Week 7 Inference for regression

3. Stat140 Essentials and Beyond

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

The way the data are/were collected determines the scope of inference

- For generalizing to the population: was it a random sample? Was there sampling bias?
- For assessing causality: was it a randomized experiment?

Collecting good data is crucial to making good inferences based on the data

Before doing inference, always explore your data with descriptive statistics

- Always visualize your data! Visualize your variables and relationships between variables
- Calculate summary statistics for variables and relationships between variables – these will be key for later inference
- The type of visualization and summary statistics depends on whether the variable(s) are categorical or quantitative

For good estimation, provide not just a point estimate, but an interval estimate which takes into account the uncertainty of the statistic

Confidence intervals are designed to capture the true parameter for a specified proportion of all samples

A P% confidence interval can be created by

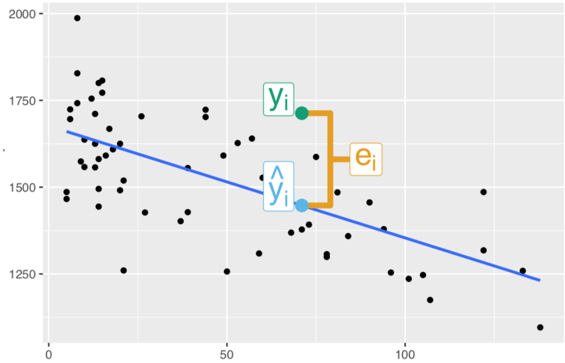
- bootstrapping (sampling with replacement)
- statistic $\pm z^* \times SE$

A p-value is the probability of getting a statistic as extreme as observed, if H_0 is true

The p-value measures the strength of the evidence the data provide against H0

- If the p-value is low, reject H_0
- If p-value is not low, then the test is inconclusive

So far, regression is a way to predict one response variable with one explanatory variables

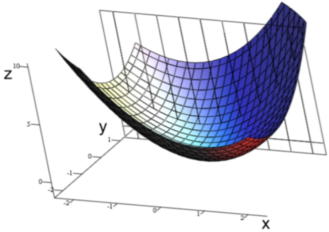


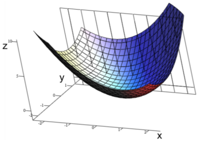
Write the steps out mathematically,

- Given a finite data set: $(x_i, y_i)_{i=1}^{n}$
- We model with $y = b_0 + b_1 x$
- Find b_0 and b_1 so that $L(b_0, b_1) := \sum_{i=1}^{n} (y_i - b0 - b_1 x)^2$ is minimized
- This is a problem of **optimization**!

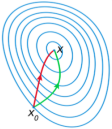
Add a new variable z

- Given a finite data set: $(x_i, y_i, z_i)_{i=1}^n$
- We model with $z = b_0 + b_1 x + b_2 y$
- Find b_0, b_1, b_2 so that $L(b_0, b_1, b_2) := \sum_{i=1}^n (z_i - b_0 - b_1 x_i - b_2 y_i)^2$ is minimized
- This is again a problem of **optimization**!





Using the gradient, which is a generalization of the derivative to multiple dimensions, we can find a way to descend on the surface step by step. **Take Multivariable Calculus (MATH 203)!**



Since our loss function $L(b_0, b_1, b_2)$ is convex, we will eventually reach the line of best fit. **Take Optimization (MATH 339)!**

- The variable you want to predict Y (say the price of Tesla stock tomorrow).
- The features used to predict $X_1, X_2, \ldots, X_k$ (say the weather, the stock prices of a 100 different related stocks on the previous day, etc.)
- The form of the regression function and the parameters defining them $F_\theta : X_1 \times X_2 \cdots \times X_n \to Y$ (this varies for every kind of regression strategy).
- Large quantities of training data.
- A loss function based on the data $L(\theta)$, which we are trying to minimize in order to find the best $F(\theta)$
- An optimization algorithm for minimizing $L(\theta)$.
- Validating the function on test data.

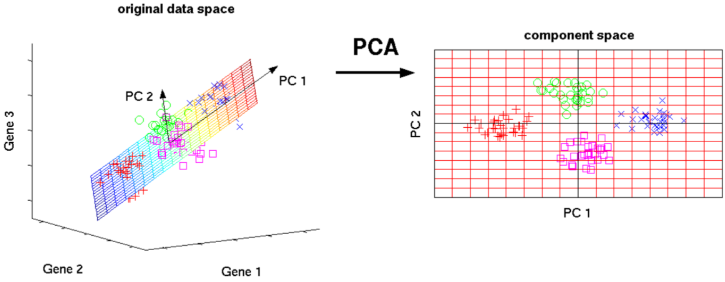
How to teach a robot to be able to recognize images as either a cat or a non-cat? This sounds like a biology problem. How can we formulate this as a regression problem?

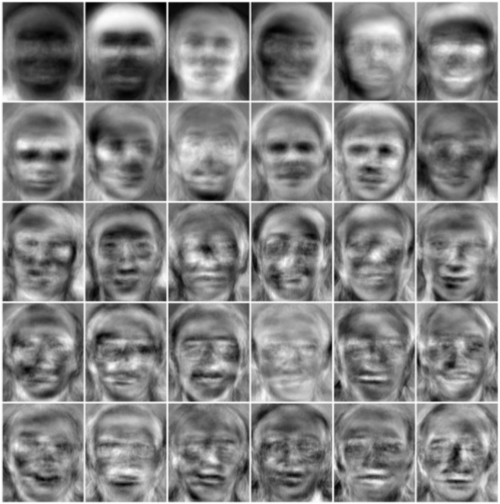
- Everything is data!
- $\mathbb{R}^{3\times1000\times1000}$ is a space of 1000 by 1000 rgb images
- $C \subset \mathbb{R}^{3\times1000\times1000}$ is the cat subspace
- Try to learn the classifier function $f_C : \mathbb{R}^{3000000} \to \{1, -1\}$ so that $f_C(x) = 1$ if $x \in C$.

Take Linear Algebra (MATH 211) and Machine learning (CS335)!

Say we want to classify 32 $\times$ 32 faces. That means 1024 features or dimensions. Hard problem! Curse of dimensionality.

"Dimension Reduction" or "Representation Learning"

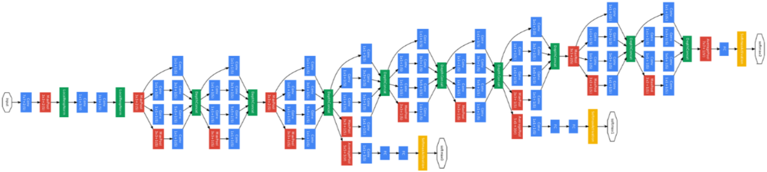




Now we can classify faces:

- Raw images to Eigenface basis coordinates
- $\mathbb{R}^{32\times 32} \to X_1 \times \ldots X_k \to Y$
- We learn the feature representation $F : \mathbb{R}^{32\times 32} \to X_1 \times \ldots X_k$ first
- Then, we learn classifier $X_1 \times \ldots X_k \to Y$

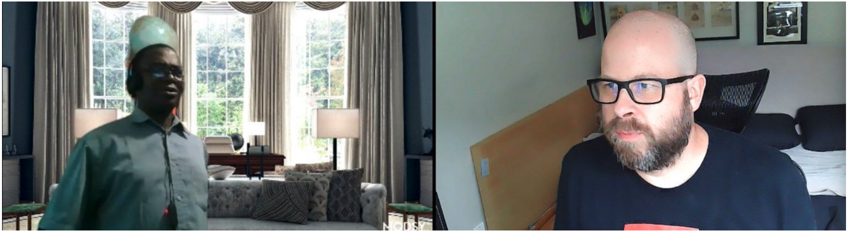
Deep learning



We don't really understand why it works, it is very hard to analyze non-convex heuristic optimization.

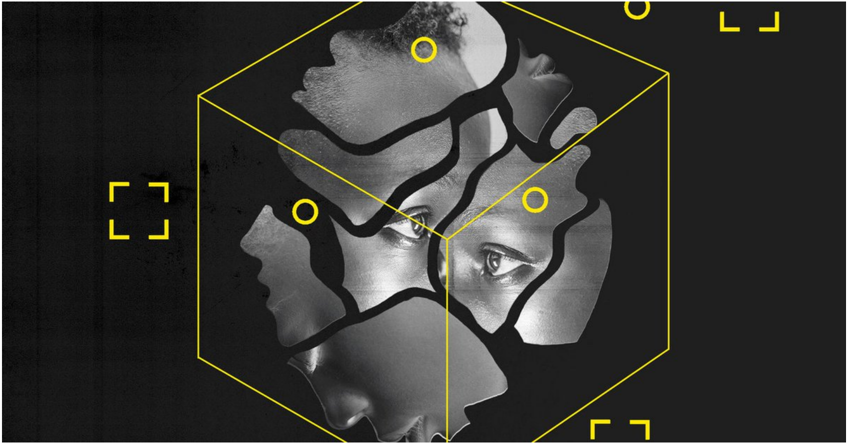
The Connection Between Applied Mathematics and Deep Learning
https://sinews.siam.org/Details-Page/the-connection-between-applied-mathematics-and-deep-learning

Where is my face?

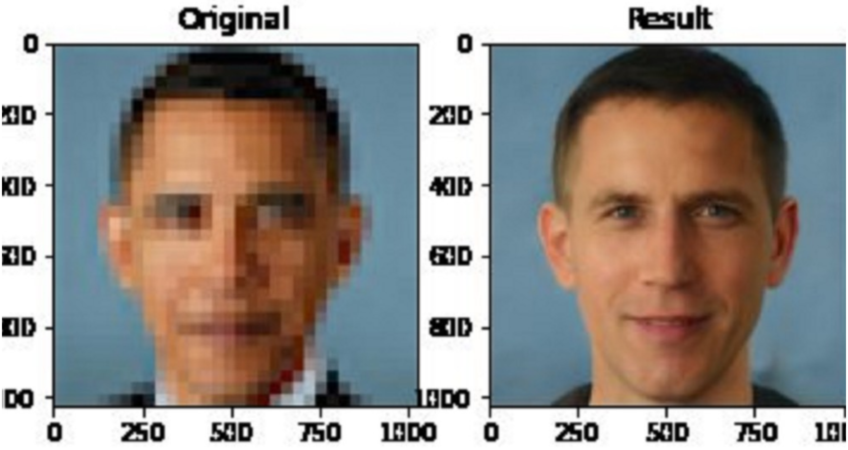


Turns out Zoom has a crappy face-detection algorithm that erases black faces...and determines that a nice pale globe in the background must be a better face than what should be obvious.

The best algorithm still struggles to detect black faces



https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/


Original
Result
0
200
400
600
800
1000
0
250
500
750
1000

Bias in the AI system

- A training dataset that isn't representative
- A training dataset that has societal bias baked in
- A poorly chosen objective function in an ML model

What can you do?

- Defining and following a set of AI principles: `https://ai.google/responsibilities/responsible-ai-practices/`
- Investing in tools and technology approaches to support the operationalization of the principles, e.g, AI Fairness 360 `https://aif360.mybluemix.net`
- Diversify your team `https://arxiv.org/pdf/2002.11836.pdf`


WITH GREAT POWER
COMES GREAT
RESPONSIBILITY...