UNION COLLEGE

CSC321 Data Mining & Machine Learning

Prof. Nick Webb
webbn@union.edu

Linear models

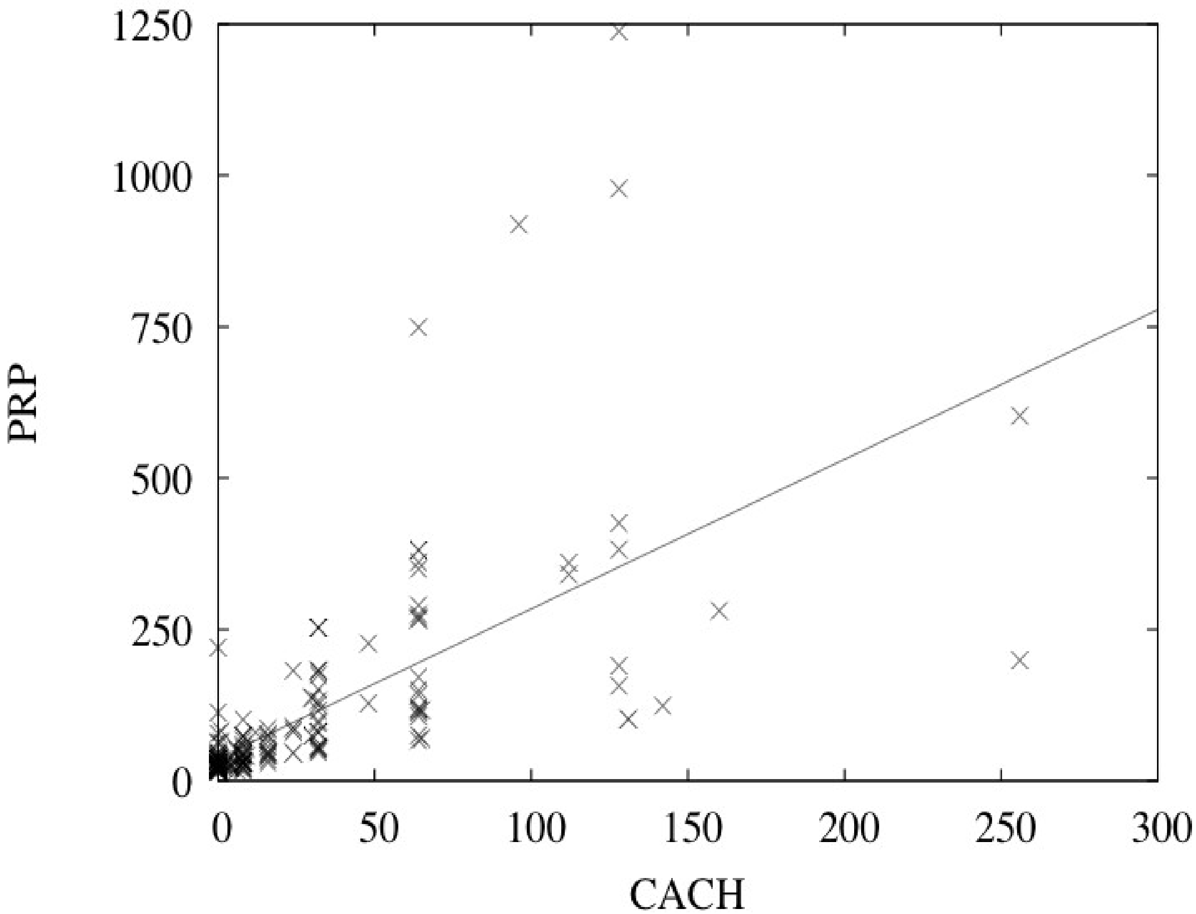
- A simple representation
- Regression model
  - Inputs (attribute values) and output are all numeric
- Output is the sum of weighted attribute values
  - The trick is to find good values for the weights

Simple ML methods

- Linear models for regression
- …and then for classification

- Regression -> where we are predicting a numeric value
- Linear regression -> a technique for doing that

A linear regression function for the CPU data



PRP = 37.06 + 2.47CACH

Simple Linear Regression

- More than 200 years old
- Assumes straight line (linear) relationship between input variable(s) and output variable
- Single input variable -> simple linear regression
- The line for a simple linear regression is:

$$y = b0 + b1 \times x$$

Simple Linear Regression

- Where b0 and b1 are the coefficients
- We learn the values of the coefficients from the data

- Once known we can use these coefficients to estimate output values y for given new input examples of x

Simple Linear Regression

- To calculate the coefficients we need:
  - Mean
  - Variance
  - Covariance

- MEAN: Average value of the numbers

Simple Linear Regression

- To calculate the coefficients we need:
  - Mean
  - Variance
  - Covariance

- VARIANCE: How spread out are the numbers
- Sum of the squared differences from the mean

UNION COLLEGE

Simple Linear Regression

- To calculate the coefficients we need:
  - Mean
  - Variance
  - Covariance

- VARIANCE: How spread out are the numbers

$$\text{variance} = \sum_{i=1}^{n} (x_i - mean(x))^2$$

Simple Linear Regression

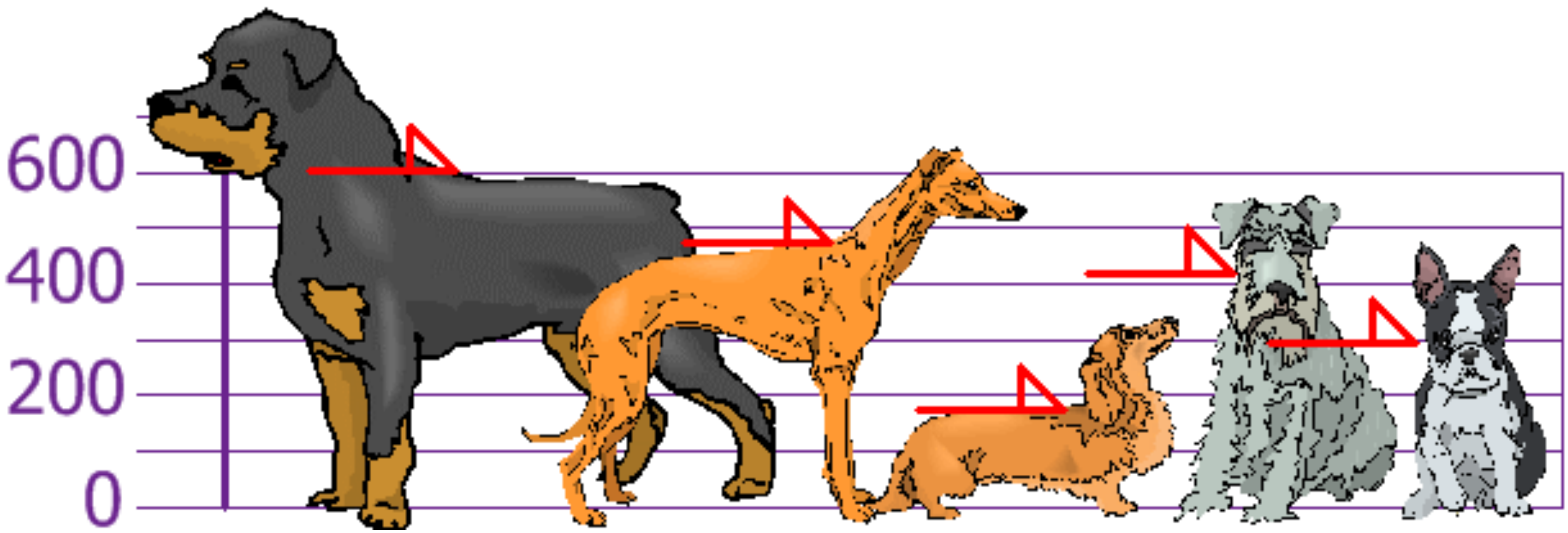
- To calculate the coefficients we need:
  - Mean
  - Variance
  - Covariance

- COVARIANCE: How numbers change together

$$\text{covariance} = \sum_{i=1}^{n} (x_i - mean(x)) \times (y_i - mean(y))$$

Variance & Standard Deviation

Heights are: 600mm, 470mm, 170mm, 430mm and 300mm

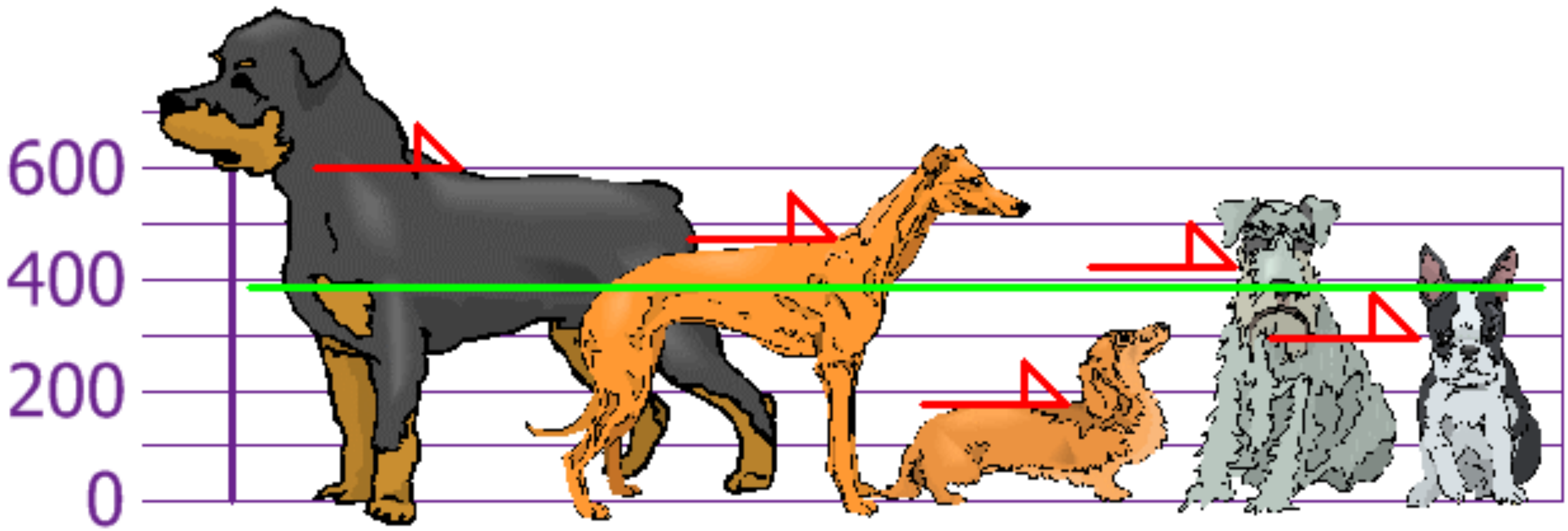


Variance & Standard Deviation

- First step, calculate the mean
  - Heights are: 600mm, 470mm, 170mm, 430mm and 300mm
- Answer
  - 394mm

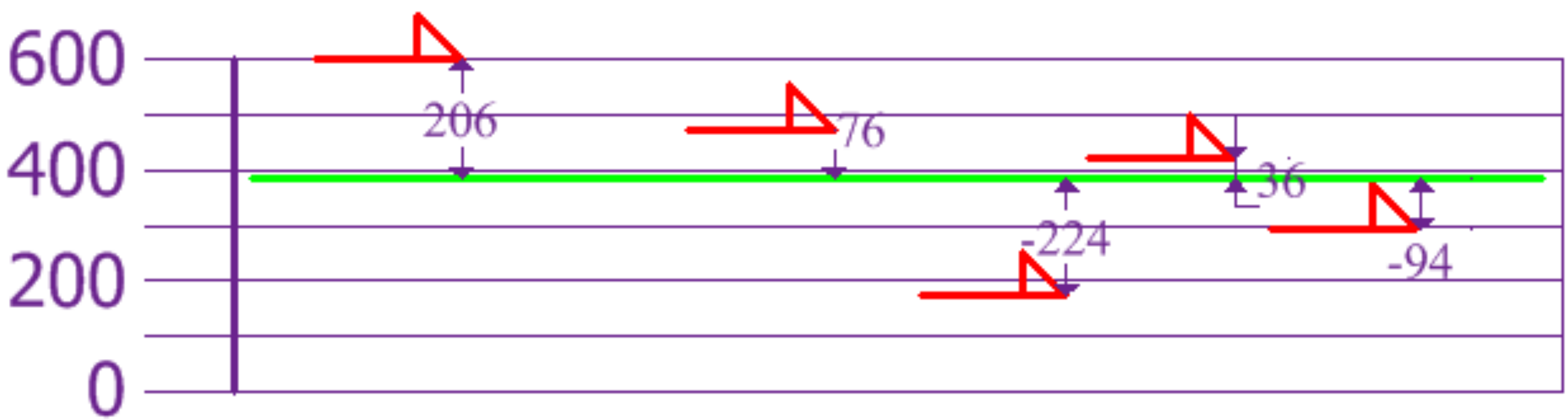
UNION COLLEGE

Variance & Standard Deviation



Now calculate each dog's difference from the mean

Variance & Standard Deviation



Variance & Standard Deviation

- To calculate average variance
  - Take each difference
  - Square it*
  - And average the result

- Variance is….
  - 21704

UNION COLLEGE

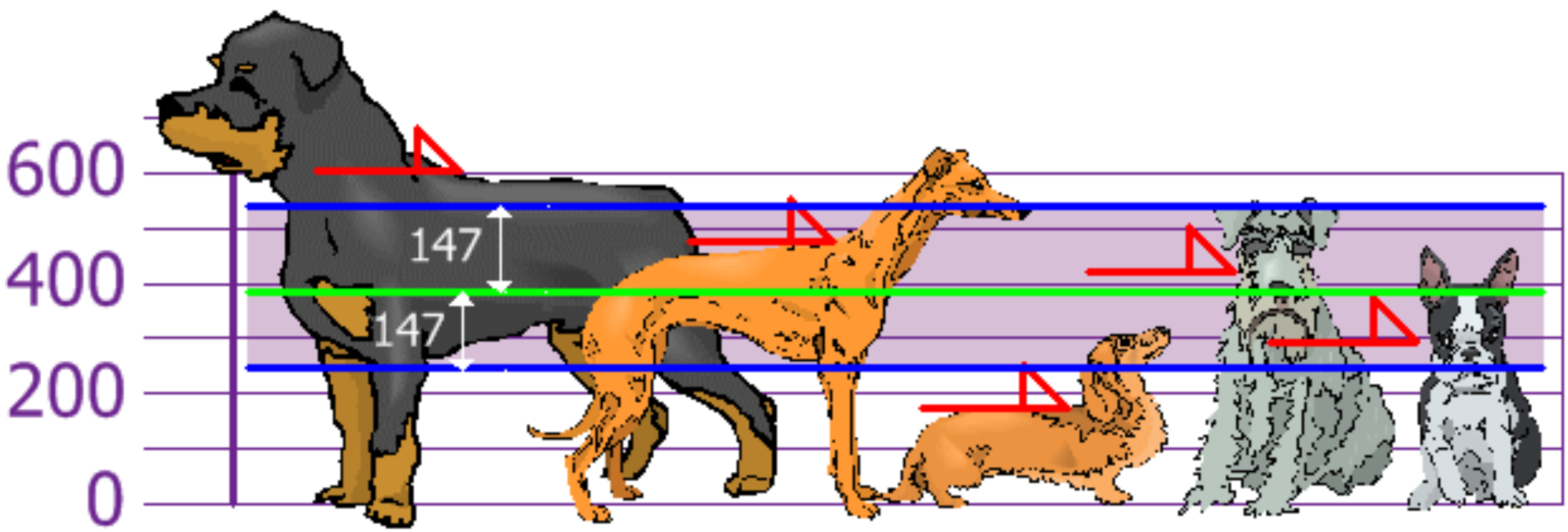
Variance & Standard Deviation

- To calculate standard deviation
- Take the square root of the average variance

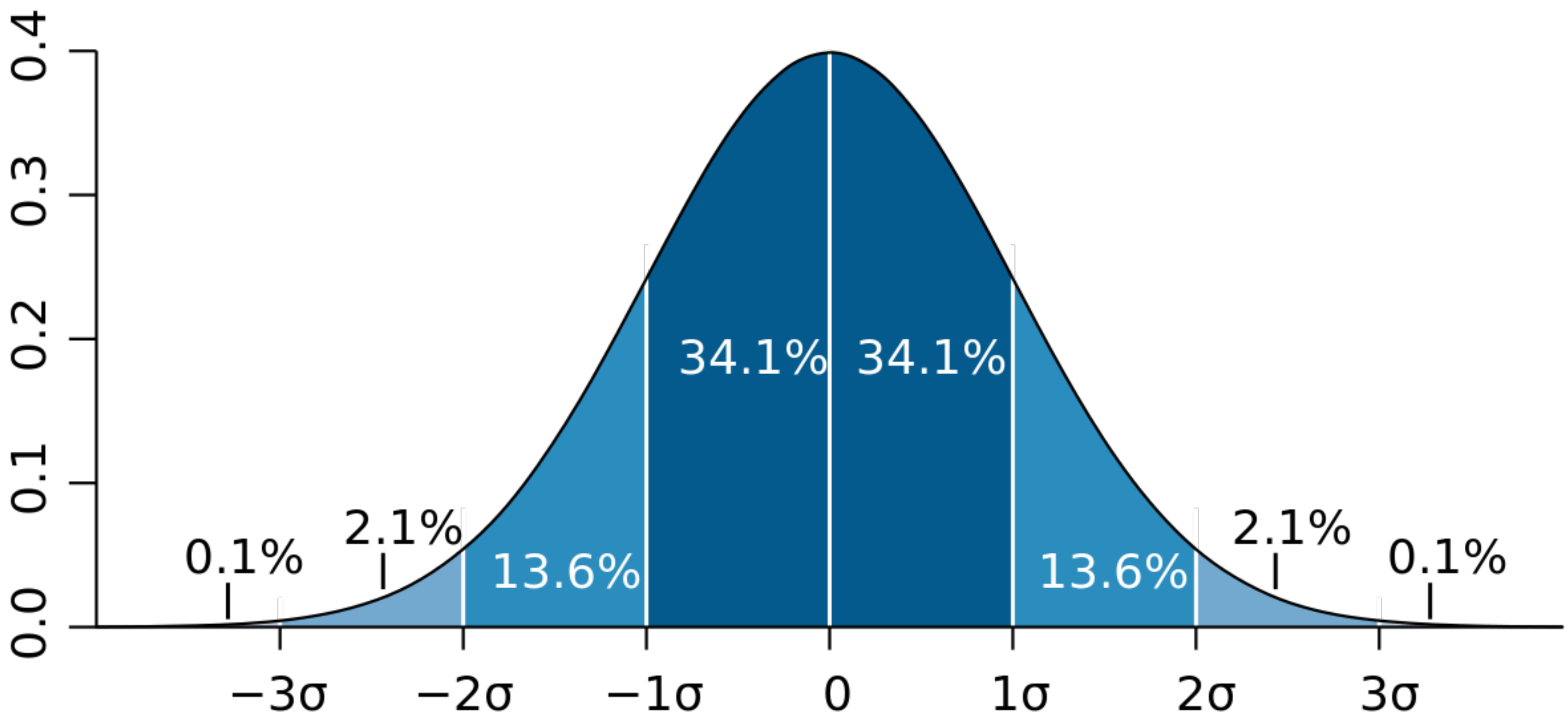
- Answer
  - 147.32mm

- Can now show which heights are within one standard deviation of the mean

Variance & Standard Deviation



Variance & Standard Deviation



Variance & Standard Deviation

- Standard deviation gives us a way of knowing what is normal, for our data
- And a way of identifying outliers
- There is a difference between a SAMPLE and a POPULATION
- Our example is a POPULATION
  - It contains all the instances we care about
- More generally we're dealing with a SAMPLE

Variance & Standard Deviation

- For a population, divide by N when calculating variance

- For sample, divide by N-1

- Consider it as a ‘correction’ when dealing with a sample

Variance & Standard Deviation

- *so why square the differences for variance?

- What if we just add the differences?
  - Negatives cancel out positives

- What about absolute values?
  - +4, +4, -4, -4 → ? Variance ?

Variance & Standard Deviation

- *so why square the differences for variance?

- What if we just add the differences?
  - Negatives cancel out positives

- What about absolute values?
  - +4, +4, -4, -4 → 4

Variance & Standard Deviation

- *so why square the differences for variance?

- What if we just add the differences?
  - Negatives cancel out positives

- What about absolute values?
  - +4, +4, -4, -4 → 4
  - +7, +1, -6, -2 → ? Variance ?

Variance & Standard Deviation

- *so why square the differences for variance?
- What if we just add the differences?
  - Negatives cancel out positives
- What about absolute values?
  - +4, +4, -4, -4 → 4
  - +7, +1, -6, -2 → 4

UNION COLLEGE

Simple linear regression

$$B1 = \frac{\sum_{i=1}^{n} (x_i - mean(x) \times (y_i - mean(y))}{\sum_{i=1}^{n} (x_i - mean(x))^2}$$

But because we were paying attention, we know this simplifies to:

$$B1 = \frac{\text{covariance}(x, y)}{\text{variance}(x)}$$

Simple Linear Regression

- We still need to estimate B0
  - Called the intercept
  - Controls where the start of the line is with respect to the y axis

$$B0 = mean(y) - B1 \times mean(x)$$

Simple Linear Regression

- Now we have the coefficients B0 and B1
- Apply them to input variables to predict output value

$$y = b0 + b1 \times x$$

- Which you'll be doing in your homework

Measuring Performance

- So we have a method of performing simple linear regression
- How well does it do?
- We need a way of measuring performance
- AND we need something to compare it to
  - Another simple machine learning algorithm

ZeroR

- For regression, use the mean of the output variable

- For classification, use the most frequently occurring class

Measuring Performance

- Calculate the size of error
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)

- MAE:
  - Sum the absolute differences from the correct score for each instance
  - Take the mean

Measuring Performance

- Calculate the size of error
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)

- RMSE:
  - Sum the square of the differences from the correct score for each instance
  - Take the square root of the mean

Measuring Performance

- Both metrics give the error in the same units as the input
- RMSE gives a larger penalty to larger errors

- In the homework, you will
  - Implement Simple Linear Regression
  - Implement ZeroR
  - Implement RMSE

UNION COLLEGE

Multivariate Linear Regression

- More than one input variable
- Still want to draw a ‘line’ between input and output
- With more dimensions, this becomes a plane (often called a hyperplane)
- Just as with simple linear regression, each input gets a weighting coefficient
- Goal of learning is to discover values for those coefficients

UNION COLLEGE

Multivariate Linear Regression

$$y = b0 + b1 \times x1 + b2 \times x2 + \ldots$$

- For simple linear regression we could simply 'read' the coefficients off the data
- For multivariate linear regression, we have a much bigger search space
- We need to estimate the coefficients

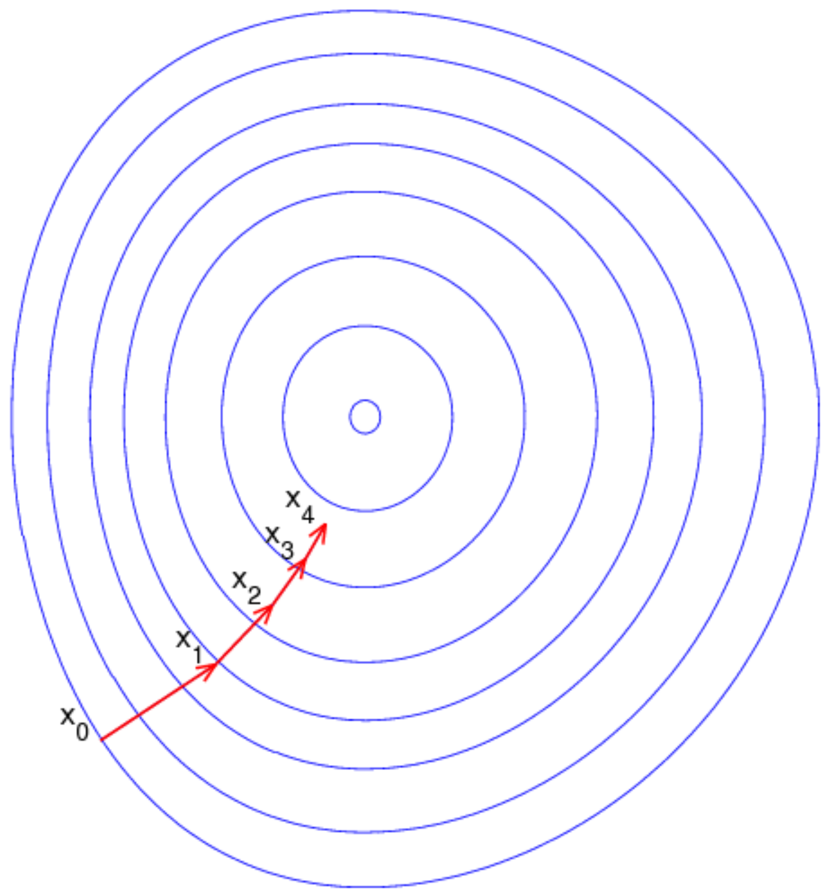
Stochastic Gradient Descent

- Popular optimization strategy
- Used in all kinds of machine learning
  - Including deep learning
- An algorithm that tweaks coefficients to find optimal values for a function
- A gradient is the SLOPE of a function
  - Higher gradient = steeper slope = faster learning
  - Zero slope = an end to learning

Stochastic Gradient Descent

- Imagine wearing a blindfold
- Trying to find the top of a mountain
- In the fewest steps possible
- Start climbing the mountain, taking big steps in the steepest direction
- As you think you're getting close to the top, take smaller steps so as not to overshoot the peak

Stochastic Gradient Descent



Stochastic Gradient Descent

- The reverse of our example:
  - We are racing to the bottom of a valley

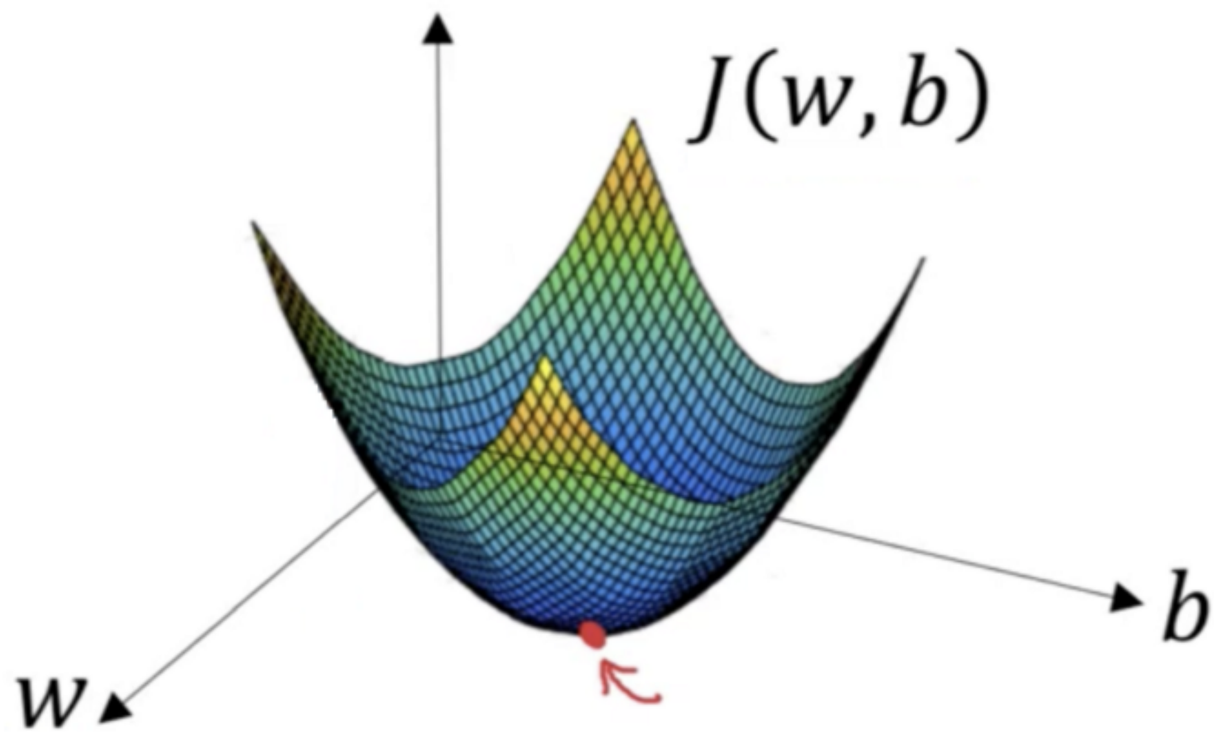
$$\mathbf{b} = \mathbf{a} - \gamma \nabla \mathbf{f}(\mathbf{a})$$

  - b is next position
  - a is previous position
  - gamma is the learning rate
  - gradient term is direction of steepest descent

Stochastic Gradient Descent

- Imagine a function with some parameters
  - J(w,b)
- Want to reach the optimal version of that function (it's minimum)
- By tweaking parameters w and b
- Shown as a red arrow on the following graph

Stochastic Gradient Descent



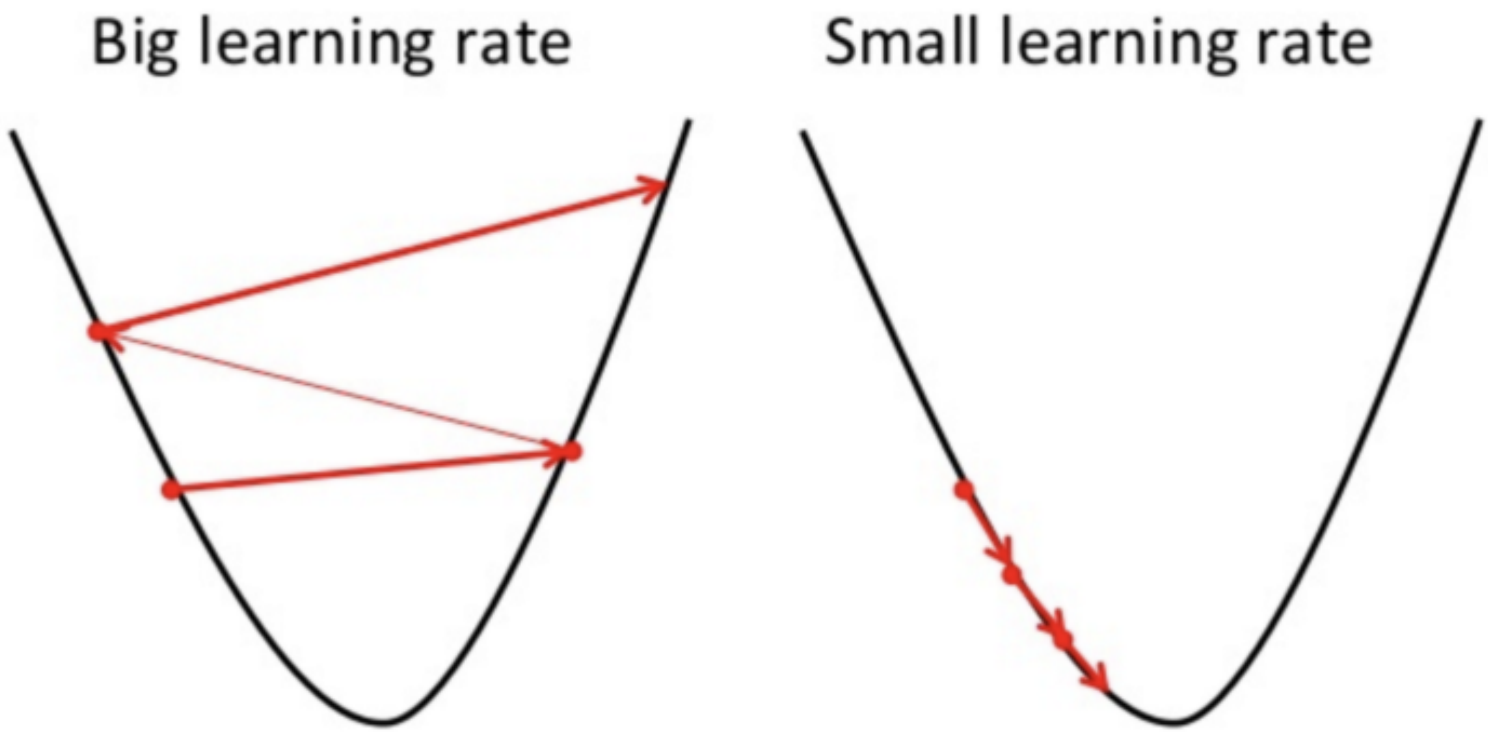
Stochastic Gradient Descent

- Initialize w and b to some values
- Stochastic gradient descent starts at that point
- Takes one step after another in the steepest downward direction
- Until it reaches the point when the function is as small as possible
- In our case, the function here is ERROR over the training data

Stochastic Gradient Descent

- The importance of learning rate
  - The size of the steps the algorithm takes in the direction of the minimum
  - Must be neither too low
  - OR too high
  - Welcome to the world of Goldilocks in machine learning

Stochastic Gradient Descent



Stochastic Gradient Descent

- Too big
  - Faster learning
  - But can bounce back and forward
  - Never reaching the minimum

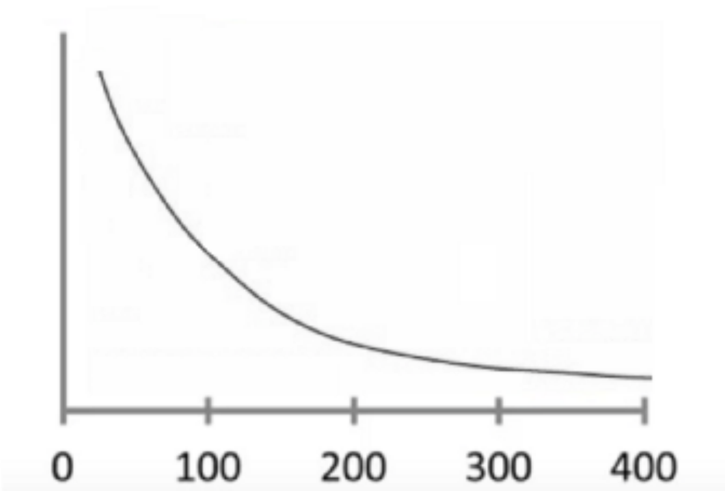
- Too small
  - Will reach the minimum
  - Can take a very long time

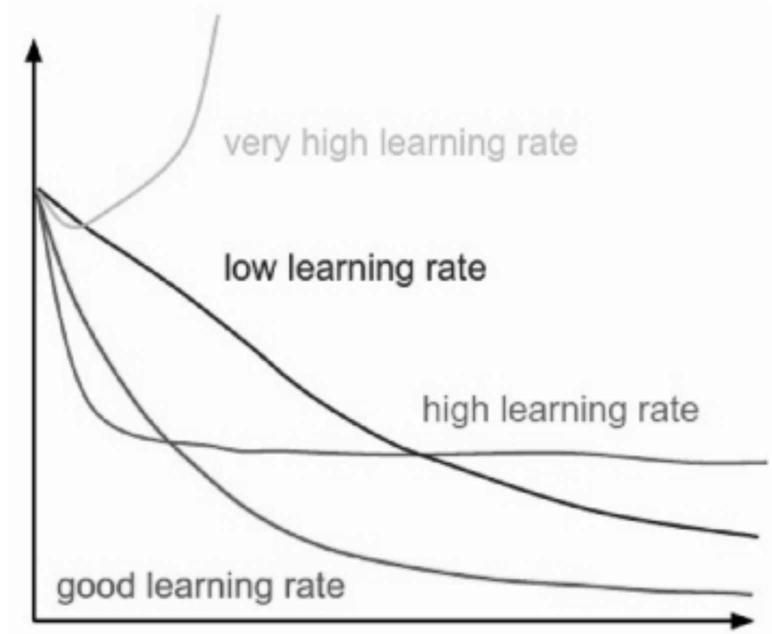
Stochastic Gradient Descent

- It's possible to plot the learning rate on a graph
- Iterations on the x-axes
- Value of cost function on y-axes

- If SGD is working properly, the cost should decrease after each iteration
- BUT we also don't know how many iterations it will take

Stochastic Gradient Descent





Stochastic Gradient Descent

- So we have two NEW parameters which we have to work with
  - Learning rate
    - 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1…
  - Iterations (called epochs)

- In each iteration, we're going to adjust each coefficient on each input variable by the learning rate, until we reach a minimum (covergence)

Multivariate Linear Regression

- Algorithm
  - Loop over each epoch
    - Loop over each row of the training data for an epoch
      - Loop over each coefficient and update it for a row in an epoch

- Where

  b = b – learning rate * error * x

Multivariate Linear Regression

$$y = b0 + b1 \times x1 + b2 \times x2 + \ldots$$

- We're going to use SGD to estimate each of our coefficients:
  - b0, b1, b2…

Working with real data

- We’re mostly going to work with csv files
  - Not all csv obey the c part
- We’ll need to load csv files
  - Can use csv module in python
- We’ll need to change data types
  - We’re going to get data as strings, and we don’t want that
- We’ll need to normalize

Scaling our data

- Most of the time we want data scaled appropriately

- Normalization
  – Rescale data in the range (0,1)
  – Requires that we know the minimum, maximum for our data

Normalization

- Once we have minimum and maximum values
- Rescale each value in the data

$$\text{scaled value} = \frac{value - \min}{\max - \min}$$

Scaling our data

- Can also center the data around 0, and fix max and min at 1 standard deviation
- Requires data to have a normal distribution
- Normalization does not have that requirement

- Should record max,min in case you need to scale any future data