CS-GY 6923: Lecture 2
Multiple Linear Regression + Feature Transformations + Model Selection

NYU Tandon School of Engineering, Prof. Christopher Musco

COURSE ADMIN

- First lab assignment `lab1.ipynb` due **Monday, by midnight.**
- First written assignment will be released this weekend.
- TA's will start office hours next week – thanks for everyone who filled out the poll.

REMINDER: SUPERVISED LEARNING

Training Dataset:

- Given input pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.
- Each $\mathbf{x}_i$ is an input data vector (the predictor).
- Each y_i is a continuous output variable (the target).

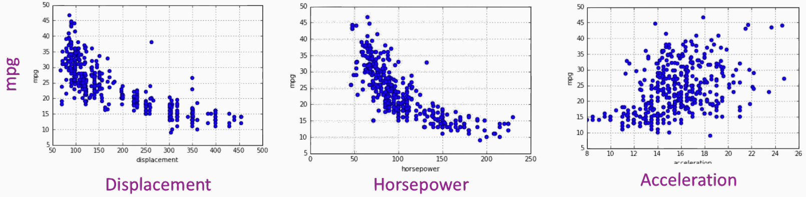
Objective:

- Have the computer automatically find some function $f(\mathbf{x})$ such that $f(\mathbf{x}_i)$ is close to y_i for the input data.

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

EXAMPLE FROM LAST CLASS

Predict miles per gallon of a vehicle given information about its engine/make/age/etc.



SUPERVISED LEARNING DEFINITIONS

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

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

Empirical Risk Minimization: Choose parameters $\boldsymbol{\theta}^*$ which minimize the Loss Function:

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

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 Model:

Predict $$y_i \approx \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{id}$$

Data matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ x_{21} & x_{22} & \ldots & x_{2d} \\ x_{31} & x_{32} & \ldots & x_{3d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nd} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \ldots & x_{1d} \\ 1 & x_{22} & \ldots & x_{2d} \\ 1 & x_{32} & \ldots & x_{3d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \ldots & x_{nd} \end{bmatrix}$$

Linear algebraic form:

$$y_i \sim \langle \mathbf{x}, \boldsymbol{\beta} \rangle$$
$$\mathbf{y} \sim \mathbf{X}\boldsymbol{\beta}$$

MULTIPLE LINEAR REGRESSION

Linear <u>Least-Squares</u> Regression.

- Model Parameters:

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_d]$$

- Model:

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

- Loss Function:

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

LINEAR ALGEBRAIC FORM OF LOSS FUNCTION

LOSS MINIMIZATION

Machine learning goal: minimize the loss function $L(\boldsymbol{\beta}) : \mathbb{R}^d \to \mathbb{R}$.

Find possible optima by determining for which $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_d]$ all the **gradient** equals **0**. I.e. when do we have:

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

GRADIENT

Loss function:

$$L(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Gradient:

$$-2 \cdot \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

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

SINGLE VARIABLE WARMUP

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

GRADIENT

Loss function:

$$L(\boldsymbol{\beta}) = \|\mathsf{y} - \mathsf{X}\boldsymbol{\beta}\|_2^2$$

MULTIPLE LINEAR REGRESSION SOLUTION

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

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

- 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:

$$\begin{aligned} x_1 &= [42, 4, 104, \texttt{hybrid}, \texttt{ford}] \\ x_2 &= [18, 8, 307, \texttt{gas}, \texttt{bmw}] \\ x_2 &= [31, 4, 150, \texttt{gas}, \texttt{honda}] \\ &\vdots \end{aligned}$$

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

$$\begin{aligned} x_1 &= [42, 4, 104, \texttt{hybrid}, \texttt{ford}] \\ x_2 &= [18, 8, 307, \texttt{gas}, \texttt{bmw}] \\ x_2 &= [31, 4, 150, \texttt{gas}, \texttt{honda}] \\ &\vdots \end{aligned}$$

$$\begin{aligned} x_1 &= [42, 4, 104, \texttt{1}, \texttt{ford}] \\ x_2 &= [18, 8, 307, \texttt{0}, \texttt{bmw}] \\ x_2 &= [31, 4, 150, \texttt{0}, \texttt{honda}] \\ &\vdots \end{aligned}$$

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?**

$$\begin{bmatrix} \texttt{ford} \\ \texttt{ford} \\ \texttt{honda} \\ \texttt{bmw} \\ \texttt{honda} \\ \texttt{ford} \end{bmatrix} \rightarrow \begin{bmatrix} \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

ONE HOT ENCODING

Better approach: One Hot Encoding.

$$\begin{bmatrix} \texttt{ford} \\ \texttt{ford} \\ \texttt{honda} \\ \texttt{bmw} \\ \texttt{honda} \\ \texttt{ford} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

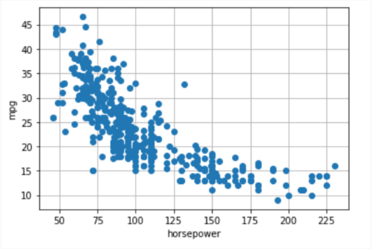
- Create a separate feature for every category, which is 1 when the variable is in that category, zero otherwise.
- Not too hard to do by hand, but you can also use library functions like `sklearn.preprocessing.OneHotEncoder`.

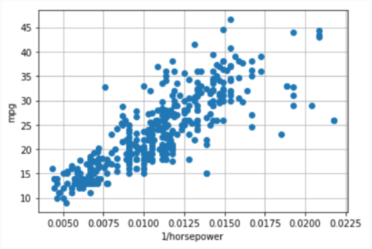
Avoids adding inadvertent linear relationships.

TRANSFORMED LINEAR MODELS

EXAMPLE FROM LAST TIME

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





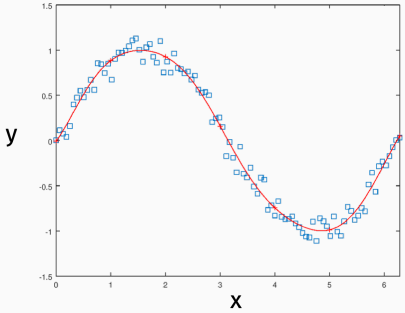
How would you know to make such a transformation?

Better approach: Choose a more flexible non-linear model class.

TRANSFORMED LINEAR MODELS

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

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



Claim: This can be done using an algorithm for multivariate regression! No need to compute another gradient or write good to optimize $\beta_0, \ldots, \beta_3$.

TRANSFORMED LINEAR MODELS

Transform into a multiple linear regression problem:

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

What is the output of $\mathbf{X}\beta$?

TRANSFORMED LINEAR MODELS

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

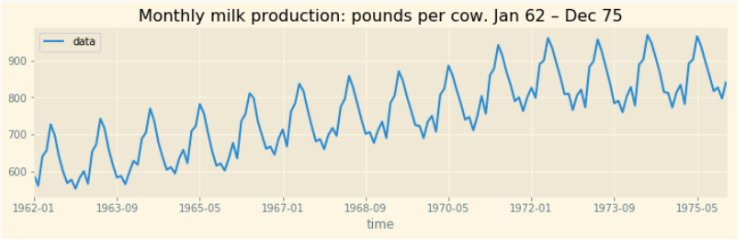
- $\phi_j(x) = x^q$
- $\phi_j(x) = sin(x)$
- $\phi_j(x) = cos(10x)$
- $\phi_j(x) = 1/x$

When might you want to include sins and cosines?

TRANSFORMED LINEAR MODELS

When might you want to include sins and cosines?

Time series data:



There is usually not much harm in including irrelevant variable transformation.

Multinomial model

Transformations can also be for multivariate data.

Example: Multinomial model.

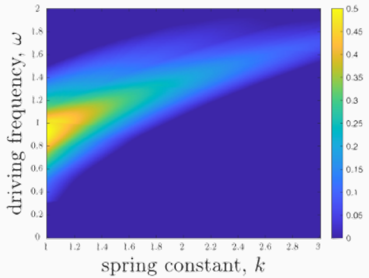
- Given a dataset with target y and predictors x, z.
- For inputs $(x_1, z_1), \ldots, (x_n, z_n)$ construct the data matrix:

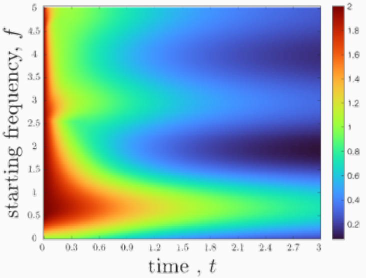
$$\begin{bmatrix} 1 & x_1 & x_1^2 & z_1 & z_1^2 & x_1 z_1 \\ 1 & x_2 & x_2^2 & z_2 & z_2^2 & x_2 z_2 \\ \vdots & \vdots & & \vdots & & \\ 1 & x_n & x_n^2 & z_n & z_n^2 & x_n z_n \end{bmatrix}$$

- Captures non-linear interaction between x and z.

MULTINOMIAL MODEL

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





MODEL SELECTION

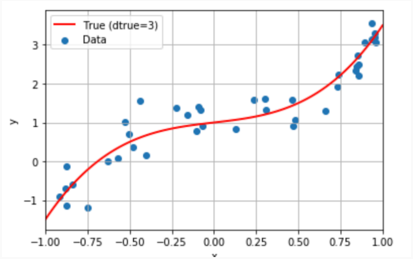
Remainder of lecture: Learn about model selection, test/train paradigm, and cross-validation through a simple example.

I have a Python demo working through this example.

FITTING A POLYNOMIAL

Simple experiment:

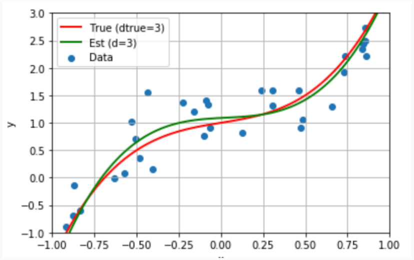
- Randomly select data points $x_1, \ldots, x_n \in [-1, 1]$.
- Choose a degree 3 polynomial $p(x)$.
- Create some fake data: $y_i = p(x_i) + \eta$ where η is a random number (e.g random Gaussian).



FITTING A POLYNOMIAL

Simple experiment:

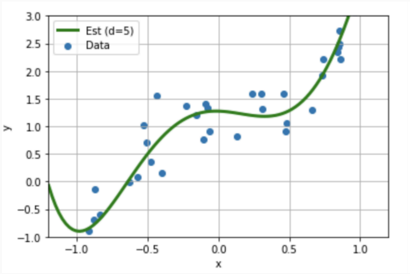
- Use multiple linear regression to fit a degree 3 polynomial.

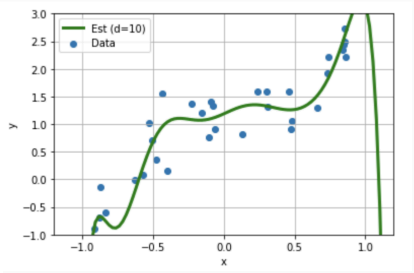


FITTING A POLYNOMIAL

What if we fit a higher degree polynomial?

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

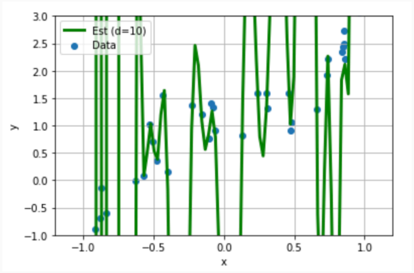




FITTING A POLYNOMIAL

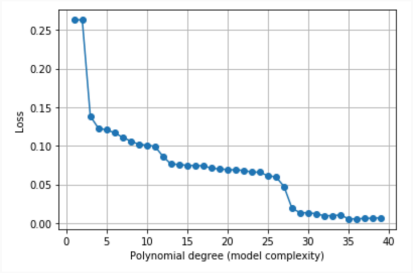
Even higher?

- Fit degree 40 polynomial under squared loss.



MODEL SELECTION

The more **complex** our model class (i.e. the higher degree we allow) the better our loss:



MODEL SELECTION

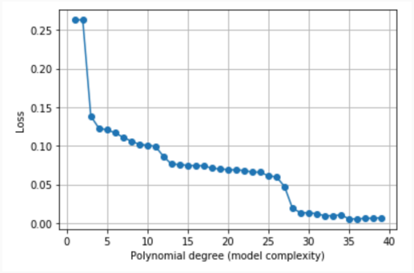
Consider $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\bar{\mathbf{X}} = [\mathbf{X}, \mathbf{z}] \in \mathbb{R}^{n \times d+1}$ with one additional column appended on.

Claim:

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

MODEL SELECTION

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



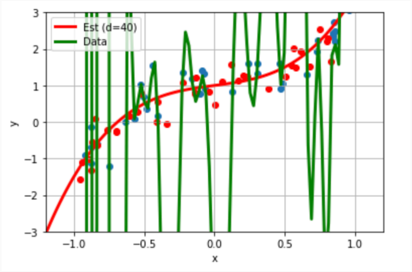
So training loss alone is not usually a good metric for model selection. Small loss does not imply generalization.

Generalization: How well do we do on new data.

MODEL SELECTION

Problem: Loss alone is not informative for choosing model.

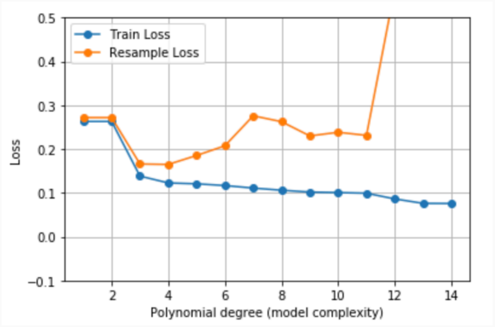
For more complex models, we get smaller loss on the training data, but don't expect to perform well on "new" data:



In other words, the model does not generalize.

MODEL SELECTION

Solution: Directly test model on “new data”.



- Loss continues to decrease as model complexity grows.
- Performance on new data “turns around” once our model gets too complex. Minimized around degree 4.

TRAIN-TEST PARADIGM

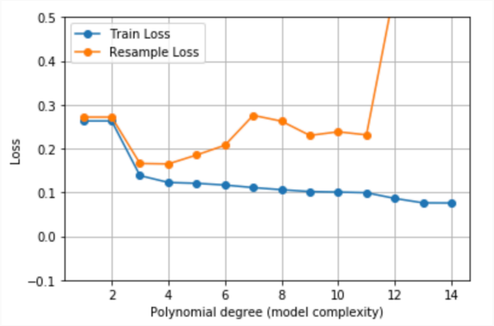
More reasonable approach: Evaluate model on fresh test data which was not used during training.

Test/train split:

- Given data set $(\mathbf{X}, \mathbf{y})$, split into two sets $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ and $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$.
- Train q models $f^{(1)}, \ldots, f^{(q)}$ by finding parameters which minimize the loss on $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$.
- Evaluate loss of each trained model on $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$.

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.

TRAIN-TEST PARADIGM



- **Train loss** continues to decrease as model complexity grows.
- **Test loss** “turns around” once our model gets too complex. Minimized around degree $3-4$.

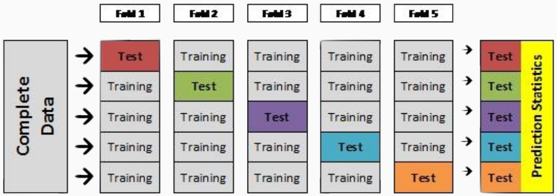
GENERALIZATION ERROR

If the test loss remains low, we say that the model **generalizes**. Test lost 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

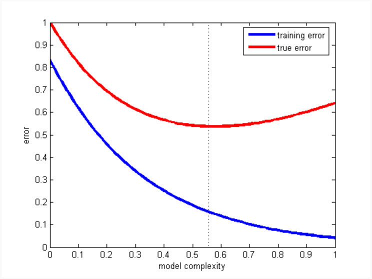


- Randomly divide data in K parts.
    - Typical choice: $K = 5$ or $K = 10$.
- Use $K - 1$ parts for training, 1 for test.
- For each model, compute test loss L_{ts} for each “fold”.
- Choose model with best average loss.
- Retrain best model on entire dataset.

Is there any disadvantage to choosing K larger?

THE FUNDAMENTAL CURVE OF ML

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



TRAIN-TEST INTUITION

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

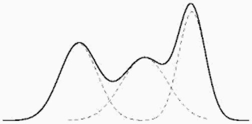
Intuition: Models which perform better on the test set will generalize better to future data.

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 $(\mathbf{x}, y) \sim \mathcal{D}$.


✓

✓

✓

E.g. $x_1, \ldots, x_d$ are Gaussian random variables with parameters $\mu_1, \sigma_1, \ldots, \mu_d, \sigma_d$.

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

RISK

Statistical Learning Model:

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

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

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

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

RISK

- (Population) Risk:

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

- **Empirical Risk**: Draw $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n) \sim \mathcal{D}$

$$R_E(f, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L\left(f(\mathrm{x}, \boldsymbol{\theta}), y \right)$$

EMPIRICAL RISK

For any fixed model f and parameters $\boldsymbol{\theta}$,

$$\mathbb{E}\left[R_E(f, \boldsymbol{\theta})\right] = R(f, \boldsymbol{\theta}).$$

Only true if f and $\boldsymbol{\theta}$ are chosen *without looking at the data used to compute the empirical risk.*

MODEL SELECTION

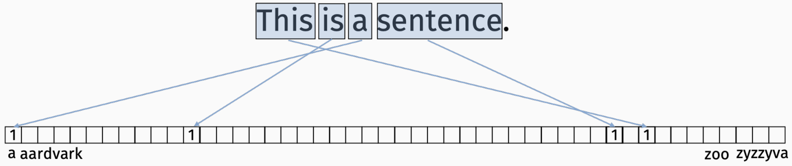
- Train q models $(f^{(1)}, \boldsymbol{\theta}_1^*), \ldots, (f^{(q)}, \boldsymbol{\theta}_q^*)$.
- For each model, compute empirical risk $R_E(f^{(i)}, \boldsymbol{\theta}_i^*)$ using test data.
- Since we assume our original dataset was drawn independently from $\mathcal{D}$, so is the random test subset.

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

MODEL SELECTION EXAMPLE

bag-of-words models and **n-grams**

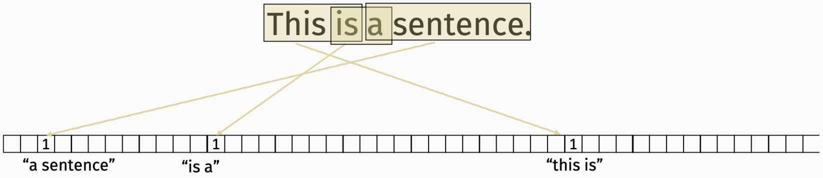
Common way to represent documents (emails, webpages, books) as numerical data. The ultimate example of 1-hot encoding.



bag-of-words

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.

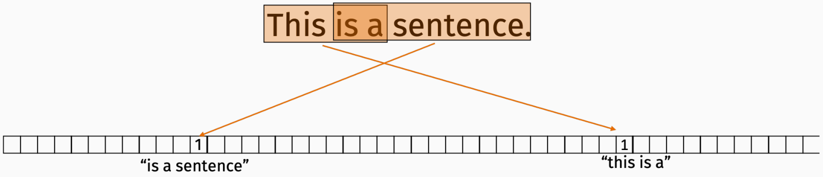


bi-grams

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.



tri-grams

MODEL SELECTION EXAMPLE

Models of increasing order:

- Model $f^{(1)}_{\boldsymbol{\theta}_1}$: spam filter that looks at **single words**.
- Model $f^{(2)}_{\boldsymbol{\theta}_2}$: spam filter that looks at **bi-grams**.
- Model $f^{(3)}_{\boldsymbol{\theta}_3}$: spam filter that looks at **tri-grams**.
- ...

"interest" "low interest" "low interest loan"

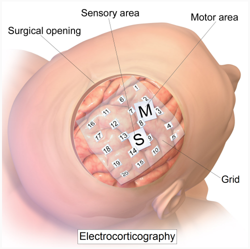
Increased length of **n-gram** means more expressive power.

Will also be relevant in our first generative ML lab!

MODEL SELECTION EXAMPLE

Electrocorticography ECoG (upcoming lab):

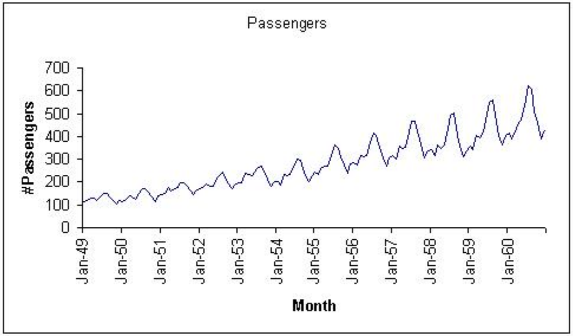
- Implant grid of electrodes on surface of the brain to measure electrical activity in different regions.



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

AUTOREGRESSIVE MODEL

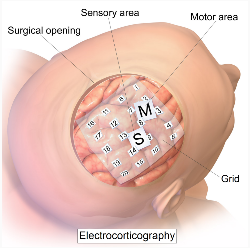
Predicting time t based on a linear function of the signals at time $t, t-1, \ldots, t-q$ is not the same as fitting a line to the time series. It's much more expressive.



Predecessor of modern "recurrent neural networks".

MODEL SELECTION LAB TIP

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.

ADAPTIVE DATA ANALYSIS

Slight caveat: This is typically not how machine learning or scientific discovery works in practice!

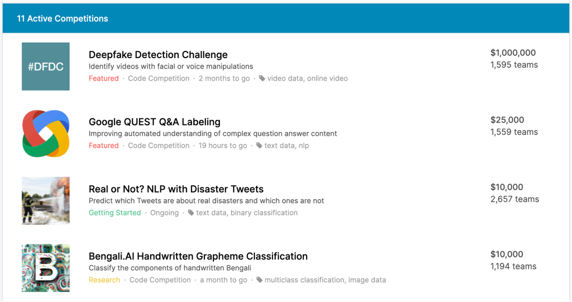
Typical workflow:

- 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.

ADAPTIVE DATA ANALYSIS

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



11 Active Competitions

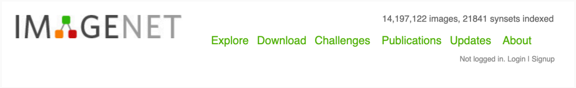
Deepfake Detection Challenge
Identify videos with facial or voice manipulations
Featured · Code Competition · 2 months to go · video data, online video
$1,000,000
1,595 teams

Google QUEST Q&A Labeling
Improving automated understanding of complex question answer content
Featured · Code Competition · 19 hours to go · text data, nlp
$25,000
1,559 teams

Real or Not? NLP with Disaster Tweets
Predict which Tweets are about real disasters and which ones are not
Getting Started · Ongoing · text data, binary classification
$10,000
2,657 teams

Bengali.AI Handwritten Grapheme Classification
Classify the components of handwritten Bengali
Research · Code Competition · a month to go · multiclass classification, image data
$10,000
1,194 teams

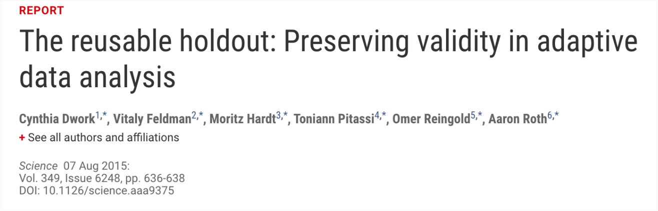
Kaggle (various competitions)



14,197,122 images, 21841 synsets indexed

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.

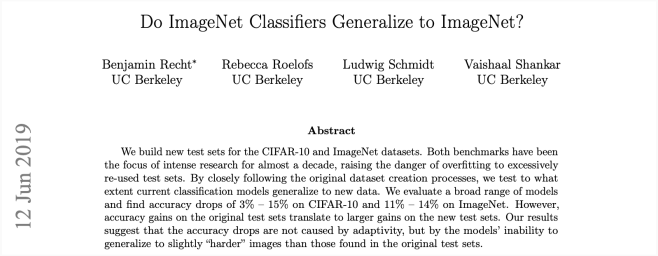

REPORT

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?

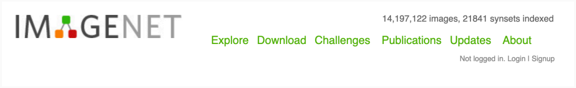
Benjamin Recht* UC Berkeley
Rebecca Roelofs UC Berkeley
Ludwig Schmidt UC Berkeley
Vaishaal Shankar UC Berkeley

12 Jun 2019

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

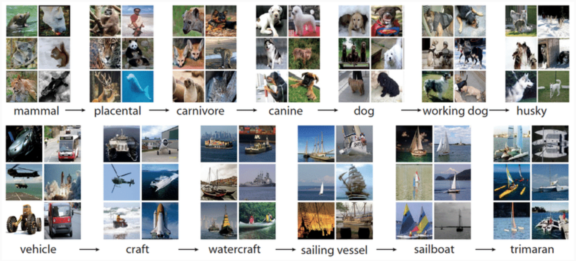

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications Updates About

Not logged in. Login I Signup

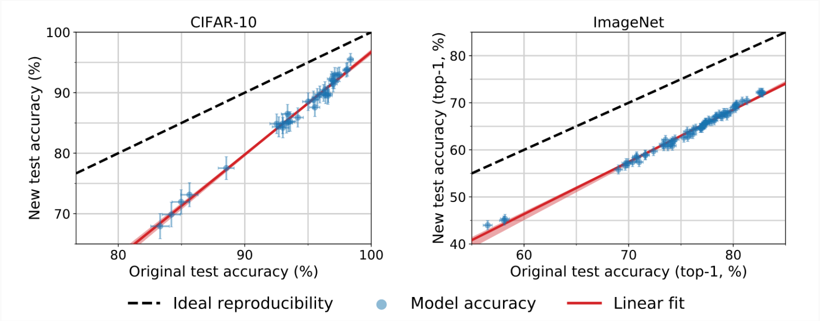
Collected by Fei-Fei Li's group at Stanford in 2006ish and labeled using Amazon Mechanical Turk.



mammal → placental → carnivore → canine → dog → working dog → husky
vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

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.

REGULARIZATION

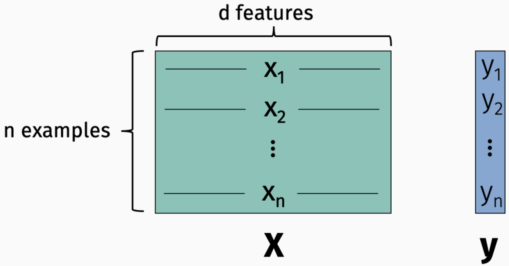
over-parameterized models

In all the model selection examples we discussed we had full control over the complexity of the model: could range from underfitting to overfitting.

In practice, you often don't have this freedom. Even the most basic model might lead to overfitting.

OVER-PARAMETERIZED MODELS

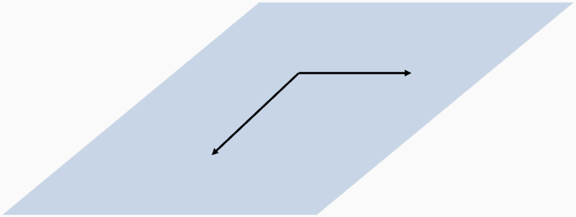
Example: Linear regression model where $d \geq n$. Almost always the case e.g. when using bag-of-words features.



Can (almost) always find β so that $\mathbf{X}\beta = \mathbf{y}$ exactly.

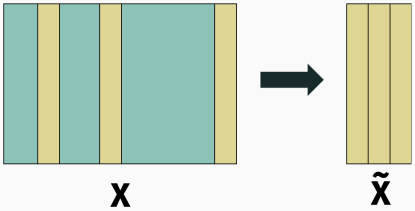
HIGH DIMENSIONAL LINEAR MODELS

Claim: For almost all sets of n length n vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, we can write any vector $\mathbf{y}$ as a linear combination of these vectors.



FEATURE SELECTION

Select some subset of features to use in model:



Filter method: Compute some metric for each feature, and select features with highest score.

- Example: compute loss or R^2 value when each feature in $\mathbf{X}$ is used in single variate regression.

Any potential limitations of this approach?

FEATURE SELECTION

FEATURE SELECTION

Exhaustive approach: Pick best subset of q features.

Faster approach: Greedily select q features.

Stepwise Regression:

- **Forward:** Step 1: pick single feature that gives lowest loss. Step k: pick feature that when combined with previous $k-1$ chosen features gives lowest loss.
- **Backward:** Start with all of the features. Greedily eliminate those which have least impact on model performance.

Feature selection deserves more than two slides, but we won't go into too much more detail!

ALTERNATIVE APPROACH

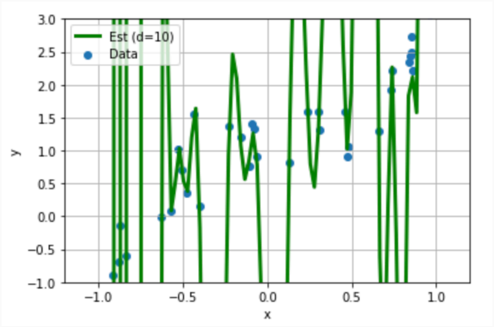
Regularization: Explicitly discourage overfitting by adding a regularization penalty to the loss minimization problem.

$$\min_{\boldsymbol{\theta}} \left[L(\boldsymbol{\theta}) + Reg(\boldsymbol{\theta})\right].$$

Example: Least squares regression. $L(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$.

- Ridge regression (ℓ_2): $Reg(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2^2$
- LASSO (least absolute shrinkage and selection operator) (ℓ_1): $Reg(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$
- Elastic net: $Reg(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$

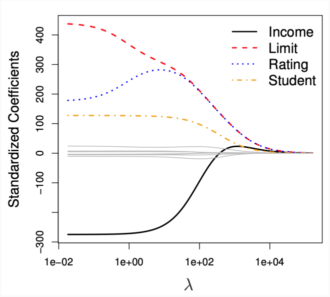
REGULARIZATION



RIDGE REGULARIZATION

Ridge regression: $\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_2^2$.

- As $\lambda \to \infty$, we expect $\|\beta\|_2^2 \to 0$ and $\|\mathbf{X}\beta - \mathbf{y}\|_2^2 \to \|\mathbf{y}\|_2^2$.
- Feature selection methods attempt to set many coordinates in β to 0. Ridge regularizations encourages coordinates to be small.



DUALITY WITH CONSTRAINED REGRESSION

$$\text{Ridge regression: } \min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

- Can be viewed as shrinking the size of our model class. Relaxed version of $\min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2^2<c} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$.

Claim: For any λ, let $\boldsymbol{\beta}_\lambda^* = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$. Then there is some $c(\lambda)$ such that:

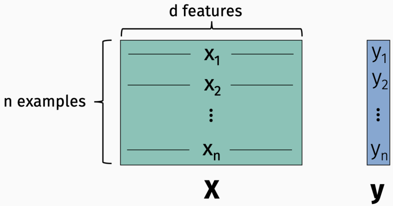
$$\boldsymbol{\beta}_\lambda^* = \mathop{\arg\min}_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2^2<c(\lambda)} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2.$$

Moreover, we have the for $\lambda' > \lambda$, $c(\lambda') < c(\lambda)$.

RIDGE REGULARIZATION

Ridge regression: $\min_{\beta} \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2$.

- $\min_{\beta:\|\beta\|_2^2<c} \|X\beta - y\|_2^2$ won't have zero error solution for all y, even when over-parameterized.



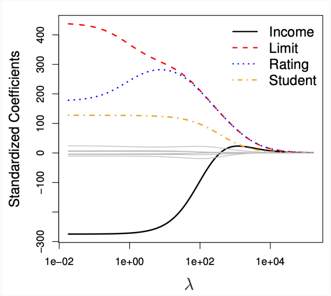
RIDGE REGULARIZATION

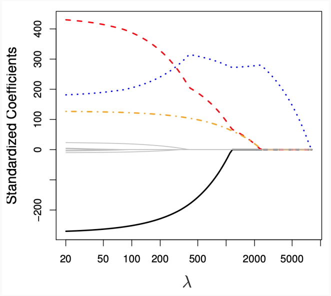
How do we minimize: $L_R(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$?

LASSO REGULARIZATION

Lasso regularization: $\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$.

- As $\lambda \to \infty$, we expect $\|\boldsymbol{\beta}\|_1 \to 0$ and $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \to \|\mathbf{y}\|_2^2$.
- Typically encourages subset of $\boldsymbol{\beta}_i$'s to go to zero, in contrast to ridge regularization.





LASSO REGULARIZATION

Pros:

- Simpler, more interpretable model.
- More intuitive reduction in model order.

Cons:

- No closed form solution because $\|\beta\|_1$ is not differentiable.
- Can be solved with iterative methods, but generally not as quickly as ridge regression.

REGULARIZATION

Notes:

- Model selection/cross validation used to choose optimal scaling λ on $\lambda\|\boldsymbol{\beta}\|_2^2$ or $\lambda\|\boldsymbol{\beta}\|_1$.
- Often grid search for best parameters is performed in "log space". E.g. consider $[\lambda_1, \ldots, \lambda_q] = 1.5^{[-4,-3,-2,-1,-0,1,2,3,4]}$.
- Regularization methods are not invariant to data scaling. Typically when using regularization we mean center and scale columns to have unit variance.