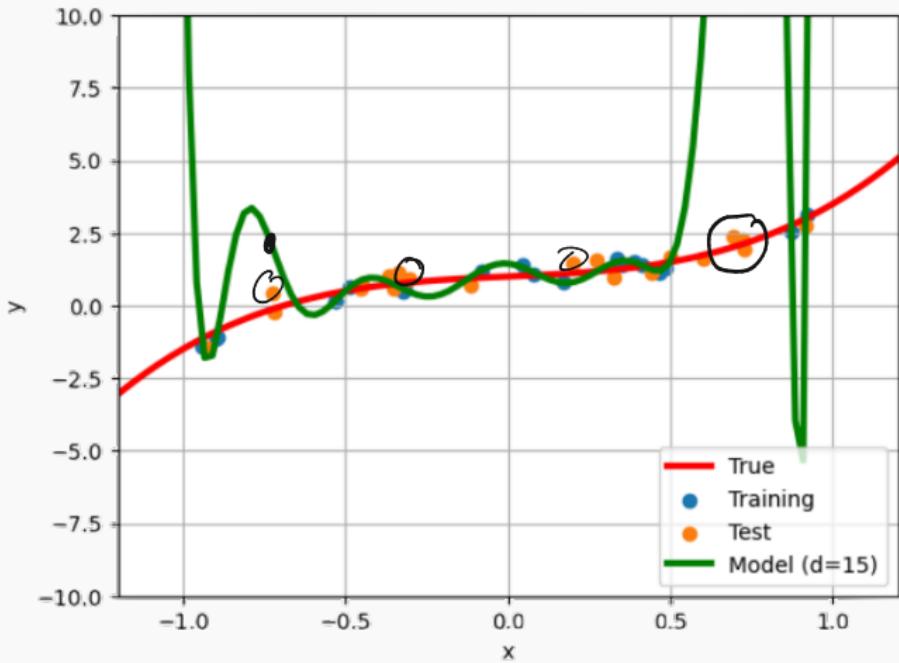


So training loss alone is not usually a good metric for model selection.

MODEL SELECTION

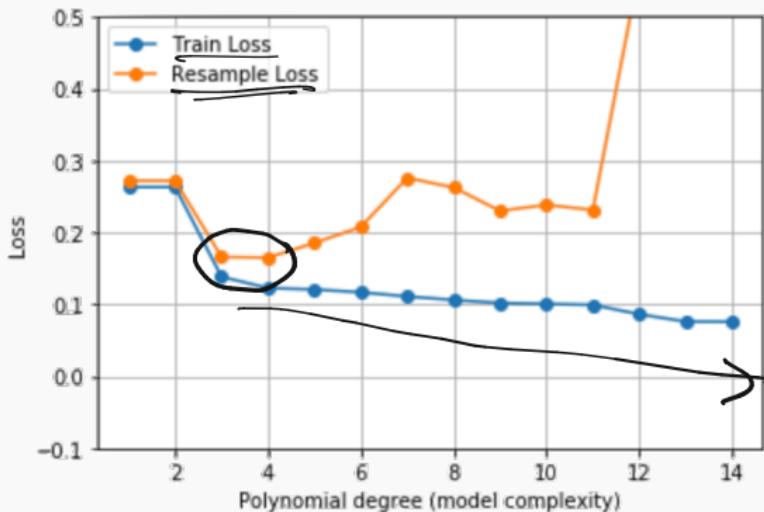
Problem: Small loss does not imply generalization

Generalization: How well do we do on new data?



MODEL SELECTION

Solution: Directly test model on “new data”.



- **Train loss** decreases as model complexity grows.
- **Test loss** “turns around” once our model gets too complex. Minimized around degree 3 – 4.

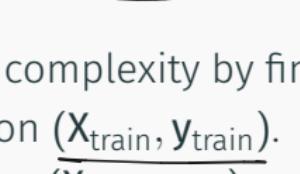
TRAIN-TEST PARADIGM

More reasonable approach: Evaluate model on fresh test data which was not used during training.



Test/train split:

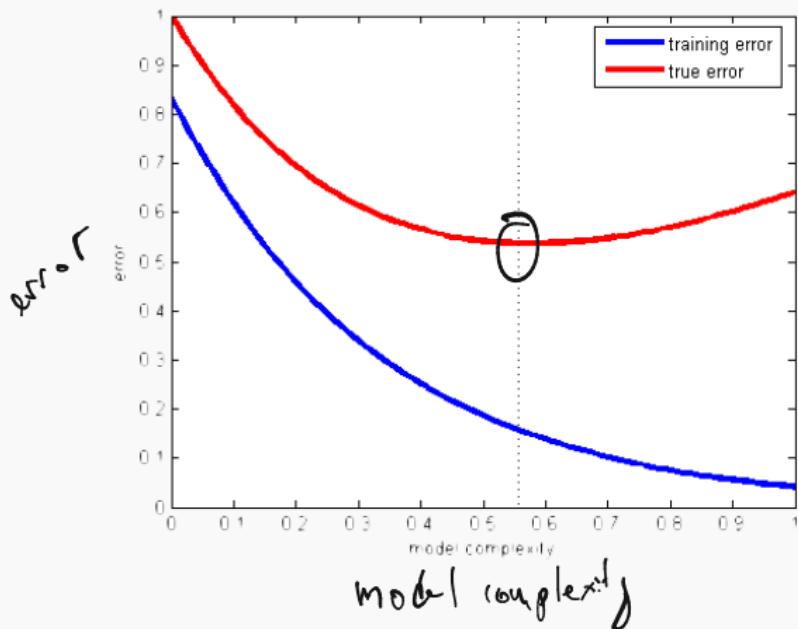
- Given data set (X, y) , split into two sets $(\underline{X_{train}}, \underline{y_{train}})$ and $(\underline{X_{test}}, \underline{y_{test}})$.
- Train q models $\underline{f^{(1)}}, \dots, \underline{f^{(q)}}$ of varying complexity by finding parameters which minimize the loss on $(\underline{X_{train}}, \underline{y_{train}})$.
- Evaluate loss of each trained model on $(\underline{X_{test}}, \underline{y_{test}})$.
- Pick model with lowest test loss.



Sometimes you will see the term **validation set** instead of test set. Sometimes there will be both: use validation set for choosing the model, and test set for getting a final performance measure.

THE FUNDAMENTAL CURVE OF ML

The above trend is fairly representative of what we tend to see across the board:



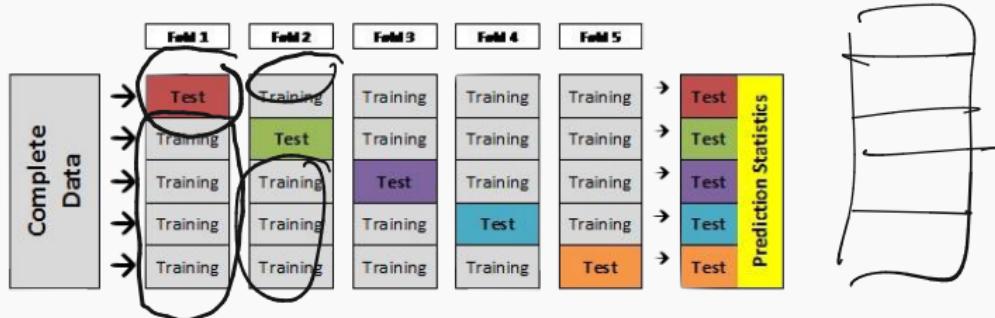
GENERALIZATION ERROR

If the test loss remains low, we say that the model **generalizes**.
Test loss is often called **generalization error**.

TRAIN-TEST PARADIGM

Typical train-test split: 90-70% / 10-30%. Trade-off between between optimization of model parameters and better estimate of model performance.

K-FOLD CROSS VALIDATION



- Randomly divide data in K parts.
 - Typical choice: $K = 5$ or $K = 10$.
- Use $K - 1$ parts for training, 1 for test.
- For each model, compute test loss L_{ts} for each “fold”.
- Choose model with best average loss.
- Retrain best model on entire dataset.

Is there any disadvantage to choosing K larger?

Is “test error” the end goal though? Don’t we care about “future” error?

Intuition: Models which perform better on the test set will **generalize** better to future data.

Goal: Introduce a little bit of formalism to better understand what this means. What is “future” data?

STATISTICAL LEARNING MODEL

Statistical Learning Model:

- Assume each data example is randomly drawn from some distribution $(x, y) \sim \mathcal{D}$.

E.g. x_1, \dots, x_d are Gaussian random variables with parameters



This is not really a simplifying assumption! The distribution could be arbitrarily complicated.

Statistical Learning Model:

- Assume each data example is randomly drawn from some distribution $(x, y) \sim \mathcal{D}$.
- Define the Risk of a model/parameters:

$$\underline{R(f, \theta)} = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\underline{L(f(x, \theta), y)}]$$

here L is our loss function (e.g. $L(z, y) = |z - y|$ or $\underline{L(z, y) = (z - y)^2}$).

Ultimate Goal: Find model $f \in \{f^{(1)}, \dots, f^{(q)}\}$ and parameter vector $\underline{\theta}$ to minimize the $R(f, \theta)$.

$$\mathcal{R}(f, \theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n L(f(x_i, \theta), y_i)$$

- (Population) Risk:

$$\underline{\mathcal{R}(f, \theta)} = \mathbb{E}_{(x, y) \sim \mathcal{D}} [L(f(x, \theta), y)]$$

- Empirical Risk: Draw $(x_1, y_1), \dots, (x_n, y_n) \sim \mathcal{D}$

$$\underline{\mathcal{R}_E(f, \theta)} = \frac{1}{n} \sum_{i=1}^n \underline{L(f(x_i, \theta), y_i)}$$

$$\begin{aligned} \mathbb{E}(\mathcal{R}_E(f, \theta)) &= \mathbb{E}[\dots] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[L(f(x_i, \theta), y_i)] \\ &= \frac{1}{n} \sum_{i=1}^n \mathcal{R}(f, \theta) = \mathcal{R}(f, \theta). \end{aligned}$$

For any fixed model f and parameters θ ,

$$\underbrace{\mathbb{E}[R_E(f, \theta)]}_{\text{---}} = R(f, \theta).$$

Only true if f and θ are chosen *without looking at the data used to compute the empirical risk*.

MODEL SELECTION

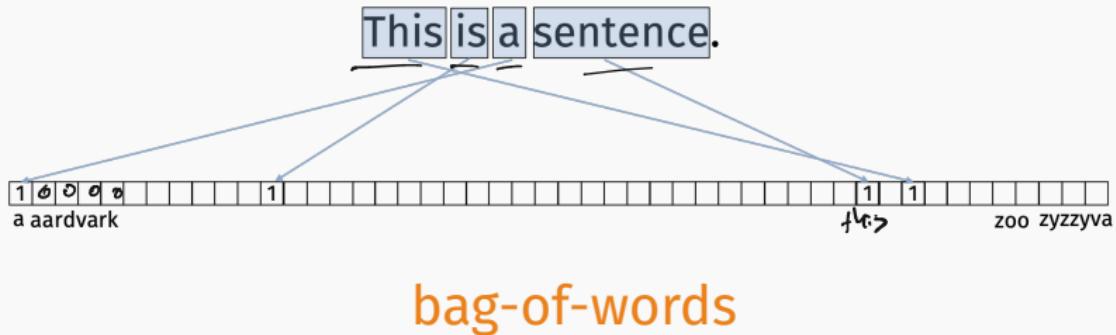
- Train q models $(f^{(1)}, \theta_1^*), \dots, (f^{(q)}, \theta_q^*)$.
- For each model, compute empirical risk $R_E(f^{(i)}, \theta_i^*)$ using test data.
- Since we assume our original dataset was drawn independently from \mathcal{D} , so is the random test subset.

No matter how our models were trained or how complex they are, $R_E(f^{(i)}, \theta_i^*)$ is an unbiased estimate of the true risk $R(f^{(i)}, \theta_i^*)$ for every i . Can use it to distinguish between models.

MODEL SELECTION EXAMPLE

bag-of-words models and n-grams

Common way to represent documents (emails, webpages, books) as numerical data. The ultimate example of 1-hot encoding.



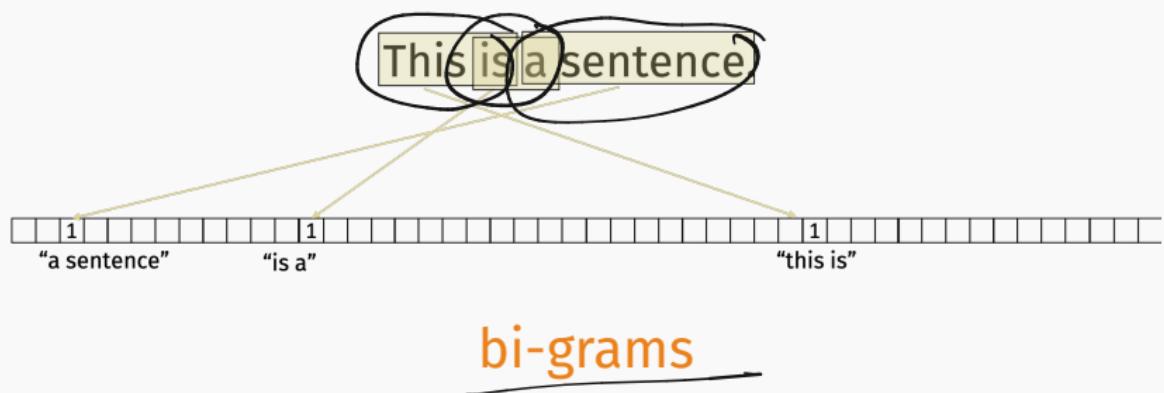
MODEL SELECTION EXAMPLE

bag-of-words models and n-grams

$m = \# \text{ of words}$

Common way to represent documents (emails, webpages, books) as numerical data. The ultimate example of 1-hot encoding.

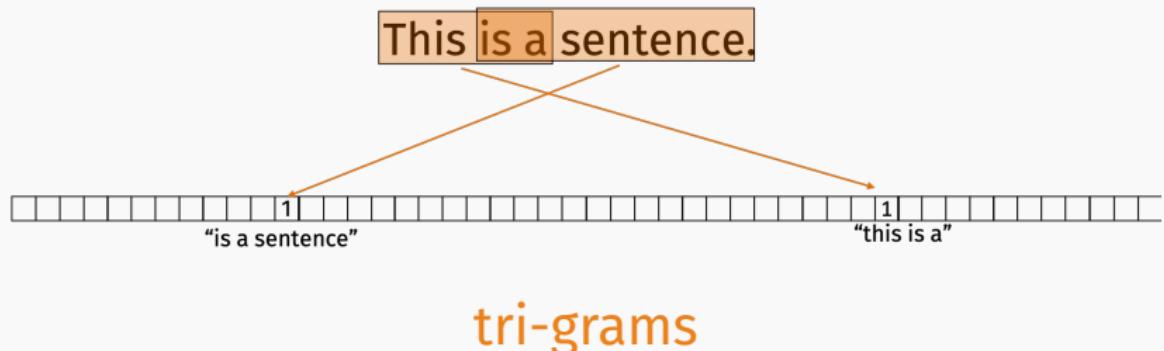
m^2



MODEL SELECTION EXAMPLE

bag-of-words models and n-grams

Common way to represent documents (emails, webpages, books) as numerical data. The ultimate example of 1-hot encoding.



MODEL SELECTION EXAMPLE

Models of increasing order:

- Model $f_{\theta_1}^{(1)}$: spam filter that looks at **single words**.
- Model $f_{\theta_2}^{(2)}$: spam filter that looks at **bi-grams**.
- Model $f_{\theta_3}^{(3)}$: spam filter that looks at **tri-grams**.
- ...

“interest”

“low interest”

“low interest loan”

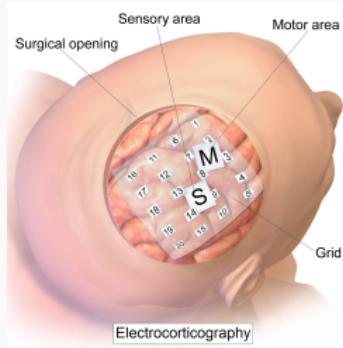
Increased length of **n-gram** means more expressive power.

Will be very relevant in our lab on generative language models!

MODEL SELECTION EXAMPLE

Electrocorticography ECoG (next lab):

- Implant grid of electrodes on surface of the brain to measure electrical activity in different regions.



- Predict hand motion based on ECoG measurements.
- Model order:** predict movement at time t using brain signals at time $t, t - 1, \dots, t - q$ for varying values of q .

ELECTROCORTICOGRAPHY

Our lab uses data collected from monkeys. Precursor to technologies like Braingate, Neuralink, etc.

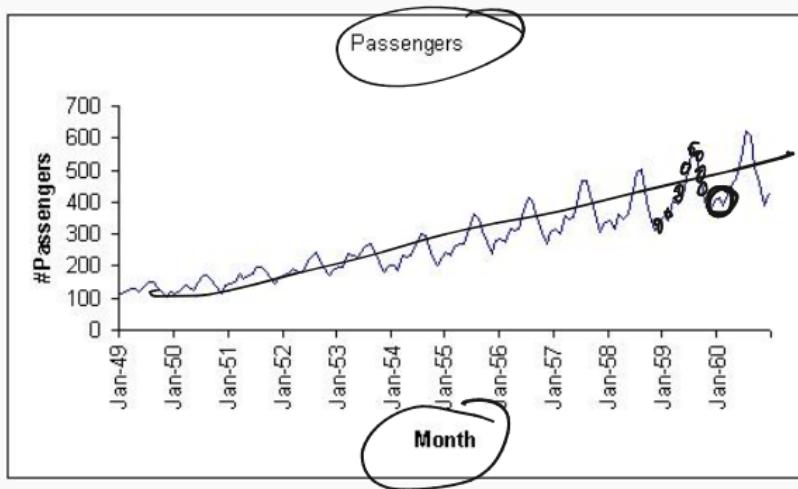


Paralyzed or impaired person could control computer cursor, robotic arm, etc. simply by thinking about it.

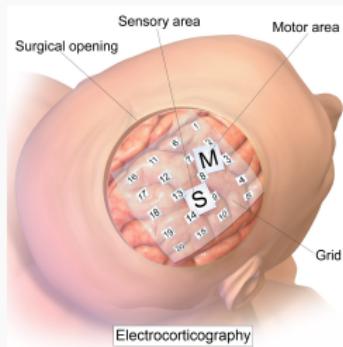
Small implant reads brainwaves and recognizes their intent.

AUTOREGRESSIVE MODEL

Predicting time t based on a linear function of the signals at time $t, t-1, \dots, t-q$ is not the same as fitting a line to the time series. It's much more expressive.



Electrocorticography ECoG lab:



First lab where computation actually matters (solving regression problems with $\sim 40k$ examples, ~ 1500 features)

Makes sense to test and debug code using a subset of the data.

Slight caveat: The train-test paradigm is typically not how machine learning or scientific discovery works in practice!

Typical workflow:

- Train a class of models.
- Test.
- Adjust class of models.
- Test.
- Adjust class of models.
- Cont...

Final model implicitly depends on test set because performance on the test set guided how we changed our model.

Popularity of ML benchmarks and competitions leads to adaptivity at a massive scale.

11 Active Competitions

	#DFDC Deepfake Detection Challenge Identify videos with facial or voice manipulations <small>Featured · Code Competition · 2 months to go · video data, online video</small>	\$1,000,000 1,595 teams
	Google QUEST Q&A Labeling Improving automated understanding of complex question answer content <small>Featured · Code Competition · 19 hours to go · text data, nlp</small>	\$25,000 1,559 teams
	Real or Not? NLP with Disaster Tweets Predict which Tweets are about real disasters and which ones are not <small>Getting Started · Ongoing · text data, binary classification</small>	\$10,000 2,657 teams
	Bengali.AI Handwritten Grapheme Classification Classify the components of handwritten Bengali <small>Research · Code Competition · a month to go · multiclass classification, image data</small>	\$10,000 1,194 teams

Kaggle (various competitions)



14,197,122 images, 21841 synsets indexed

[Explore](#) [Download](#) [Challenges](#) [Publications](#) [Updates](#) [About](#)

Not logged in. [Login](#) | [Signup](#)

Imagenet (image classification and categorization)

Is adaptivity a problem? Does it lead to over-fitting? How much? How can we prevent it? All current research. Related to the problem of “p-value hacking” in science.

REPORT

The reusable holdout: Preserving validity in adaptive data analysis

Cynthia Dwork^{1,*}, Vitaly Feldman^{2,*}, Moritz Hardt^{3,*}, Toniann Pitassi^{4,*}, Omer Reingold^{5,*}, Aaron Roth^{6,*}

* See all authors and affiliations

Science 07 Aug 2015:
Vol. 349, Issue 6248, pp. 636-638
DOI: 10.1126/science.aaa9375

Do ImageNet Classifiers Generalize to ImageNet?			
Benjamin Recht*	Rebecca Roelofs	Ludwig Schmidt	Vaishaal Shankar
UC Berkeley	UC Berkeley	UC Berkeley	UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models’ inability to generalize to slightly “harder” images than those found in the original test sets.

IMAGENET DATASET

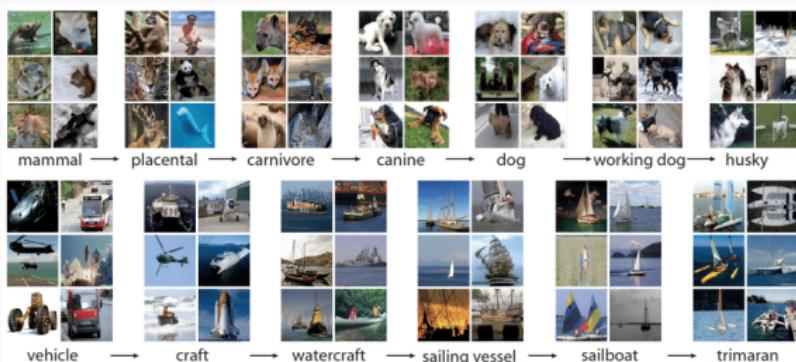


14,197,122 images, 21841 synsets indexed

[Explore](#) [Download](#) [Challenges](#) [Publications](#) [Updates](#) [About](#)

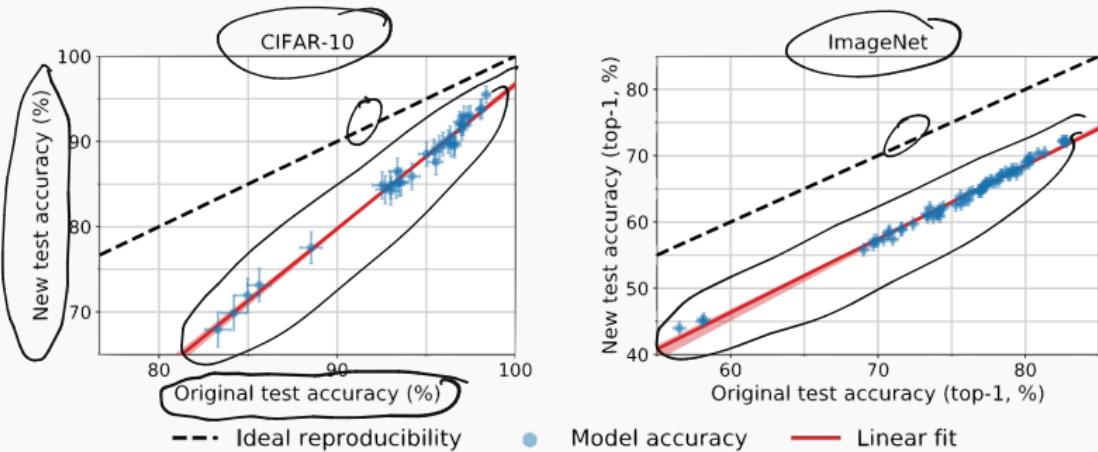
Not logged in. [Login](#) | [Signup](#)

Collected by Fei-Fei Li's group at Stanford in 2006ish and labeled using Amazon Mechanical Turk.



We now have neural network models that can solve these classification problems with > 95% accuracy.

Do ImageNet Classifiers Generalized to ImageNet?



Interestingly, when comparing popular vision models on “fresh” data, while performance dropped across the board, the relative rank of model performance did not change significantly.