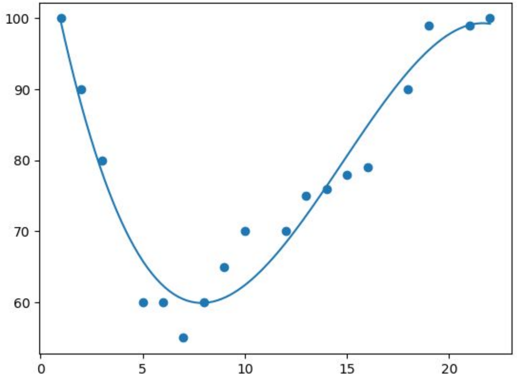
Machine Learning

Summer Springboard

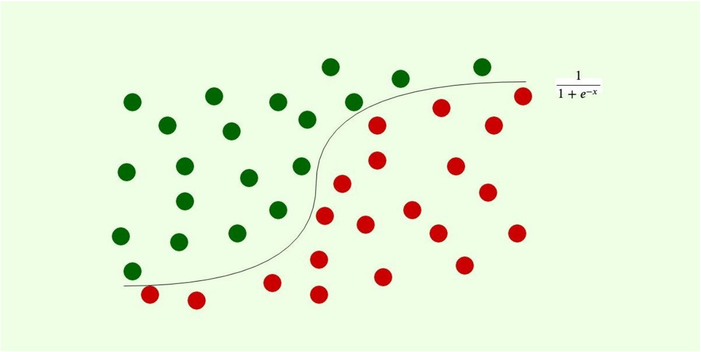
SUMMER SPRINGBOARD
Look Inward. Go Upward.

What is machine learning?

- Machine learning is a form of statistical analysis for finding trends in data, on the assumption that it will generalize to new data
- Machine learning is the search for the "best fit" lines through points



Python Machine Learning Polynomial Regression (w3schools.com)



Logistic regression and Keras for classification » AI Geek Programmer

SUMMER SPRINGBOARD
Look Inward. Go Upward.

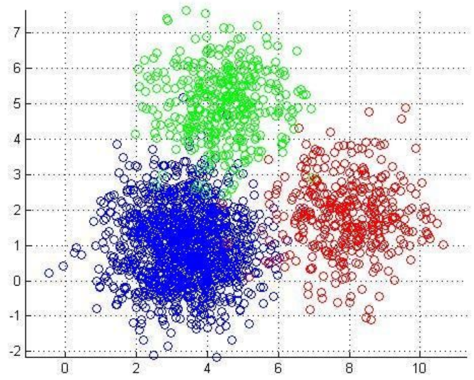
The goal

	Age	Height	Weight	…	Health Indicator
Person 1	37	5’7”	180		High
Person 2	33	5’4”	140		Low

- Learn a function f(age, height, weight,...) = Health Indicator
- Which allows me to ask the question f(New Person) = ?
- We call the columns 2 through n-1 *features* and the last column the *label*
- We call the set of all possible combinations of features the *feature space* and the set of all possible labels the *label space*

Types of machine learning

- Supervised: My dataset contains the labels
- Reinforcement: If I take an action, I can observe the label
  - Robotics
  - Ad targeting
- Unsupervised: My dataset contains no labels
  - Clustering
- Semi-Supervised: I am trying to predict multiple things, some with labels and some without



SUMMER SPRINGBOARD
Look Inward. Go Upward.

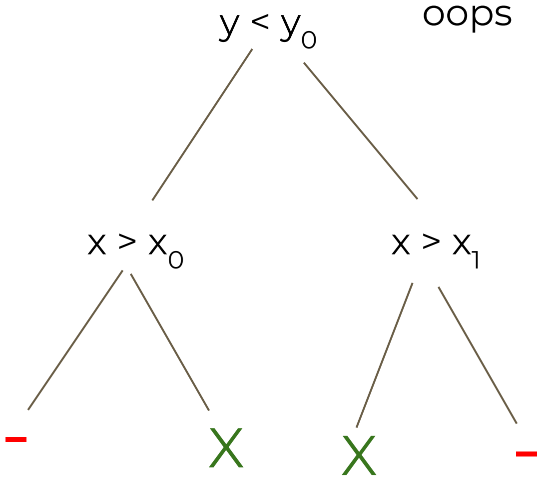
Supervised learning tasks

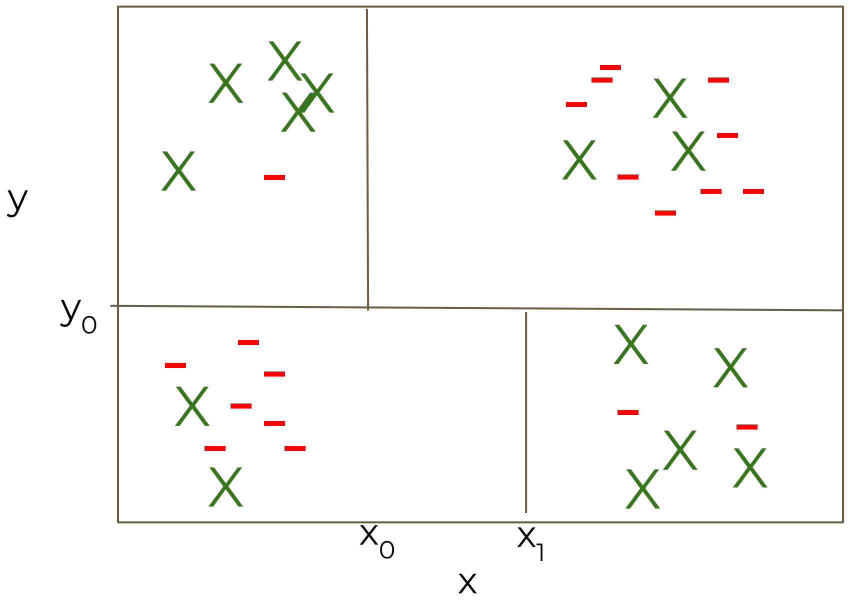
- There are primarily two supervised learning tasks
  a. Classification: Labels are *categorical*
     - apple / orange / banana
     - Disease / no disease
  b. Regression: Labels are *numerical*
     - Revenue
     - Energy consumption

SUMMER SPRINGBOARD
Look Inward. Go Upward.

Decision Trees

oops





Let's do an MVP
(minimum viable product)

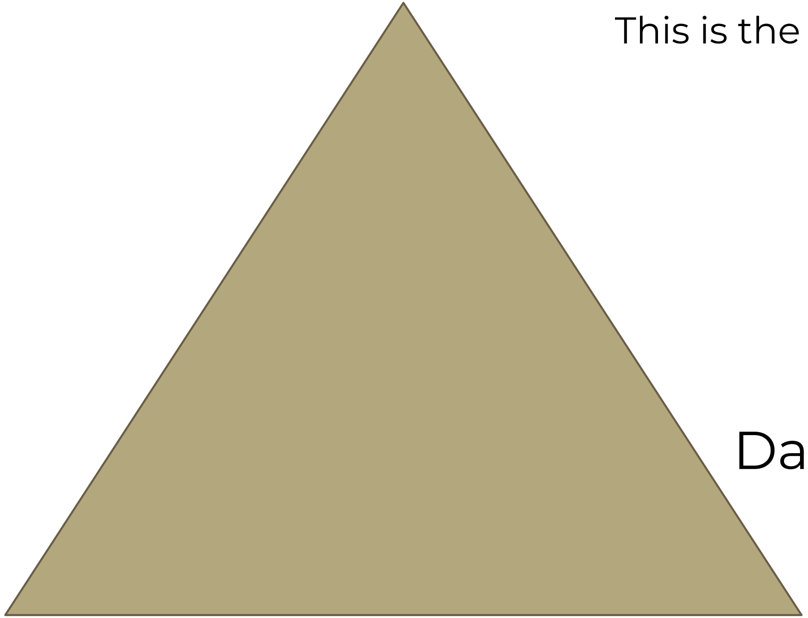
What went wrong?

SUMMER SPRINGBOARD
Look Inward. Go Upward.

First a word on evaluation

Optimization

What researchers do
This is the fun part



Data Manipulation

What practitioners do
This is the part that makes the biggest difference

Evaluation

No one really thinks about this
This is the single most important part by far

SUMMER
SPRINGBOARD
Look Inward. Go Upward.

There are four possibilities

		Predicted	
		X	–
Truth	X	True Positive	False Positive
	–	False Negative	True Negative

Pick one

Examples of preferring:

- False positives (Type I Error)
  - Medical testing (with non-invasive interventions)
  - Security
- False negatives (Type II Error)
  - Fraud detection
  - Spam filtering
- What about medical testing with *invasive* interventions?

Accuracy (isn't good enough)

- Accuracy = $\dfrac{TP + TN}{TP + FP + TN + FN}$
- Do I have Dengue Fever?
  - Most Accurate model:
    - No
- F1 score = $\dfrac{2 * Precision * Recall}{Precision + Recall}$

$$precision = \frac{tp}{tp + fp}$$

Precision — Out of all the examples that predicted as positive, how many are really positive?

[Data Science in Medicine — Precision & Recall or Specificity & Sensitivity? | by Alon Lekhtman | Towards Data Science]

Back to the question: Why were our predictions perfect?

Generalization

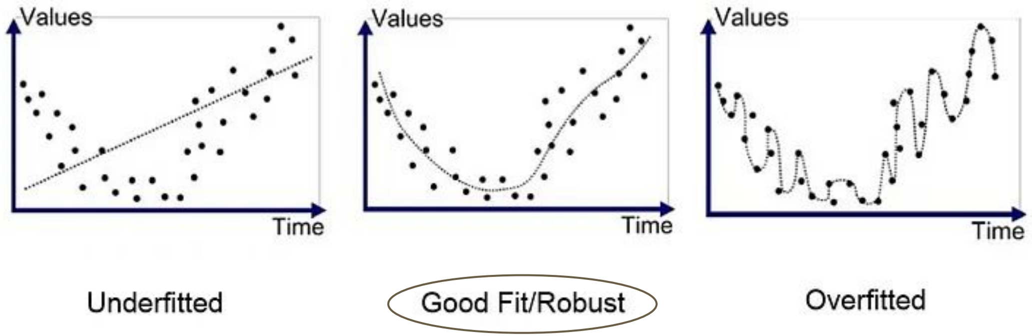
- The goal isn't to classify the data you have
  - This is labeled data! I already know the answer!
- The goal is to classify the data you *haven't collected yet*
- The problem: memorizing your training data
- The solution: evaluating on holdout data
- There are two main ways of doing this:
  - Splitting: train on most of the data and evaluate on the rest
  - Cross validation: train multiple models on different subsets of the data and average the evaluations
    - This allows you to train on more data in each "fold." We'll talk more about this later
- Let's head back to the notebook

SUMMER SPRINGBOARD
Look Inward. Go Upward.

Live in the sweet spot

I haven't actually learned anything

All I've done is memorize my training data



SUMMER SPRINGBOARD
Look Inward. Go Upward.

[Machine Learning: How to Prevent Overfitting | by Ken Hoffman | The Startup | Medium]

Let’s fix it

SUMMER SPRINGBOARD
Look Inward. Go Upward.

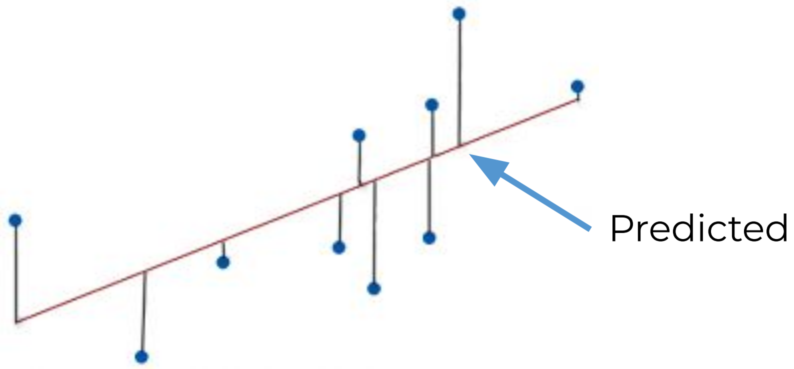
Supervised learning tasks

- There are primarily two supervised learning tasks
  a. Classification: Labels are *categorical*
     - high / medium / low
     - highly likely / somewhat likely / somewhat unlikely / highly unlikely
  b. Regression: Labels are *numerical*
     - Revenue
     - Energy consumption

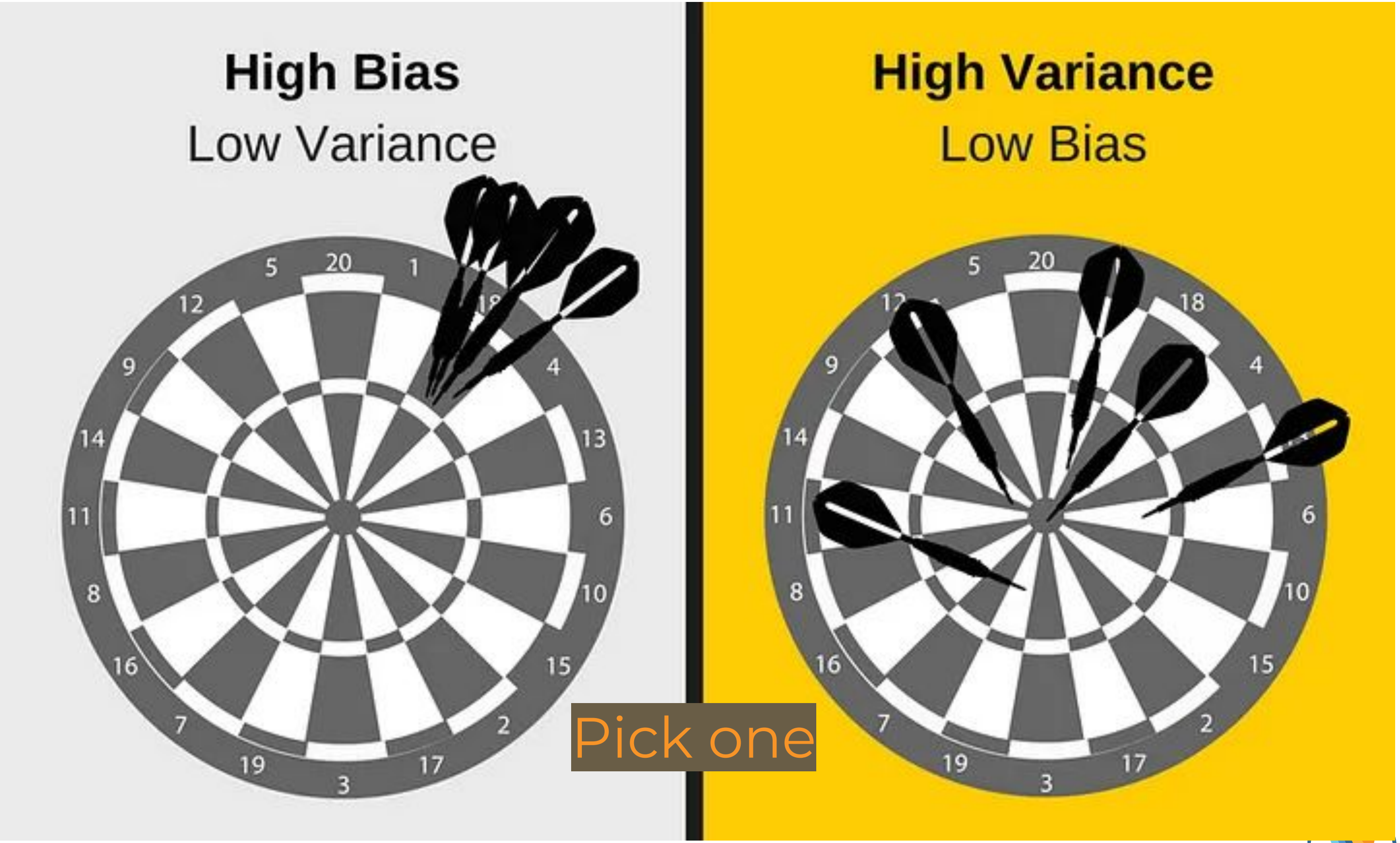
SUMMER SPRINGBOARD
Look Inward. Go Upward.

Evaluation first: RMSE

- Error: Observed - Predicted
  - Technically this is a residual, but everyone calls it error
- Squared Error:
  - Sometimes error is negative. I want this number to always be positive
- Mean Squared Error:
  - I want the *average* error
- Root Mean Square Error
  - However, that's the average error *squared* so I have to take the square root



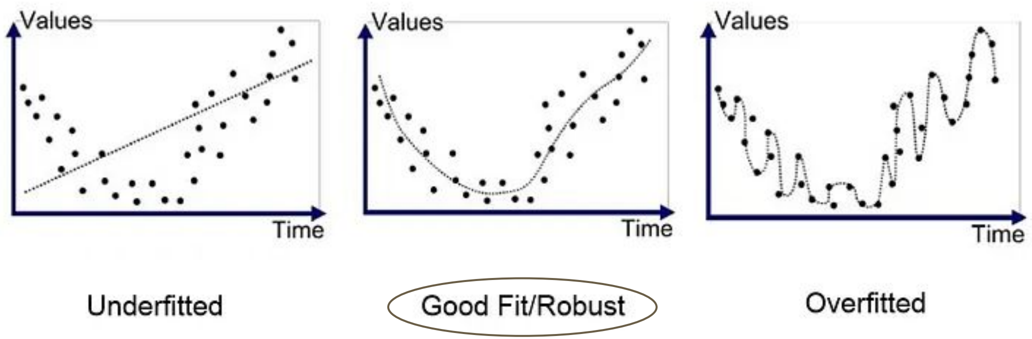
[Standard Error of the Regression vs. R-squared - Statistics By Jim]



[Bias-variance trade-off… I'm always confusing the two! | by Super Albert | The Making Of… a Data Scientist | Medium]

High bias / low variance

Low bias / high variance



SUMMER SPRINGBOARD
Look Inward. Go Upward.

Machine Learning: How to Prevent Overfitting | by Ken Hoffman | The Startup | Medium

Linear Regression

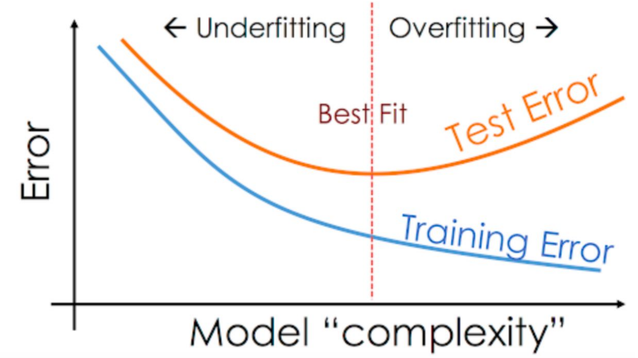
- A straight line $y = mx + b$
  - With a single input
- A straight line $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$
- The goal of linear regression is to find the vector of betas

$$\underset{B}{\operatorname{argmin}} \, ||y - BX||^2$$

- This can be found directly using linear algebra

Regularization

- You have to *earn* your complexity



[Overfitting & Underfitting in Machine Learning - Data Analytics (vitalflux.com)](#)

Degree

- This is a degree one polynomial with a single input variable

$$y = mx + b$$

- This is a degree one polynomial with n input variables

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

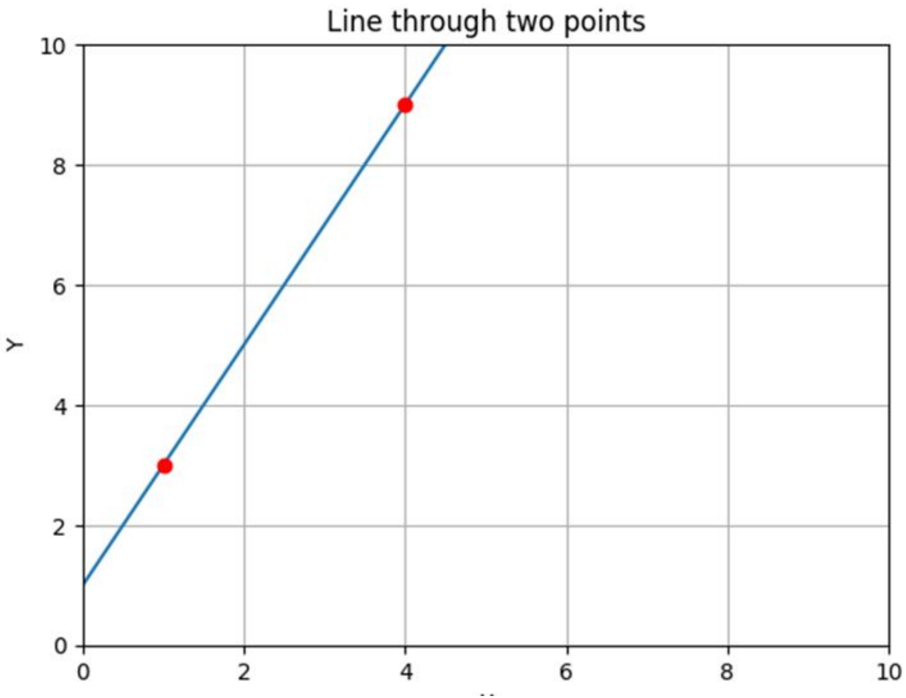
- This is a degree *two* polynomial with one input variable

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

SUMMER SPRINGBOARD
Look Inward. Go Upward.

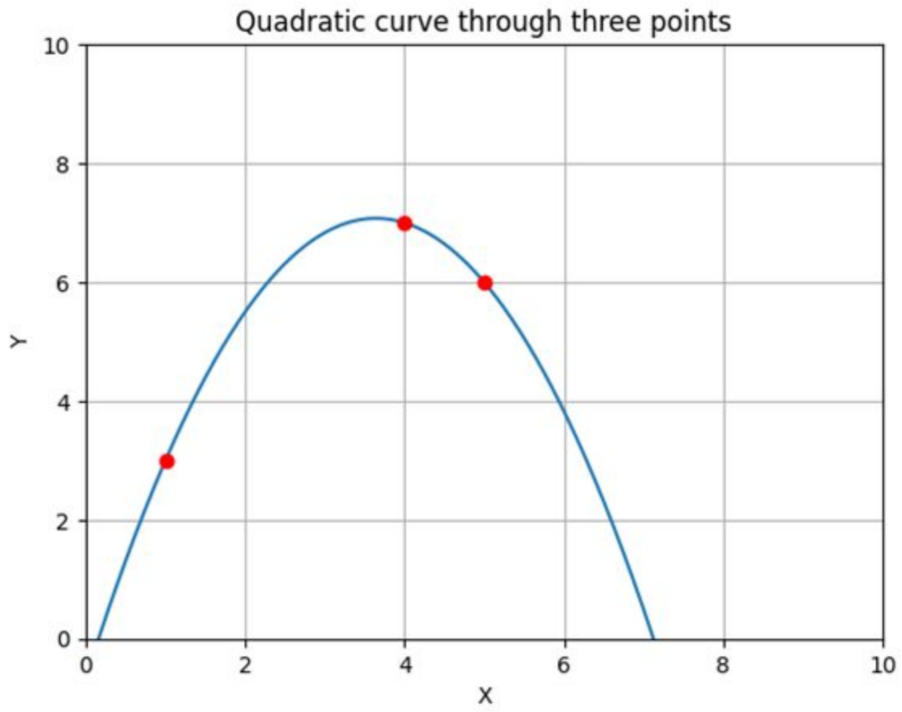
Recall

- I can draw a curve through two points with a degree one polynomial



Recall

- I can draw a curve through three points with a degree two polynomial



Technically, this is also linear

Linear!?

- You might ask in what sense is that linear, as it is not a straight line

$$y = mx + b$$

- The answer is that it is a *linear combination* of constants and variables, or the sum of the products of a constant and a variable

$$y = \sum a_i x_i$$

- For a straight line

$$a = m, b$$
$$x = x, 0$$

- For a quadratic

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

$$a = a, b, c$$
$$x = x^2, x, 0$$

- So what isn't linear?

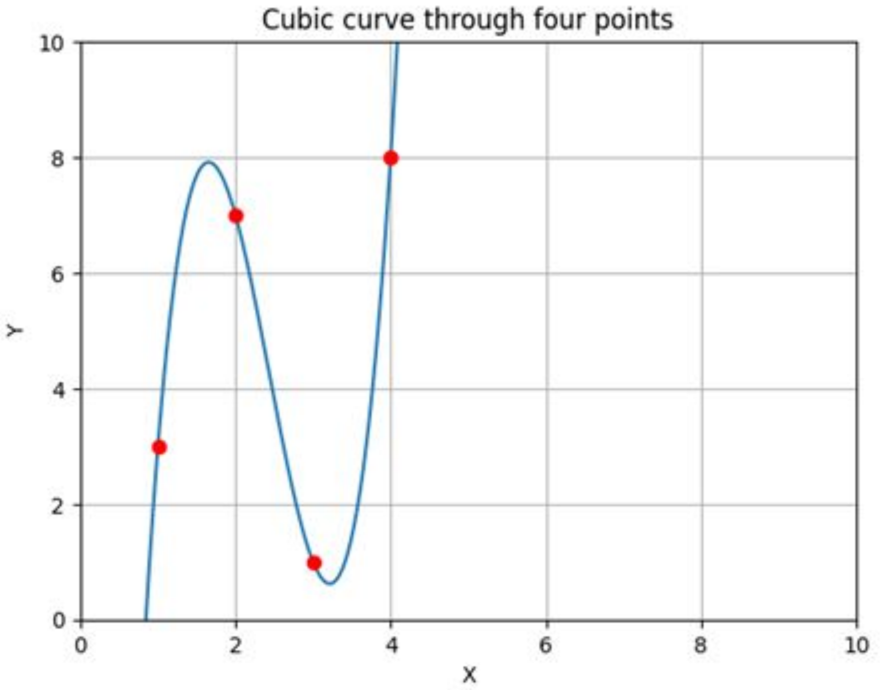
$$y = mx_1x_2$$

is not a linear combination of constants and variables

SUMMER SPRINGBOARD
Look Inward. Go Upward.

Recall

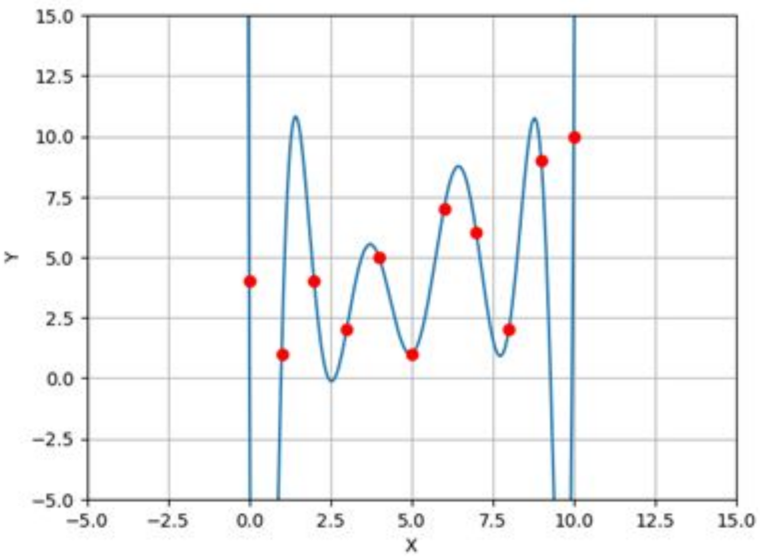
- I can draw a curve through four points with a degree three polynomial



Recall

- I can draw a curve through n+1 points with a degree n polynomial

Still linear



SUMMER SPRINGBOARD
Look Inward. Go Upward.

Recall

- I can draw a curve through two points with a degree one polynomial
- I can draw a curve through three points with a degree two polynomial
- I can draw a curve through four points with a degree three polynomial
- …
- I can draw a curve through n+1 points with a degree n polynomial
- Complexity needs to be *earned*. I shouldn't pick too high of a degree

Non-linear models and interaction effects

- This is a *non-linear model* that captures the interaction effects of x_1 and x_2

$$y = m(x_1 x_2)$$

- Complexity needs to be *earned*. I shouldn't use too many interaction effects
- (It is also no longer trivial to solve for m using linear algebra)

SUMMER SPRINGBOARD
Look Inward. Go Upward.

Regularization

- So how do I earn my complexity?
- The goal of linear regression is to find the vector of betas

$$\underset{B}{\operatorname{argmin}} ||y - BX||^2$$

- So lets add on a term that gets bigger the more complex your model is

$$\underset{B}{\operatorname{argmin}} ||y - BX||^2 + \text{“model complexity”}$$

This needs to stay small

But this can't get too big

SUMMER SPRINGBOARD
Look Inward. Go Upward.

Regression Regularization

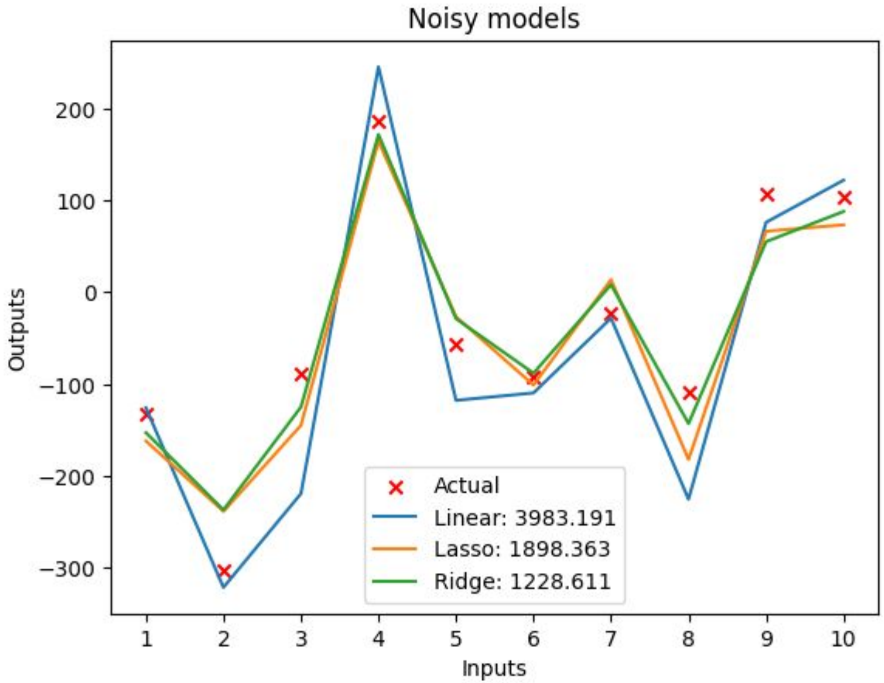
- In regression, we measure complexity in terms of the “sum” of the coefficients
- If a coefficient is large, than a small change in the input can have a large change in the output
- This sensitivity is unlikely to generalize
- The general formula is $\underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \alpha\|\beta\|^n$
- We’ll define $||\mathbf{X}||^n$ in a second, but notice how we have a new “hyper-parameter”, α, that allows us to control how much impact regularization has on our model
- If α = 0, then we are not performing any regularization
- If α is too large, then β will have to consist of all 0s
- We will talk about strategies for picking hyper-parameters soon

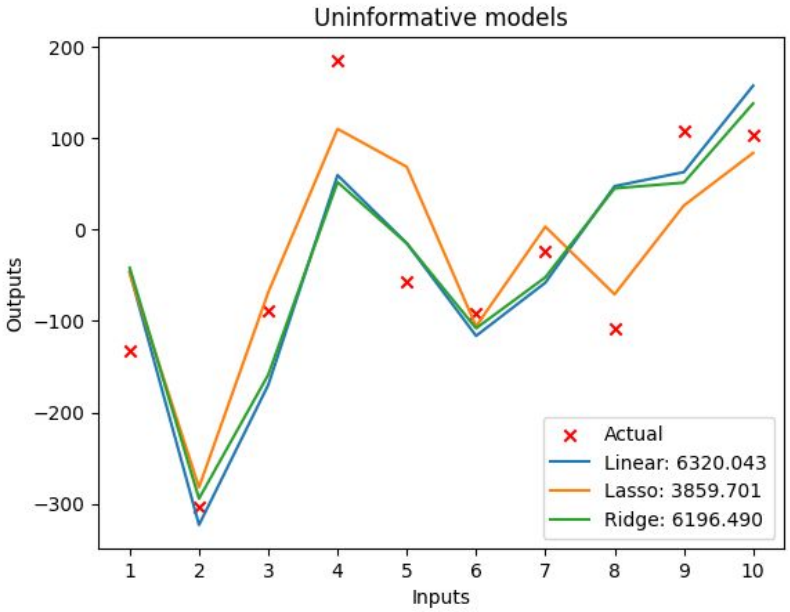
Lasso and Ridge Regression

- The two most common forms of generalization for regression models
- Lasso: n = 1 $\underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \alpha \sum |\beta_i|$
- Ridge: n = 2 $\underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \alpha \sum \beta_i^2$
- Lasso forces some coefficients towards zero
  - This allows the model to ignore uninformative features
- Ridge forces all the coefficients to shrink
  - This makes the model more robust to outliers
- (I just googled why they are called that and it is entirely uninformative. I have to look up which is which every time)

SUMMER
SPRINGBOARD
Look Inward. Go Upward.

Did it do what it was supposed to do?
How big of a difference did α make?





SUMMER SPRINGBOARD
Look Inward. Go Upward.

Logistic Regression and more Evaluation

Regression

- We *like* linear regression
  - It's fast
  - It's easy
  - It's a good place to start
    - Always do this first
      - If it doesn't work at all, you may be in trouble
      - If it's good enough, you're done
- Can we use regression for classification problems?
- Given that this is a slide, obviously yes

Logistic Regression

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

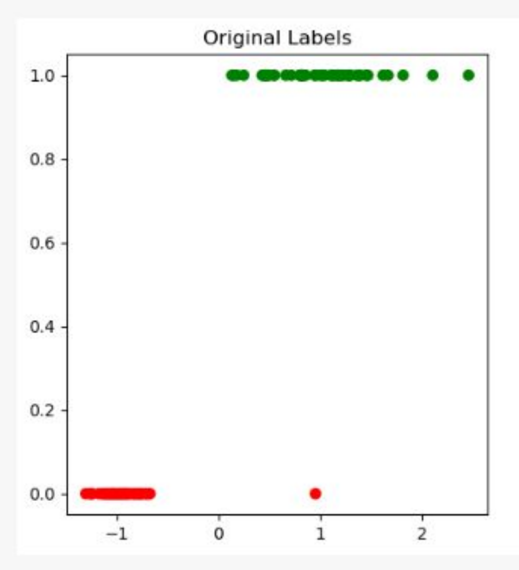
Probability

Regression Coefficients

Predicted class:

1 if p(x) > t for some t

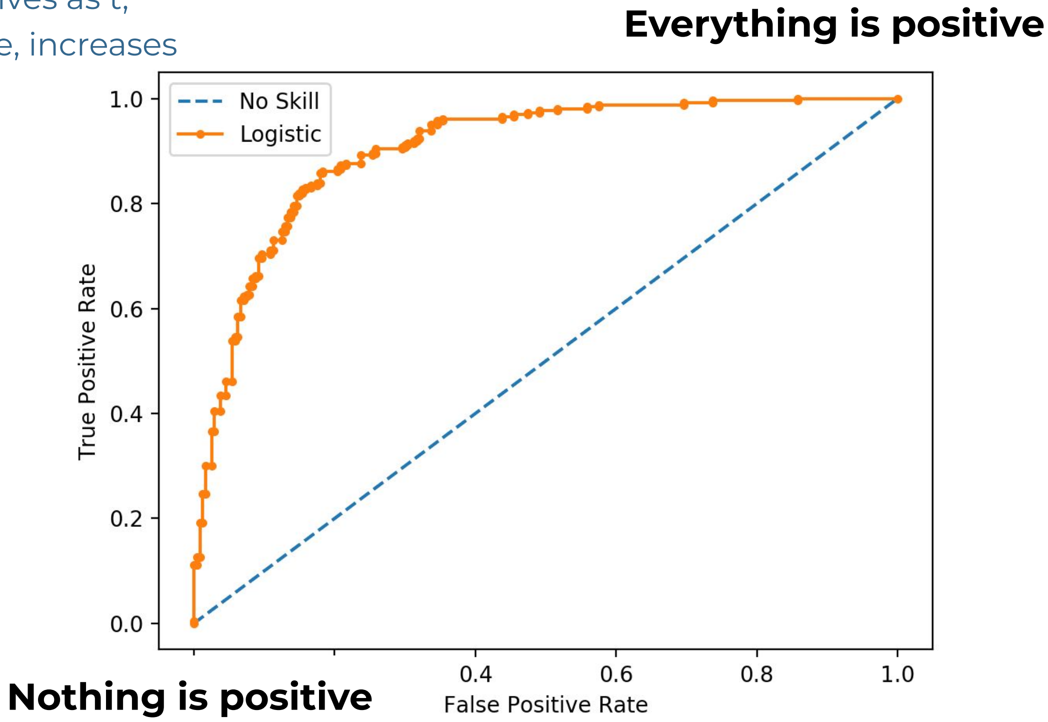
0 otherwise



ER
ARD
Upward.

ROC and AUC

- Receiver Operating Characteristic:
  - ROC(t) = (FPR, TPR)
  - Plot the true positives vs false positives as t, the threshold for predicting positive, increases
- AUC
  - Area Under the (ROC) Curve
- A random coin flip will have an AUC of .5
- An oracle will have an AUC of 1

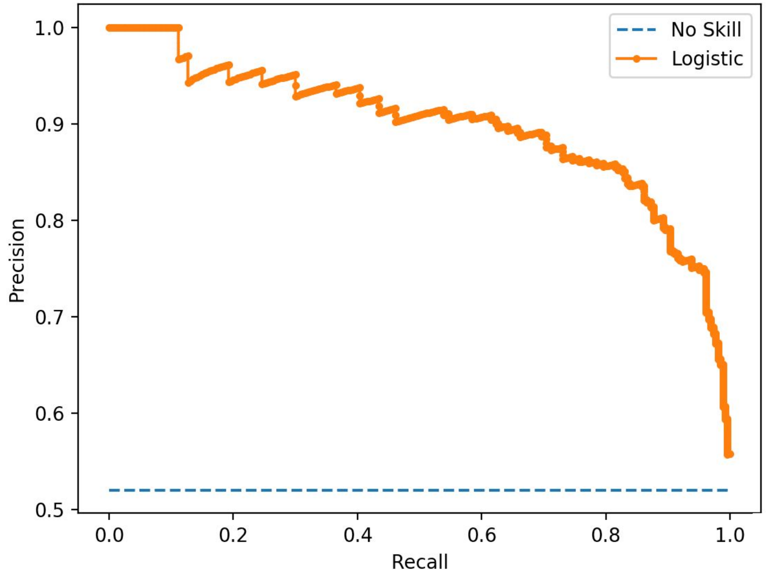


Precision vs Recall

- Precision: only give me positives and it's ok if I miss some
- Recall: I want all of them, even if it means I get some negatives
- F1 score = $\dfrac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

- If I rank my predictions, what is the precision at rank k? What is the recall at rank k?
- Mean Average Precision is the area under the Precision / Recall Curve
  - This is not the definition of MAP. This was proved by one of my thesis siblings

rank 1



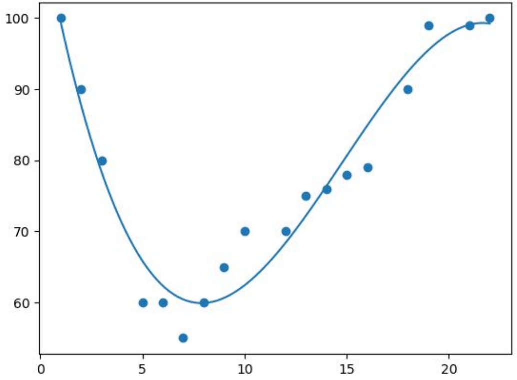
rank n

Recap

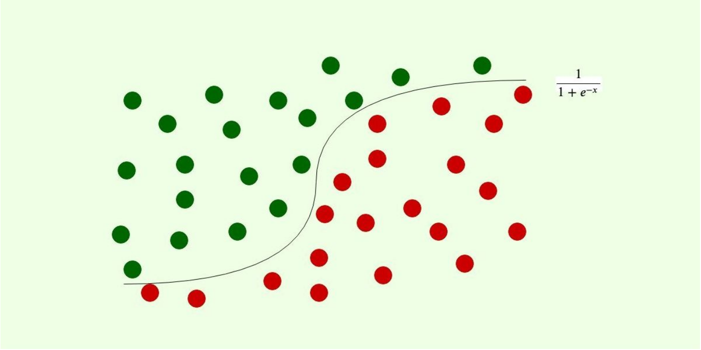
SUMMER
SPRINGBOARD
Look Inward. Go Upward.

What is machine learning?

- Machine learning is a form of statistical analysis for finding trends in data, on the assumption that it will generalize to new data
- Machine learning is the search for the "best fit" lines through points



[Python Machine Learning Polynomial Regression (w3schools.com)]



[Logistic regression and Keras for classification » AI Geek Programmer]

SUMMER SPRINGBOARD
Look Inward. Go Upward.

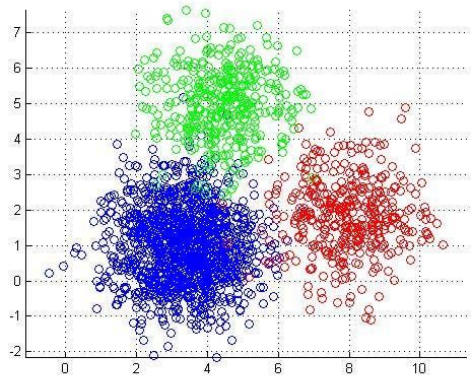
The goal

	Age	Height	Weight	…	Health Indicator
Person 1	37	5’7”	180		High
Person 2	33	5’4”	140		Low

- Learn a function f(age, height, weight,...) = Health Indicator
- Which allows me to ask the question f(New Person) = ?
- We call the columns 2 through n-1 *features* and the last column the *label*

Types of machine learning

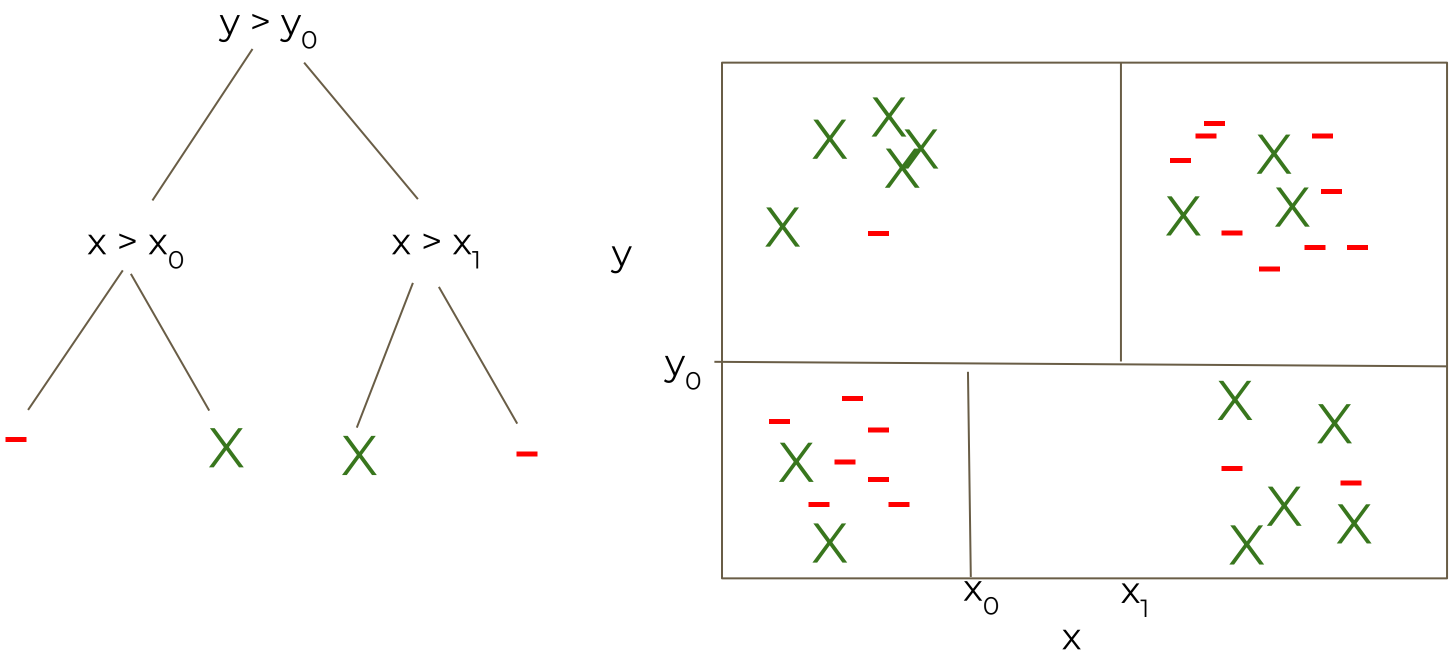
- Supervised: My dataset contains the labels
- Reinforcement: If I take an action, I can observe the label
  - Robotics
  - Ad targeting
- Unsupervised: My dataset contains no labels
  - Clustering
- Semi-Supervised: I am trying to predict multiple things, some with labels and some without



Supervised learning tasks

- There are primarily two supervised learning tasks
  a. Classification: Labels are *categorical*
     - apple / orange / banana
     - Disease / no disease
  b. Regression: Labels are *continuous*
     - Revenue
     - Energy consumption

Decision Trees



There are four possibilities

		Predicted	
		X	−
Truth	X	True Positive	False Positive
	−	False Negative	True Negative

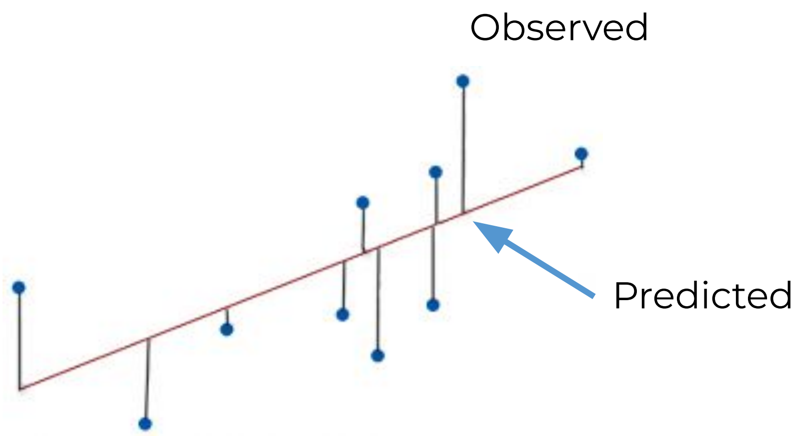
Pick one

Generalization

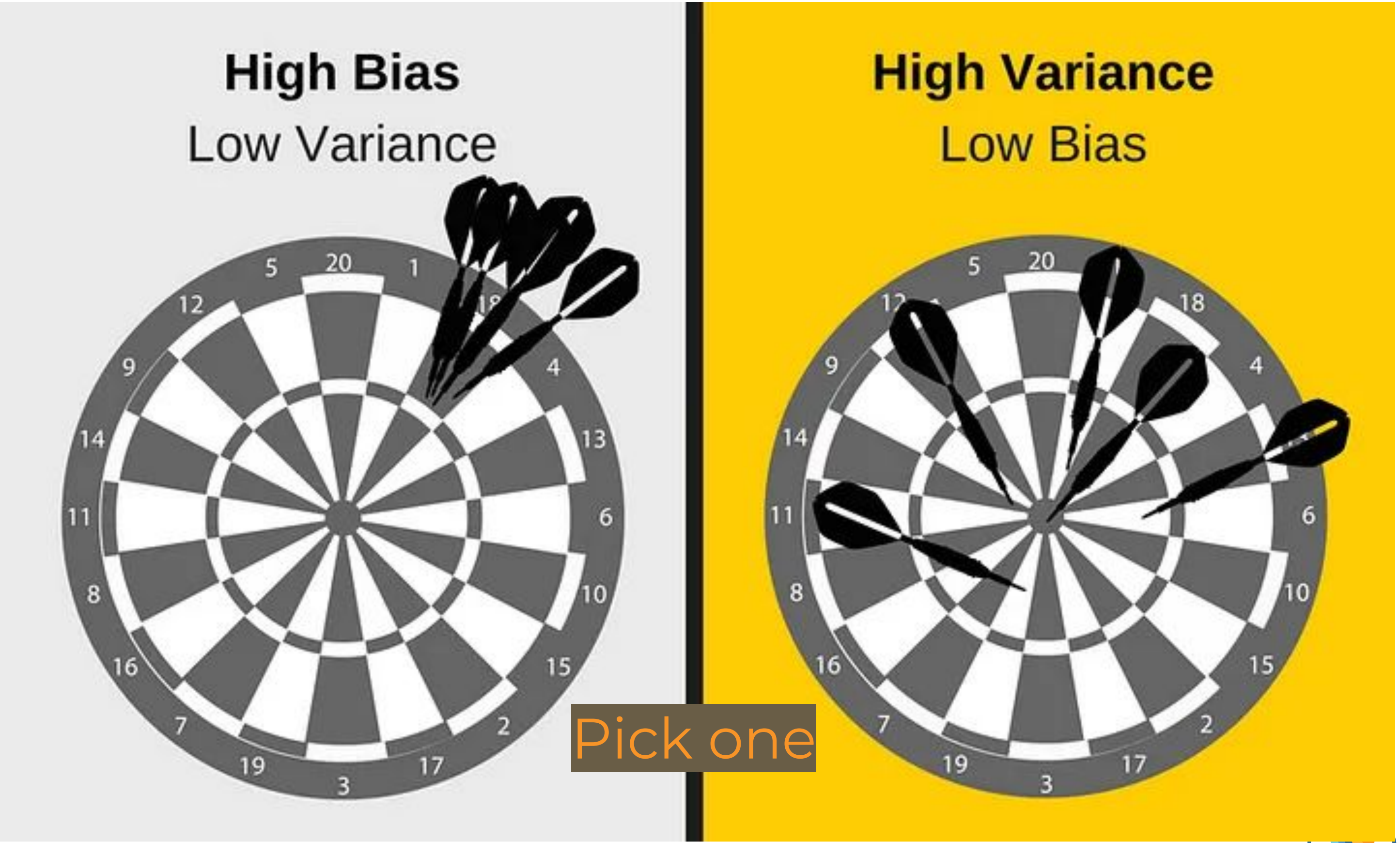
- The goal isn't to classify the data you have
  - This is labeled data! I already know the answer!
- The goal is to classify the data you *haven't collected yet*
- The problem: memorizing your training data
- The solution: evaluating on holdout data
- These are known as train / test splits (or train / test / validation splits)

Regression: Minimize RMSE (or similar)

- Error: Observed - Predicted
  - Technically this is a residual, but everyone calls it error
- Squared Error:
  - Sometimes error is negative. I want this number to always be positive
- Mean Squared Error:
  - I want the *average* error
- Root Mean Square Error
  - However, that's the average error *squared* so I have to take the square root



[Standard Error of the Regression vs. R-squared - Statistics By Jim]



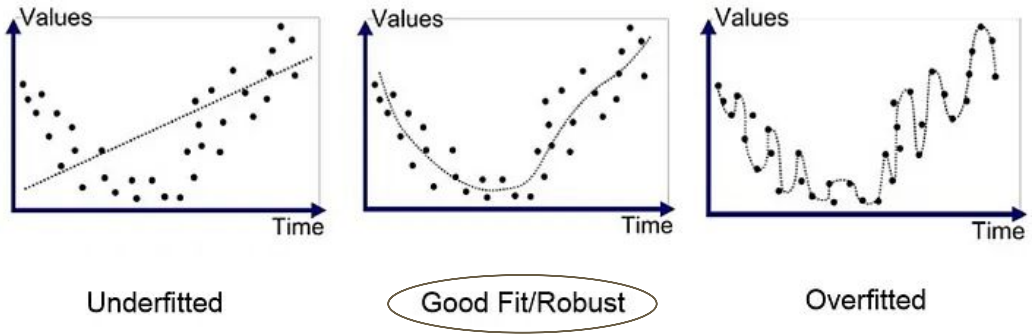
Pick one

[Bias-variance trade-off… I'm always confusing the two! | by Super Albert | The Making Of… a Data Scientist | Medium]

SUMMER SPRINGBOARD
Look Inward. Go Upward.

High bias / low variance

Low bias / high variance



SUMMER SPRINGBOARD
Look Inward. Go Upward.

[Machine Learning: How to Prevent Overfitting | by Ken Hoffman | The Startup | Medium]

Linear Regression

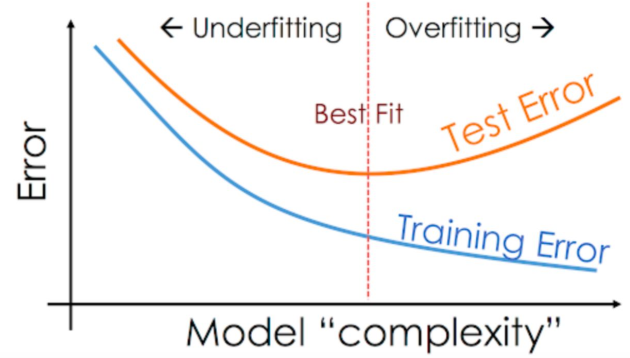
- A straight line $y = mx + b$
  - With a single input
- A straight line $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$
- The goal of linear regression is to find the vector of betas

$$\underset{B}{\operatorname{argmin}} ||y - BX||^2$$

- This can be found directly using linear algebra

Regularization

- You have to *earn* your complexity

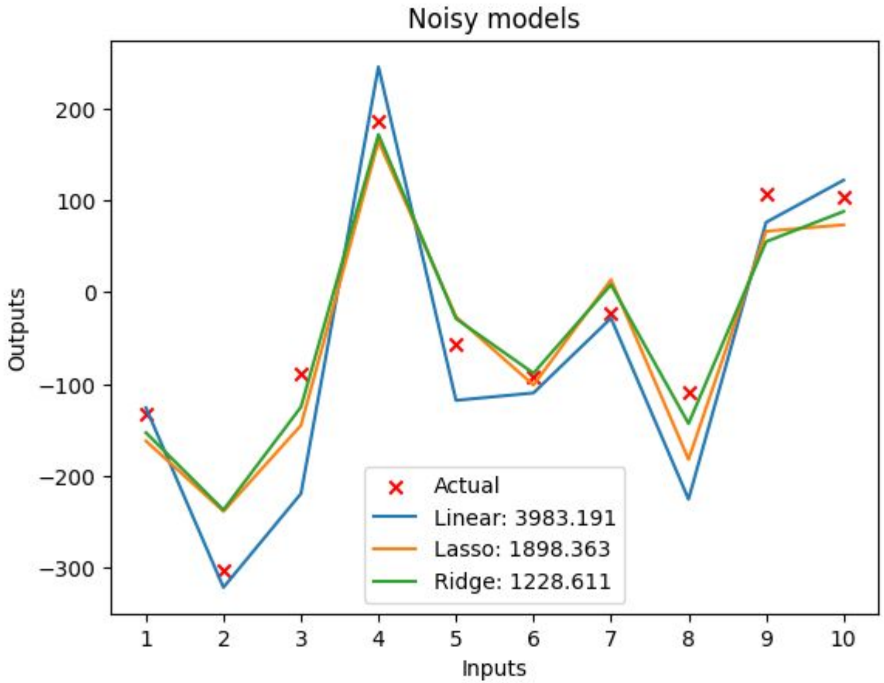


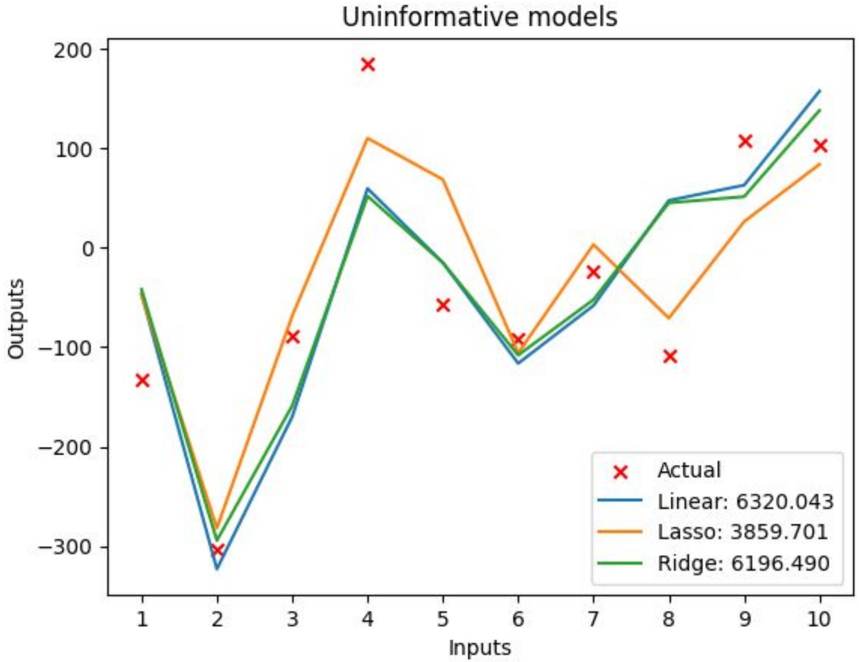
[Overfitting & Underfitting in Machine Learning - Data Analytics (vitalflux.com)]

SUMMER SPRINGBOARD
Look Inward. Go Upward.

Lasso and Ridge Regression

- The two most common forms of generalization for regression models
- Lasso: n = 1 $\underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \alpha \sum |\beta_i|$
- Ridge: n = 2 $\underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \alpha \sum \beta_i^2$
- Lasso forces some coefficients towards zero
  - This allows the model to ignore uninformative features
- Ridge forces all the coefficients to shrink
  - This makes the model more robust to outliers
- (I just googled why they are called that and it is entirely uninformative. I just look it up every time)





SUMMER SPRINGBOARD
Look Inward. Go Upward.

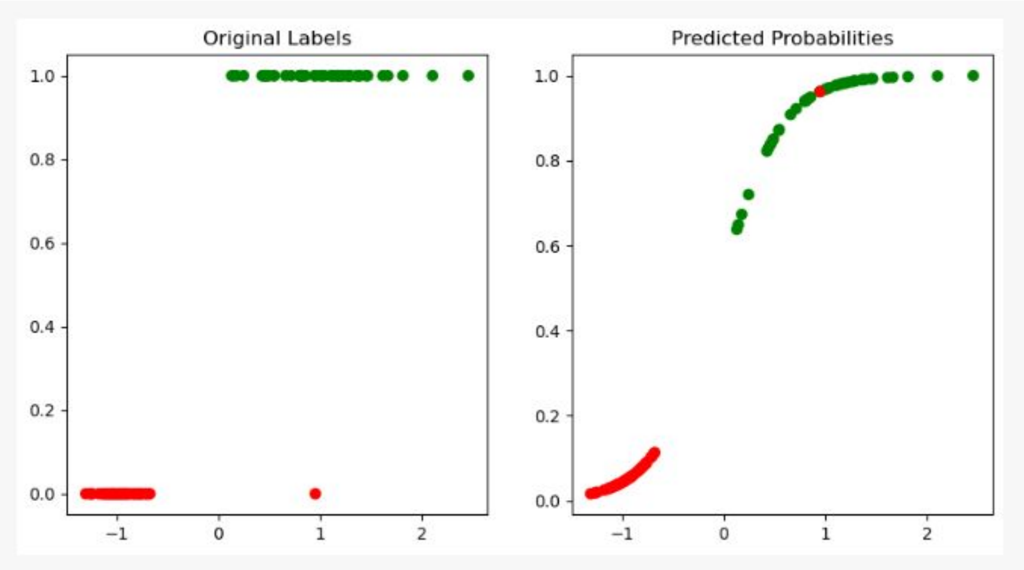
Logistic Regression

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Probability

Regression Coefficients

Predicted class:
1 if p(x) > t for some t
0 otherwise



What I learned meta-evaluating search engines

- Always make sure you can answer the following questions:
  - What *exactly* do I want to have happen?
  - How will I know if it worked?

- There's what's easy to measure and there's what's important. They are rarely the same thing.
- "When a measure becomes a target, it ceases to be a good measure"
  –Goodhart's Law
- "But if we measure just what's easy, we'll maximize just what's easy."
  [Column: You Are What You Measure (hbr.org)]

SUMMER SPRINGBOARD
Look Inward. Go Upward.