Genetic Algorithms

Winning entry for the Forecasting & Futurism 3rd Annual iPad Contest

Copyright 2013 by Jeff Heaton (reused material from several books)

Jeff Heaton

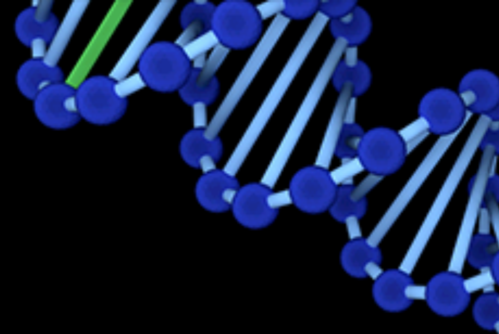
- EHR Informatics Scientist
- Blogger and Author
- EMail: jeff@jeffheaton.com
- GitHub: https://github.com/jeffheaton
- Twitter: @jeffheaton

Problem Definition

My contest entry made use of a Genetic Algorithm to devise a Radial Basis Function (RBF) Network to predict the malignancy of breast tumors. This determination is made using the following nine attributes collected from a tumor.

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

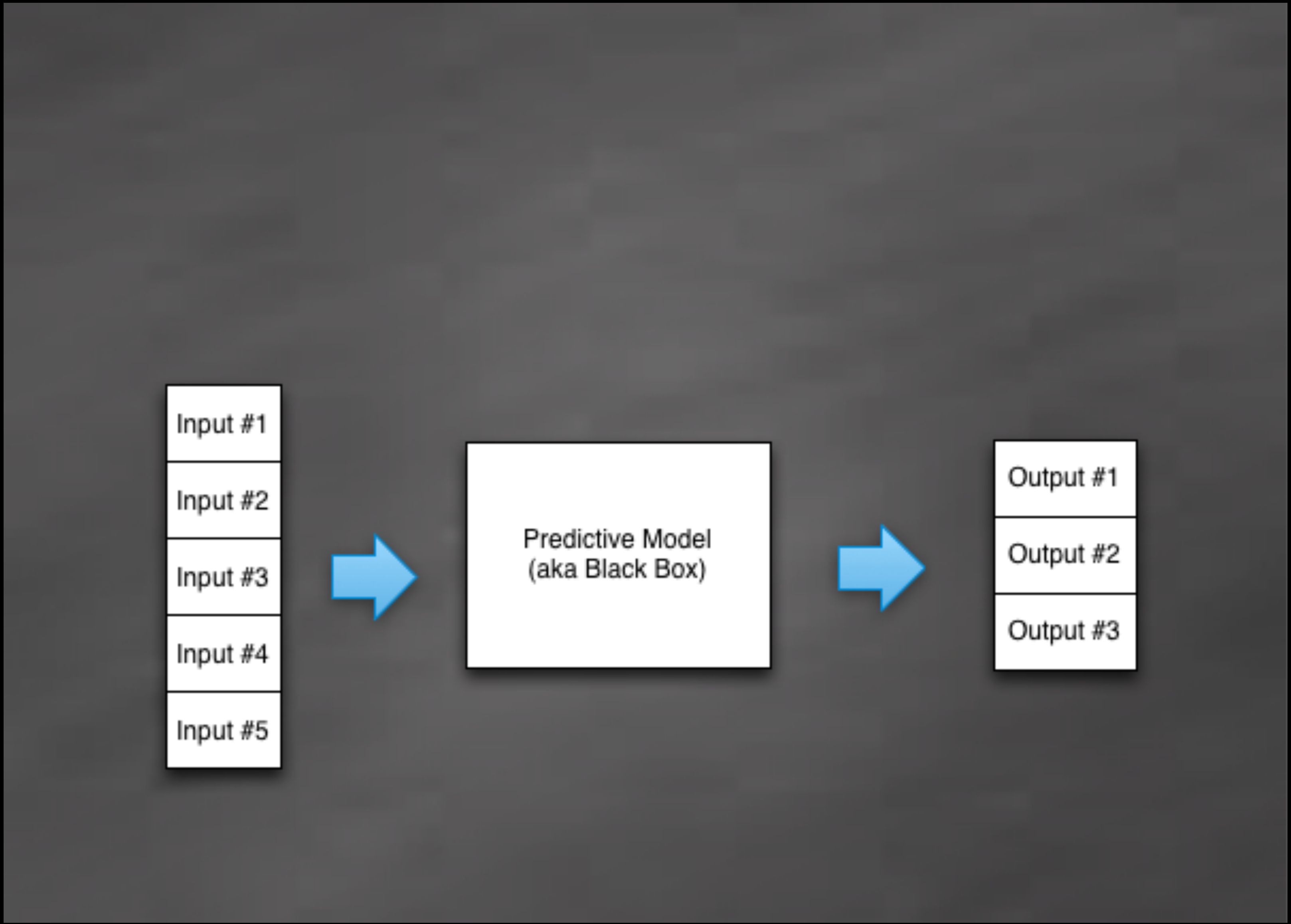
Data Used

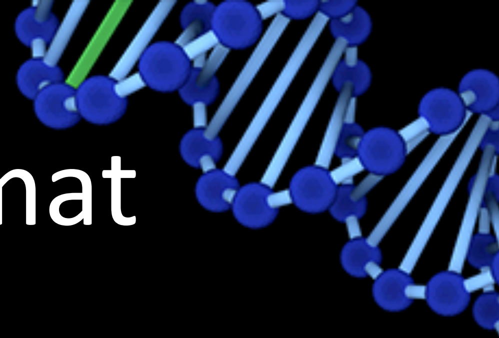


I used training obtained by Dr. William H. Wolberg, University of Wisconsin Hospitals . Using these attributes I constructed a RBF network to perform classification. The two classes were either malignant or benign.

- Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.
- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

What is Machine Learning?

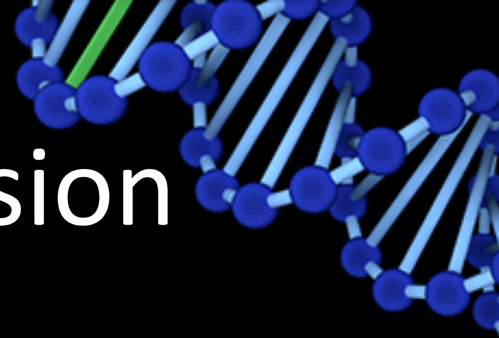




Input and Output Format

Input and output are both vectors (arrays).

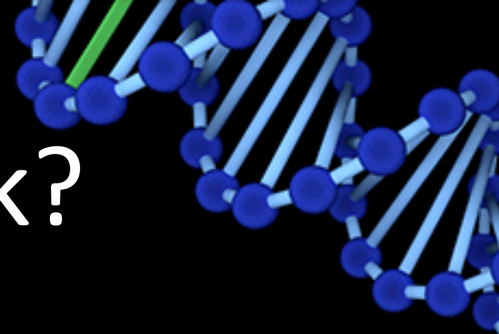
- Inputs: [0.1, 0.3, -2, 0.4, 0.3]
- Outputs: [0.5, 0.1]



Classification & Regression

Both classification and regression models accept input data (features) and produce an output.

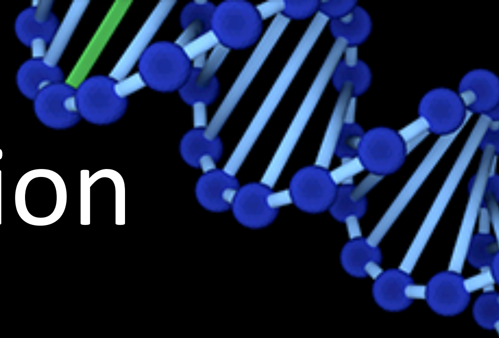
- Classification Models output a class that the input data belongs to (i.e. given tumor features is it benign or malignant)
- Regression Models output a number (i.e. given a car's features, what is the MPG)



How do Models Work?

Models are usually some form of mathematical equation. Models have coefficients that are adjusted to fit the models output to training data. Two examples of very simple models:

- Linear Regression models find linear relationship between inputs and single output.
- Polynomial Models fit the model data to a polynomial. Also a single output.



Temperature Conversion

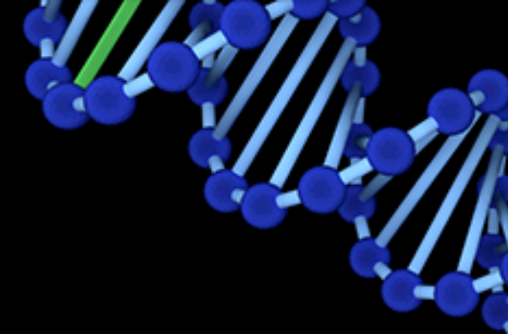
Temperature conversion can be fit to a linear regression model.

- 9/5 * °C + 32 = °F

There are two coefficients:

- 9/5
- 32

Linear Regression



The general form of linear regression is:

For temperature conversion there is only one x (Celsius value), so:

We have two coefficients: c0 and c1.

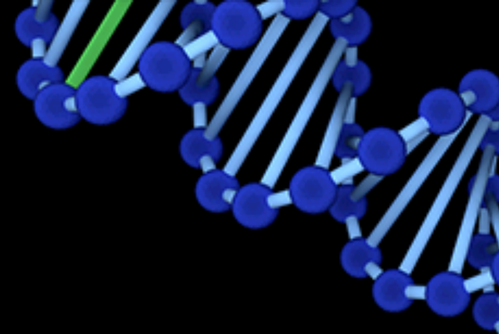
This is the same as or original conversion formula.

- 9/5 * °C + 32 = °F

Temperature Conversion Model

There are many ways to train (or fit) a model to the training data.

- For the temperature model, we know the coefficients.
- The coefficients are 9/5 and 32.
- Coefficient vector (array): [1.8,32]
- What if we did not know the coefficients? Yet we had several known temperatures in each scale.
- The computer can figure the coefficients for us!
- This is called model fitting, or training.



Training the Model

There are many ways to train (or fit) a model to the training data.

- Training is nothing more than finding coefficients that produce the desired output
- Training is usually iterative
- Marginal improvements are produced on each iteration
- Training can take a long time!
- The training method to use depends on if the model is differentiable

Non-Differentiable Models

There are many ways to train (or fit) a model to the training data if the model does not have a derivative.

- Genetic Algorithms
- Simulated Annealing
- Nelder Mead
- Hill Climbing
- Greedy Random Walk

Differentiable Models

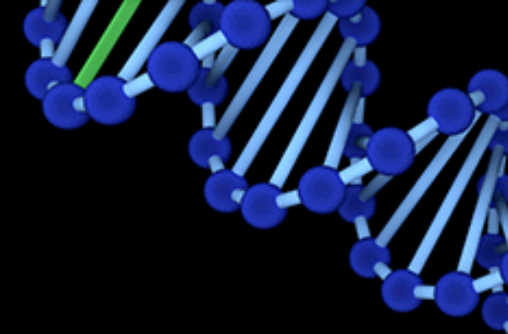
If the model does have a derivative there is likely a very mathematically efficient way to train.

- Gradient Descent
- Least Sum of Squares
- Levenberg–Marquardt
- Scaled Conjugate Gradient (SCG)

Beyond Linear Regression

Linear regression can be very useful, but it cannot fit all data sets. Other models can fit more complex data. Common models:

- General Linear Model (GLM)
- Neural Network
- Deep Belief Network
- Support Vector Machine (SVM)
- Bayesian Model



Model Similarities

Most models share many commonalities.

- Models accept inputs and produce one or more outputs
- Some models support regression, some support classification, others both
- All have a concept of weights, coefficients or other parameters
- Weights or coefficients are updated to train/fit the model
- You just need to select an appropriate training algorithm

RBF Neural Network

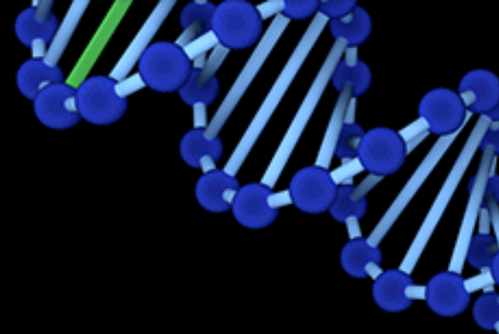
For my entry I used a RBF neural network trained by a Genetic algorithm. RBF networks…

- Can accept multiple inputs
- Are used for either classification or regression
- My entry used classification: malignant or benign
- Uses weights and variable number of RBF functions
- Training adjusts weights and RBF parameters

Radial Basis Functions

A Radial Basis Function (RBF) is a special class of function.

- The RBF’s used for RBF Networks typically have a range of 0 to 1
- The RBF’s have two parameters: center and width
- RBF’s are symmetrical about the center



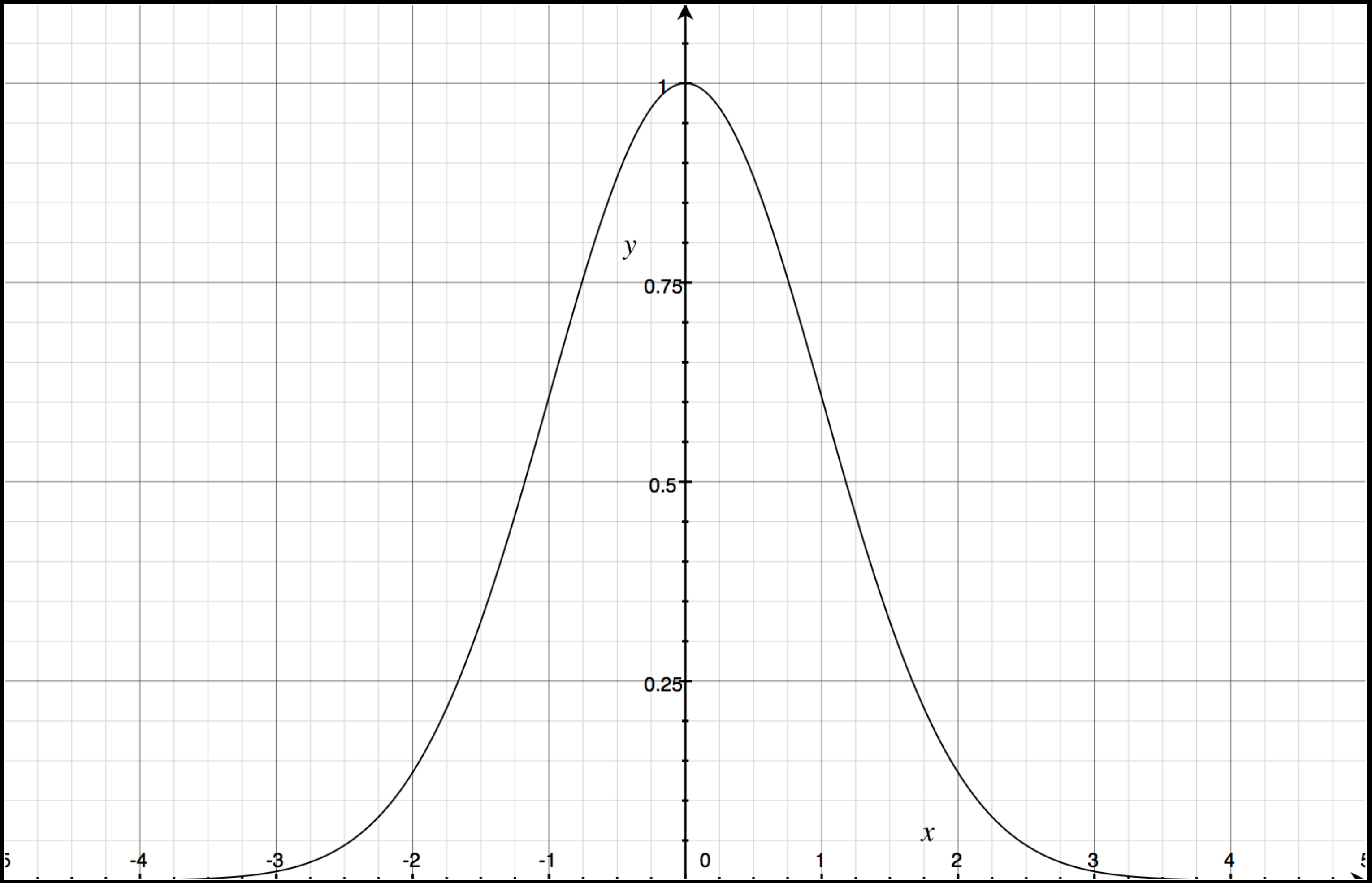
Gaussian RBF

Gaussian is a very common type of RBF. Here are its formulas:

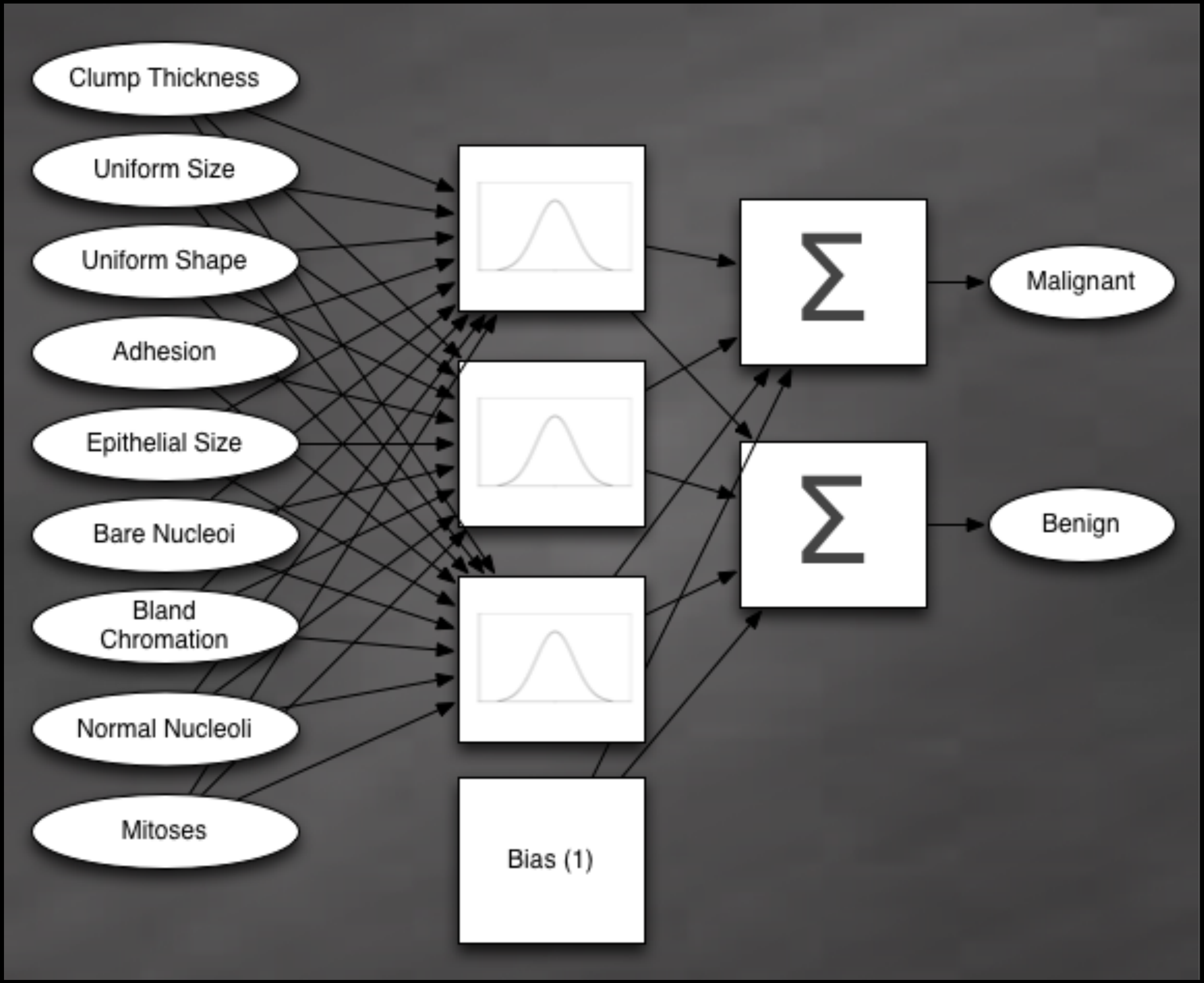
First calculate r, the sum of the distances between the center and input.

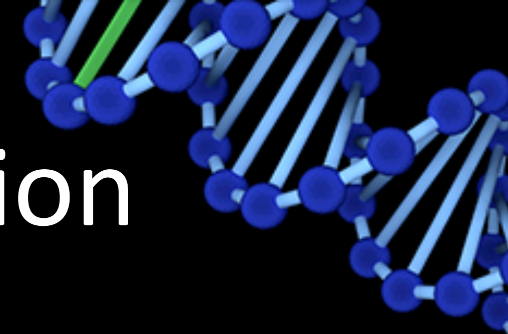
Now calculate the RBF using r. Below is the Gaussian RBF. The value w is the width.

Gaussian RBF Graph



RBF Network Diagram





RBF Network Calculation

Here is how an RBF network is calculated.

-
- X = Input vector/features
- N = Number of inputs/features/dimensions
- a, b = Connection weights
- p = RBF function, with a single defined-width
- Double-vertical bar-thingy = magnitude or distance
- c = center of the RBF

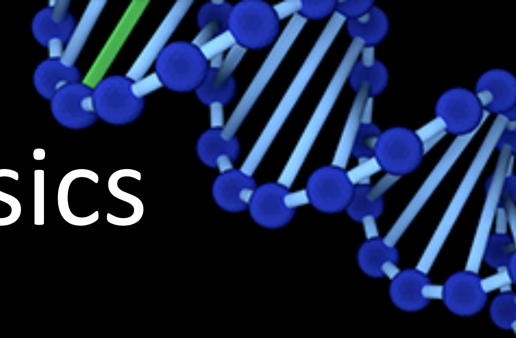
RBF Training Vector

The RBF training vector is a mix of weights and RBF parameters.

- Input coefficients/weights
- Output/Summation coefficients/weights
- RBF Width Values (same width in all dimensions)
- RBF Center Vectors

This all results in some sort of numeric vector. This vector is adjusted until the RBF network outputs what it is supposed to output given the input.

- [0.523, 1.236, … , 0.73625]



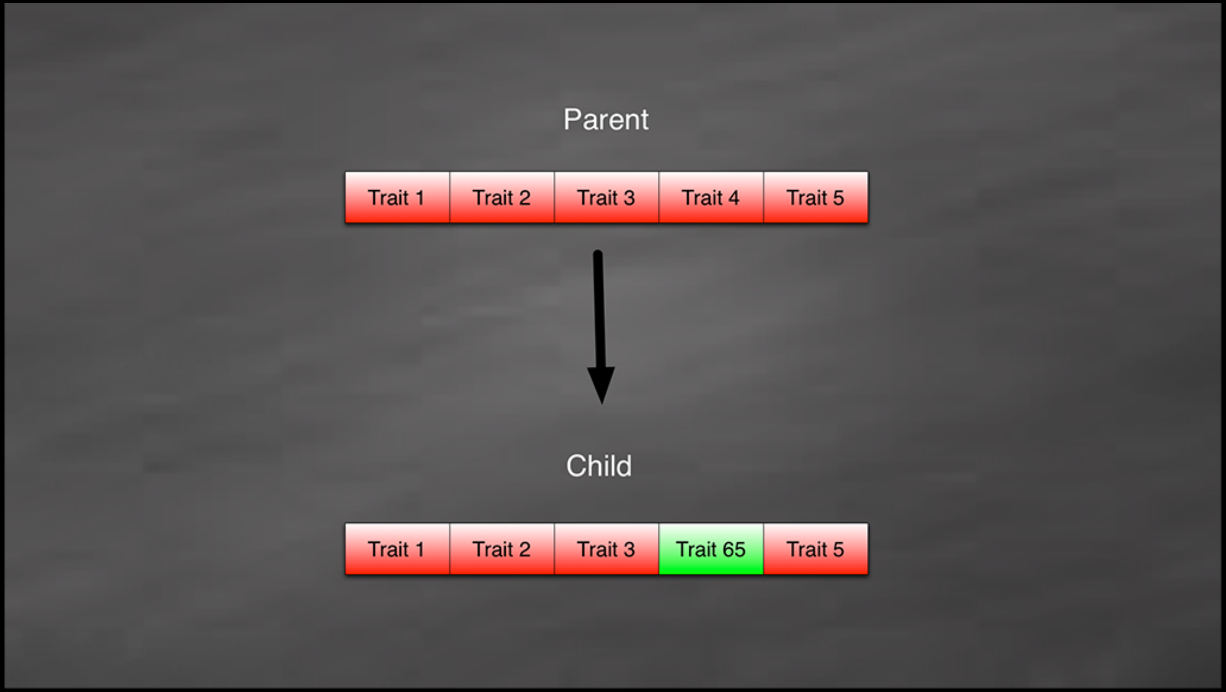
Genetic Algorithm Basics

A genetic algorithm can easily be used to fit a fixed-length training vector to a model. No derivative required.

- Create a population of some number of random individuals (i.e. 1,000)
- Devise a score function
- Select some relatively small number (i.e. 5) of the top individuals to pass into the next generation (elitism)
- Choose a certain percentage (i.e. 20%) of the top individuals to pass into the next generation with slight modifications (mutation/asexual reproduction)
- Fill the rest of the next generation with new individuals that share traits from pairs of the top individuals (crossover/sexual reproduction)

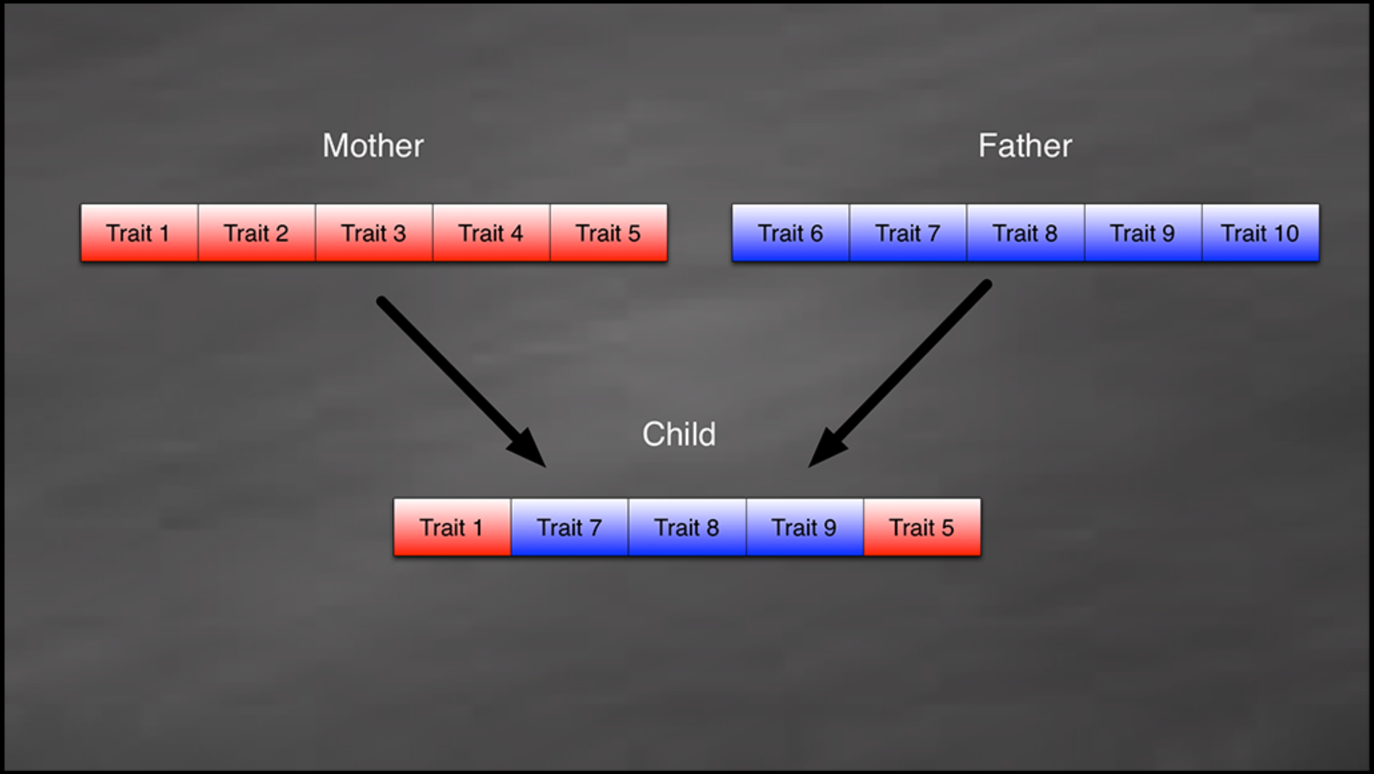
Mutation

Mutation slightly changes one individual and passes it onto the next generation. Mutation is critical for introducing “new” information.

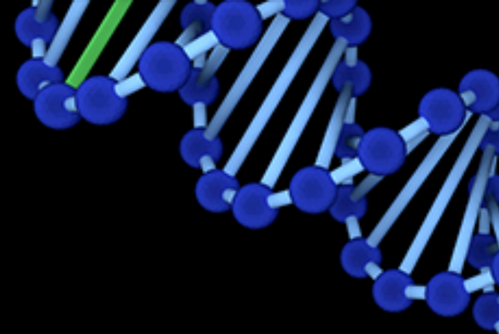


Crossover

Crossover allows two successful individuals to combine traits.



Advanced GA's



Here are some other features from very recent GA research.

- Species
- Does gender matter?
- Just two parents?
- Monogamy
- Innovation Database
- Tournaments (selection and antiselection)
- Islands and Land Bridges

Free (as in speech/beer) Code

You can download my source code from GitHub.

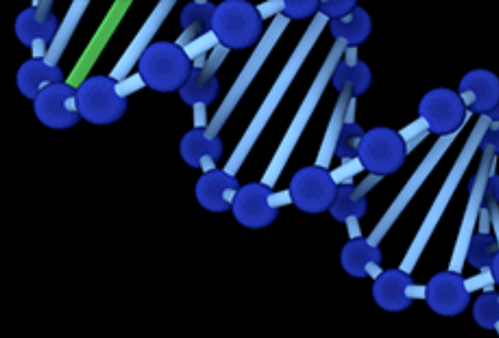
https://github.com/jeffheaton/ga-csharp

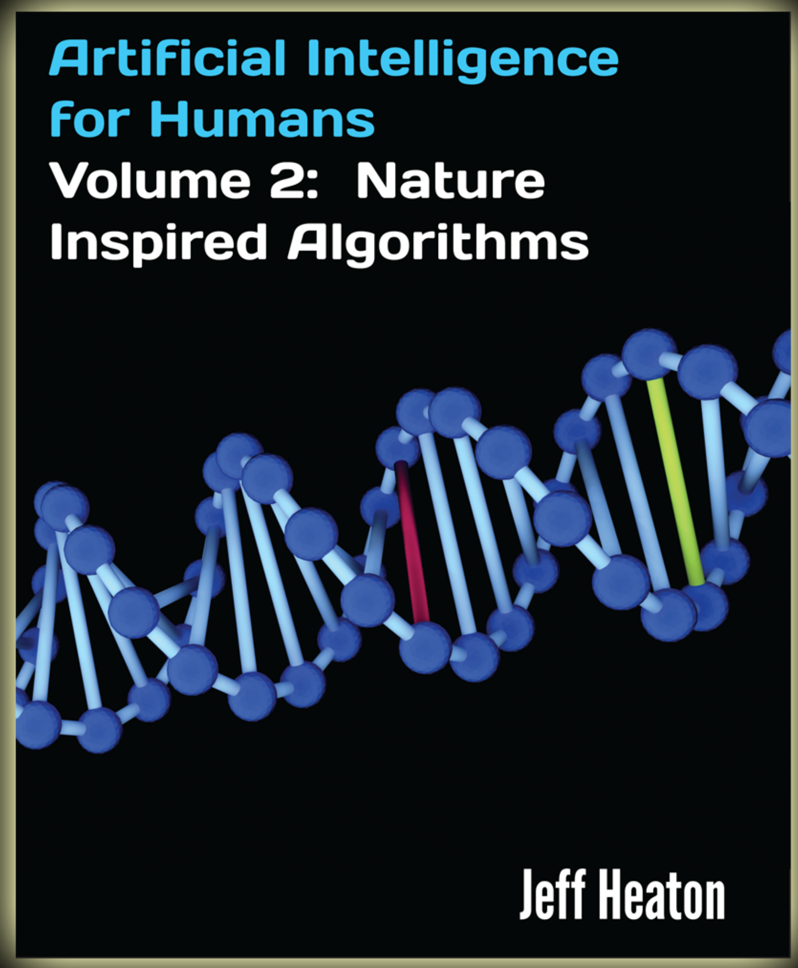
Using my GA Utility

My GA contest entry is meant to be a starting point.

- Reads/writes .xlsx files
- Scriptable via an .xml file
- Provides built in RBF-Network, but can be adapted to any model you program
- Uses C# Parallel class loops for maximum speed on multi-core processors

Artificial Intelligence for Humans





Any Questions?