Statistics

**Lesson 0**

Distributions

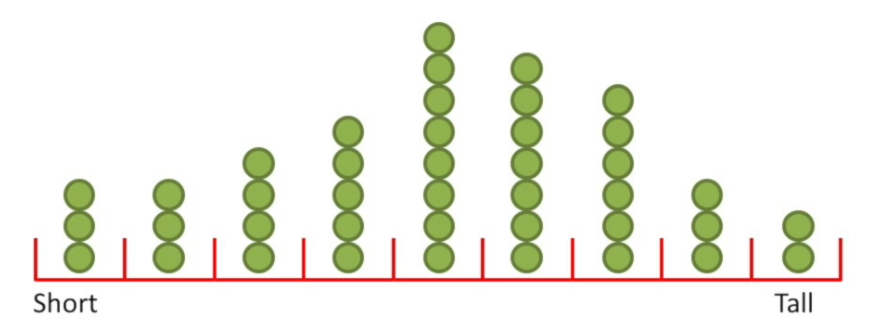
This is the process of splitting statistical data into “bins” so that they fit onto a histogram. The histogram makes it easier to measure and make insights

A graph showing the measurements of many people

A close up of a logo

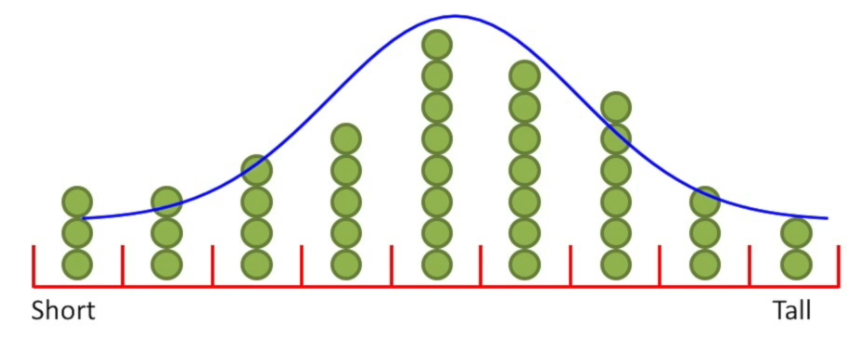
Description automatically generated

Distribution type 1 - Histogram



Figuring out how wide to make each “bin” is important. If the bin is too wide or too narrow, the data may not be equally distributed.

Distribution type 2 – Normal/Gaussian distribution curve



Advantages

* It is not limited by the width of the bins
* Helps calculate the probability for measurements of a specific size
* With a small sample size, the standard deviation + mean can help predict future values

**Lesson 1**

Normal Distribution (The central limit theorem)

μ is mean

σ is a standard deviation

π is 3.14159

e is 2.71828

A screenshot of a cell phone

Description automatically generated

x-axis: relative probability of observations

y-axis: relative probability/frequency of observations

The width of a curve is defined by the standard deviation.

Machine Learning

**Lesson 0**

Machine Learning

Application of AI that provides system the ability to automatically learn from experience without being explicitly programmed.

Machine Learning = Construct a Hypothesis/Model

Machine

Learning

Reinforcement

Learning

Supervised

Learning

Unsupervised

Learning

Classification

Regression

Logistic Regression

Linear Regression

Predict a categorical dependant variable based on values of independent variables

Predicts binary value (0 or 1)

Predict a continuous dependant variable based on values of independent variables

Predicts integer value

**Lesson 1**

Training models

First, let’s understand what overfitting is.

Overfitting

* happens when your model matches the dataset too closely
* or when your model fits the Training Dataset very well but not the Test Dataset.
* This happens because your model is trying too hard to capture the noise (random chance) in your training dataset
* To avoid overfitting, we can use regularization

This is a realistic view of the sizes of models on a very large amount of data:

Training Dataset (60%)

Contains outcomes to train a machine

Multiple Prediction Algorithms are applied

Adjusts weight in neural network

Cross Validation Dataset (20%)

compare the performances of the prediction algorithms based on training set (Using Cost Functions)

Aim to minimize over fitting

Original dataset

Test Dataset (20%)

Apply chosen algorithm on real-world data

Training data vs. Testing data

* As a rule of thumb, a dataset (training data) should be 10x it’s dimension (testing data) and independent of the model used
* This is also to say, throwing more data at a problem won’t always get better results

Bias describes the inability for a machine learning method to capture the “true” relationship.

Variance describes the difference in fits between fitting a machine learning method to the Training Dataset vs the Testing Dataset.

**Lesson 2**

Normalisation

(<https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029>)

Normalization is a technique often applied as part of data preparation for machine learning.

The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

**Lesson 3**

Linear Regression

(graphs drawn using <https://www.desmos.com/calculator>)

Linear Regression is a topic in statistics that is used as a predictive analysis tool.

Regression can be non-linear.

It aims to form relationships between data points and:

* determining the strength of predictors,
* forecasting an effect,
* trend forecasting.

Linear Regression with h(x)

Machine Learning = Construct a Hypothesis/Model

* This topic recycles concepts learned from High-school mathematics and rebrands concepts using terms found in data science.
* Our goal with h(x) is to choose parameters so that our input (x) can predict the corresponding output (y) for our training set.

Number of plot points

Number of samples in dataset

(Input)

Parameters

Hypothesis/Model

(Output)

\* A model is a dataset that represents reality

Steps:

1. Chose a regression model, apply it to dataset
2. Calculate (degree to which the data is explained by the model)
3. Calculate a p-value for (indicates if there is a significant relationship described by the model)

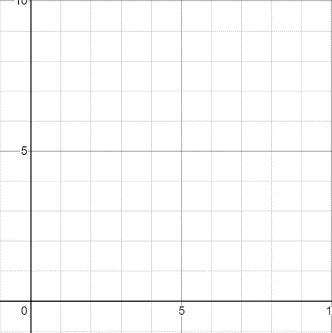
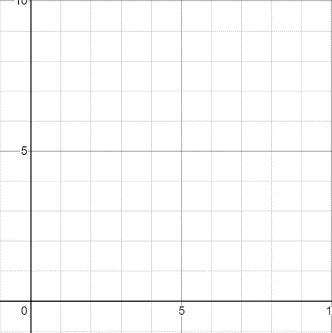
0 - Linear Regression Analysis

x axis : Independent variable

y axis : Dependent variable

graph 1 : positive relationship (m)

graph 2 : positive relationship (-m)



1 - Least-squares regression (line of best fit)

This aims to place a line across as many datapoints as possible, by using formulae to calculate the gradient and intersect with the dataset.

Mathematically, calculating distance from the line to each datapoint is called the residual. The sum of the square of these distances is . tells us how much variation of the dependant variable (y) can be explained by taking the dependant variable (x) into account.

= 1, perfect fit

= 0, no relationship



Higher

Lower

To calculate least squares, we calculate the gradient and intersect of the line of best fit using formulae. The Hypothesis model gives us the full equation for our line of best fit.

Hypothesis

Gradient

Intersect

A picture containing cabinet, electronics, furniture

Description automatically generatedExample

Dataset:

x = [1, 1, 2, 3, 4, 3, 4, 6, 4]

y = [2, 1, 0.5, 1, 3, 3, 2, 5, 4]



1. least-squares

2. Calculate

method 1:

method 2:

Notes:

1. or denotes the “sum of squared residuals”
2. or denotes the “sum of squared residuals around the mean”
3. or denotes the “variance around the mean”
4. or denotes the “variance around the fit or residuals”

3. Calculate p-value for

* probability that some random chance generated the data
* or something else that is equal
* or rarer

**Lesson 4**

T-Tests & ANOVA/Analysis of Variance (Using a design matrix)

T-Tests

The goal of a t-test is to compare means and see if they are significantly different from each other.

Steps:

1. Ignore the x-axis and find the overall mean
2. Calculate “sum of squared residuals around the mean”
3. Fit a line to the data (either least squares or the mean)

ANOVA (Analysis of Variance)

The goal of ANOVA is to analyse the differences among group means in a sample

**Lesson 5**

Logistic Regression

Logistic regression can be thought of as a special case of linear regression when the outcome variable is categorical.

Logistic Regression Models

(where p is the probability of

the hypothesis being true).

How is p derived/simplified?

(where is )

<https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>

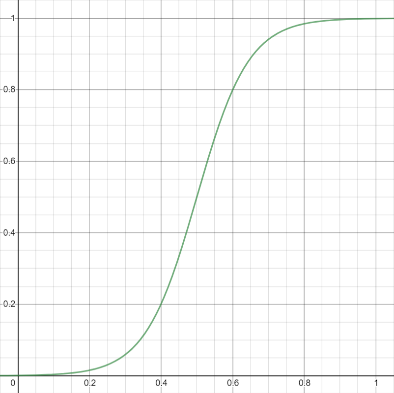
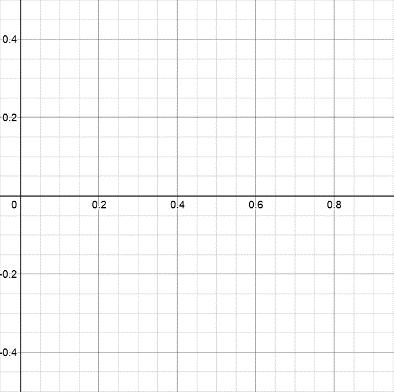
0 – Binary Logistic Regression

This method aims to predict whether something is true or false, instead of predicting something continuous (like price, if it was a dependant variable).

1 – Coefficients (using a continuous variable like weight to predict obesity)

1 true

+ infinity



0 false

- infinity

“Binary logistic model” has datapoints with a probability attached, where if 50% or more, it is categorized as true.

“log odds of y” represents a line which fits the datapoints

At y = 1, true

At y = 0, false

“Log Odds” graph has datapoints only in at positive or negative infinity.

“line of best fit” represents a line which fits the datapoints.

At y = + infinity, true

At y = - infinity, false

The coefficients are represented here by the y-intercept and the slope.

This method cannot use residuals to calculate a line of best fit nor can it calculate .

Line of best fit – Regression

Unfortunately, there is no consensus on how to calculate for logistic regression. There are more than 10 methods used today. Different researchers have proposed different measures for logistic regression. Different fields may have “best practices” regarding .

Maximum Likelihood

This calculates Logistic Regression, where the highest Maximum likelihood is used. It aims to inherit properties of from linear regression.

**Lesson 6**

Clustering – K-means

A screenshot of a cell phone

Description automatically generated

Discrete outcomes – containerized outcomes (i.e. true or false)

Continuous outcomes – time-series data (i.e. numerical intervals, infinite)

**Lesson 7**

Other types of Regression

2 – Multiple Linear Regression

This aims to add a plane or higher dimensional object to the data. This just means adding additional data to the model.

3 - Regression with Regularization – How to increase accuracy of learning model

Regularization

* This technique is used for tuning the function by adding an additional penalty term in the error function.
* this technique discourages learning a more complex or flexible model, as to avoid the risk of overfitting.

4 – Isotonic Regression – Equal Stretch Regression

This is a free-form linear model. The model is unique because it is constantly skewed one way or another. It is a complex model which can be found in python

**Lesson 8**

Cost Functions – How to measure the precision of my algorithm

* This function minimizes parameters over the dataset.
* It measures the performance of a Machine Learning model for given data.
* It quantifies the error between predicted values and expected values and presents it in the form of a single real number.

NB: Cost functions are done on the Testing Dataset.

Here are examples of popular cost functions:

Mean Absolute Error (MAE) measures the difference between the estimator (the dataset) and the estimated value (the prediction).

It calculates the average squared difference between the predictions and expected results.

MAE doesn’t add any additional weight to the distance between points — the error growth is linear.

i - index of sample,

ŷ - predicted value,

y - expected value,

m - number of samples in dataset.

i - index of sample,

ŷ - predicted value,

y - expected value,

m - number of samples in dataset.

Mean Squared Error (MSE) measures the difference between the estimator (the dataset) and the estimated value (the prediction).

It calculates the average squared difference between the predictions and expected results.

MSE errors grow exponentially with larger values of distance. It’s a metric that adds a massive penalty to points which are far away and a minimal penalty for points which are close to the expected result

A picture containing sky, map, table, indoor

Description automatically generated

Example - MSE



= estimated value. From using line of best fit

Example – MAE



**Lesson 9**

Optimization

Note: Sum of Squared Residuals or “SS” functions (like or ) are a type of Cost/Loss Function

Gradient Descent

1. Set (typically 0.001)
2. Set and to random numbers (typically 0)
3. Set (typically 1000)
4. Calculate MSE (or )
5. Calculate gradient for MSE ( and )
6. Adjust and using gradient and learning rate
7. If current MSE > MSE, then Set
8. Repeat steps 4 – 7 for the number of iterations set



Replace the intercepts as you shift the line down

Replace this with equation for line

Cost Function: MSE (Mean Squared Error) (also )

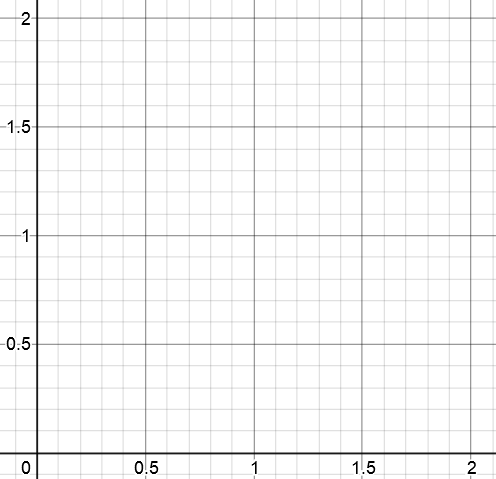
Calculate gradient for cost function:

Partial derivative with respect to b

Partial derivative with respect to m

The derivative for the cost function can determine the slope at any value for the intercept

The derivative is used for the cost function as when the slope is close to 0 (bottom of the curve), we know to stop processing. In practice, the number of steps chosen is up to you. Find a number that is efficient and can get to the answer quicker.



Each time the gradient cost function is calculated, “step” is taken in one direction and a point is plotted on the graph. The size of “step” is calculated using a learning rate.

The learning rate adjusts

Intercept

Learning Rate:

TODO

* How many correlations/Cost functions should I use to measure accuracy?
* How do I increase accuracy of my model? >> regularization?
* Confusion matrix, research this

TODO (from Hailey)

* Clustering (classification)
* Random forest (classifier + regressor)
* XG Boost
* PCA (Principle Component Analysis)

https://stackoverflow.com/questions/47577168/how-can-i-increase-the-accuracy-of-my-linear-regression-modelmachine-learning

* Add [interaction terms](https://en.wikipedia.org/wiki/Interaction_(statistics)#In_regression) to model how two or more independent variables together impact the target variable
* Add [polynomial terms](https://en.wikipedia.org/wiki/Polynomial_regression) to model the nonlinear relationship between an independent variable and the target variable
* Add [spines](http://people.stat.sfu.ca/~cschwarz/Consulting/Trinity/Phase2/TrinityWorkshop/Workshop-handouts/TW-04-Intro-splines.pdf) to approximate piecewise linear models
* Fit [isotonic regression](https://en.wikipedia.org/wiki/Isotonic_regression) to remove any assumption of the target function form
* Fit non-parametric models, such as [MARS](https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines)

What does any of this mean? ☹