Linear Regression

Jiatong

Content

- Linear Regression: Definition
- Linear Regression: Property*
- Linear Regression: Interpretation*
- Linear Regression: Some Modifications

Recap: Machine Learning Definition

- Definition1
- Machine learning allows computers to observe input and produce a desired output, either by example by identifying latent patterns in the input
- Data
  - What type of input? What type of output?
- Patterns = Algorithm
  - Intuition (empirical) / Objective (theoretical)

Recap: Machine Learning Definition

- Definition2
- Using Experience to gain expertise

- Programs that
  - Learn Rules from data
  - Adapt to changes
  - Improve performance with experience

Recap: Machine Learning Definition

- Definition3
- Fitting a function to data

- Fitting: Optimization, what parameters can we change
- Function: Model, loss function
- Data: Data/model assumptions? How we use data?

We start with Linear Regression

Why Start Here

- Well-known algorithm
- Not strictly a “machine learning” algorithm
- Can learn about fundamentals with a simple example

Example

- I run a real estate website. I want to list properties for sale and provide estimates of how much they will sell for

- Goal: Predict home prices
- Data: Previous home sales
  - Housing facts: size, bedrooms, age, neighborhood, listing price

- How do I do this

Data Model

- Assume dependent variable (y) can be modeled by a linear function of the input variables (x)

- $Y = wx + b$
- 2 dimensions
    - Compute x and b from two points

- Solution
    - Given y and x, solve for w and b

Regression

- Data $\{(x_i, y_i)\}_{i=1}^{N}$ $\quad x_i \in \mathbb{R}^M$ $y_i \in \mathbb{R}$
- Learn: a mapping from x to real valued y
  - f(x)=y
- Examples
  - GPAs
  - Stock price
  - Miles per gallon
  - Age of author

Fitting a function to data

- Fitting: solve for w given y and x
- Function: linear function
- Data: assume dependent variable linear combination of independent variables
- Minimize a function:
    - What function are we minimizing?

What is the goal?

- In Statistics:
  - linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X." (Wikipedia)

- ML goal: predict correctly the next example
  - Minimize: reduce prediction error

Loss Functions

- Machine learning algorithms minimize loss functions
  - Or some substitute for a loss function
  - The best solution minimizes the loss function*
- Definition
  - A function that maps between (true label, prediction) -> non-negative number
  - 0 = perfect prediction

Loss: What We Minimize

- Loss measures the badness of our prediction
- What’s a good loss function?
  - It depends on task and goals

- Regression loss function?
  - Proposal: How far are you from the correct answer

Sum of Squares Loss

- $f(x) - y$
- $(f(x) - y)^2$
- $\sum_{i=1}^{\mathrm{n}}(f(x_i) - y_i)^2$
- $\sum_{i=1}^{\mathrm{n}}(w \cdot x_i - y_i)^2$

Goal of Learning

- True error
- $\sum_{i=1}^{\infty}(w \cdot x_i - y_i)^2$
- We need infinite data to measure this

Goal of Learning (Cont’d)

- If we can’t measure true error, how do we judge learning success?
- Should an algorithm maximize performance on observed data?

- Proposal: Measure error on the given data
  - Call this the “training data”
  - Is this a good idea?

Measuring Error

- Very bad idea for Machine learning (not that bad for statistics)
- Recall: machine learning cares about prediction (the future)
  - How well will the system do once deployed?

- Memorizing the training data is easy

Generalization

- Generalization
  - The ability of an algorithm to generalize knowledge learned from observed data to new data

- Simple example: memory based classifier
  - Binary classification
  - Train: remember each example
  - Test: if we have seen an example before, report label
    - Otherwise, guess randomly

- Train error: 0%, test error: 50%

Assumption for Linear Regression

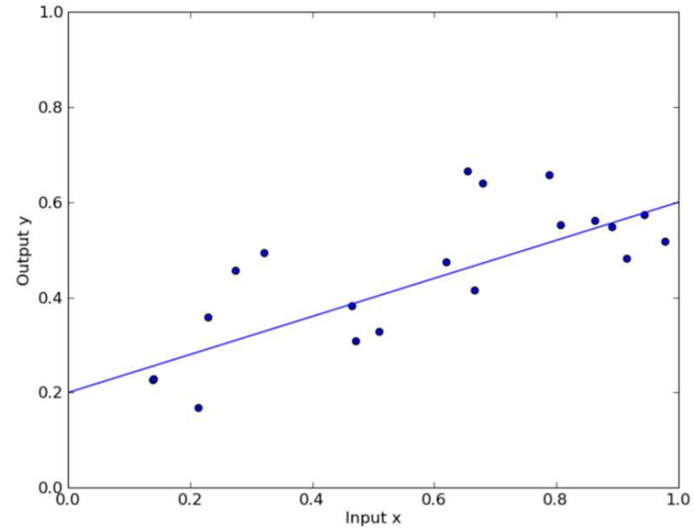
- We assumed output (y) is linear combination of inputs
- This is wrong
  - Rarely is data actually linear
- But realistic assumption may be too complex
- A reasonable middle ground?

Noise from a Gaussian Distribution

- Assume output permuted by Gaussian noise
- $y = wx + \varepsilon$
- $\varepsilon \sim N(\mu, \sigma^2), \mu = 0, \sigma^2 = 1$
- The data isn't really generated in this way
  - Assume that it is for sake of modeling (Law of Large Numbers)

Probability of Output

- $p(y|x, w, \sigma^2)$
- $= N(y, \sigma^2)$
- $= N(wx, \sigma^2)$



Two common methods

- Ordinary Least Square vs Maximum Likelihood
- They will reach the same solution to the data (mostly)

Maximum Likelihood

- Likelihood = probability of overserving data

- Writing likelihood
  - Assume data generated from our linear regression

Blackboard Session

Bias?

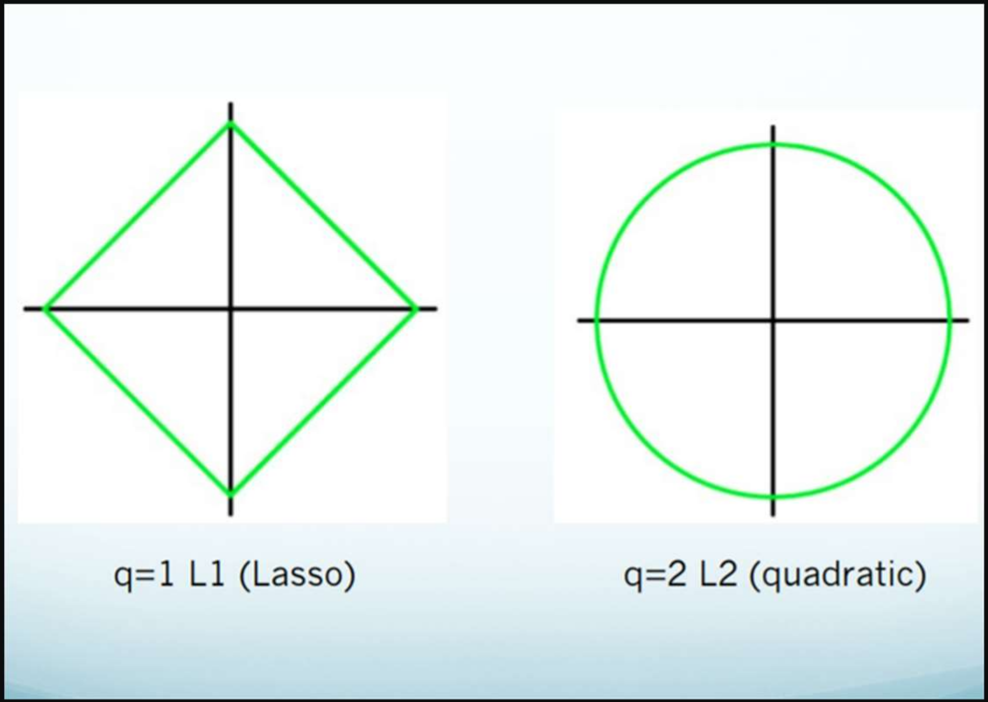
- Gaussians: Maximum likelihood estimate is biased
  - This is ok if we have infinite data
  - We never have infinite data!

- Overfitting: avoid it by favoring certain solutions

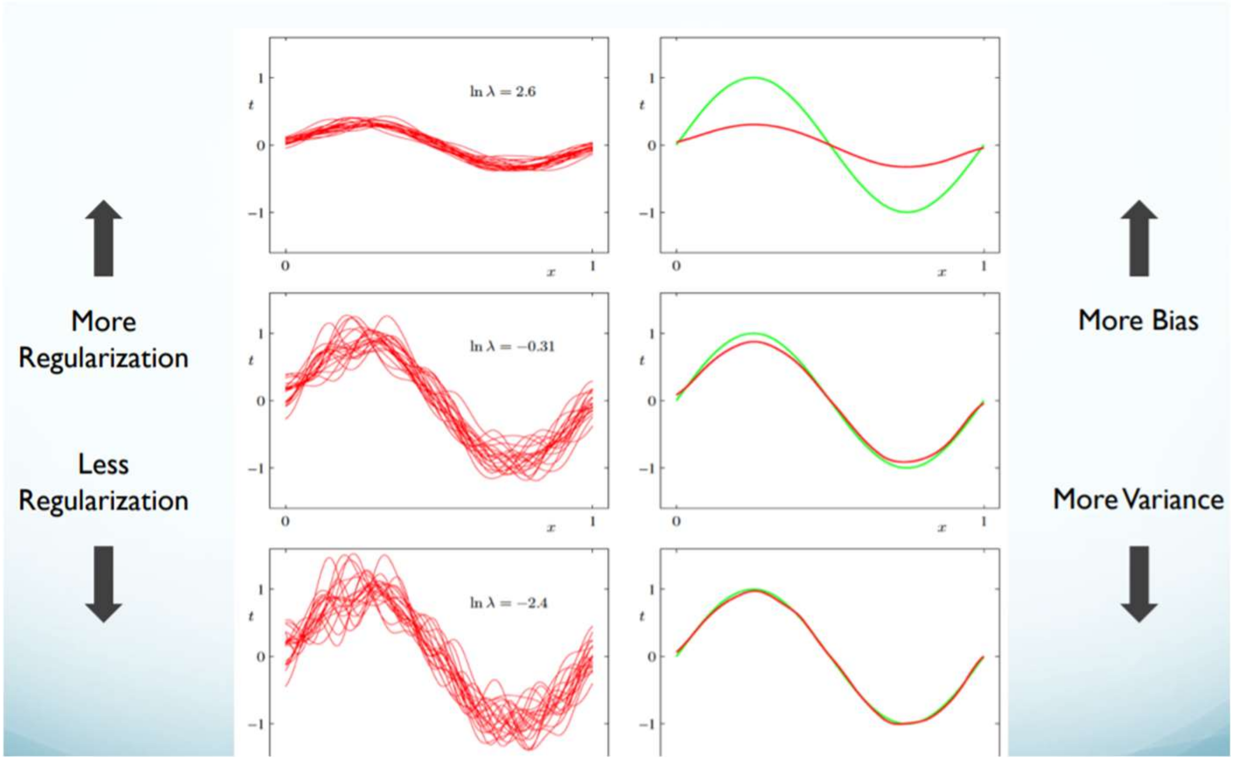
- Regularization
  - Add term to objective to favor different considerations
  - What should we favor: simpler is better → favor small weights

Blackboard Session for Regularization

Regularization Behavior



Parameter Tradeoff



Interpretation

- http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis

Property

- https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem
- https://statisticsbyjim.com/regression/gauss-markov-theorem-ols-blue/

Paper

- Find your interest
  - https://arxiv.org/
  - Some problem in real world (predicting crowdfunding success, predicting housing/car/stock price, image, speech, language processing)

- Read more papers (literature review)
- Think about what makes you different from previous work
- Prepare data/methodology
- Experiments
- Write the paper

Examples

- Some volunteers please~

My example

A different story --- Reviewing paper

- Find some fields you like
- Read many many many paper
- Reformat them
- Think about some future directions
- Write the paper

Examples~

Some Possible Fields

- Various of Predictions (stock, used car, crowdfunding, housing…
- https://www.kaggle.com/competitions
- https://www.kaggle.com/datasets
- https://elitedatascience.com/datasets
- CV, NLP, Speech