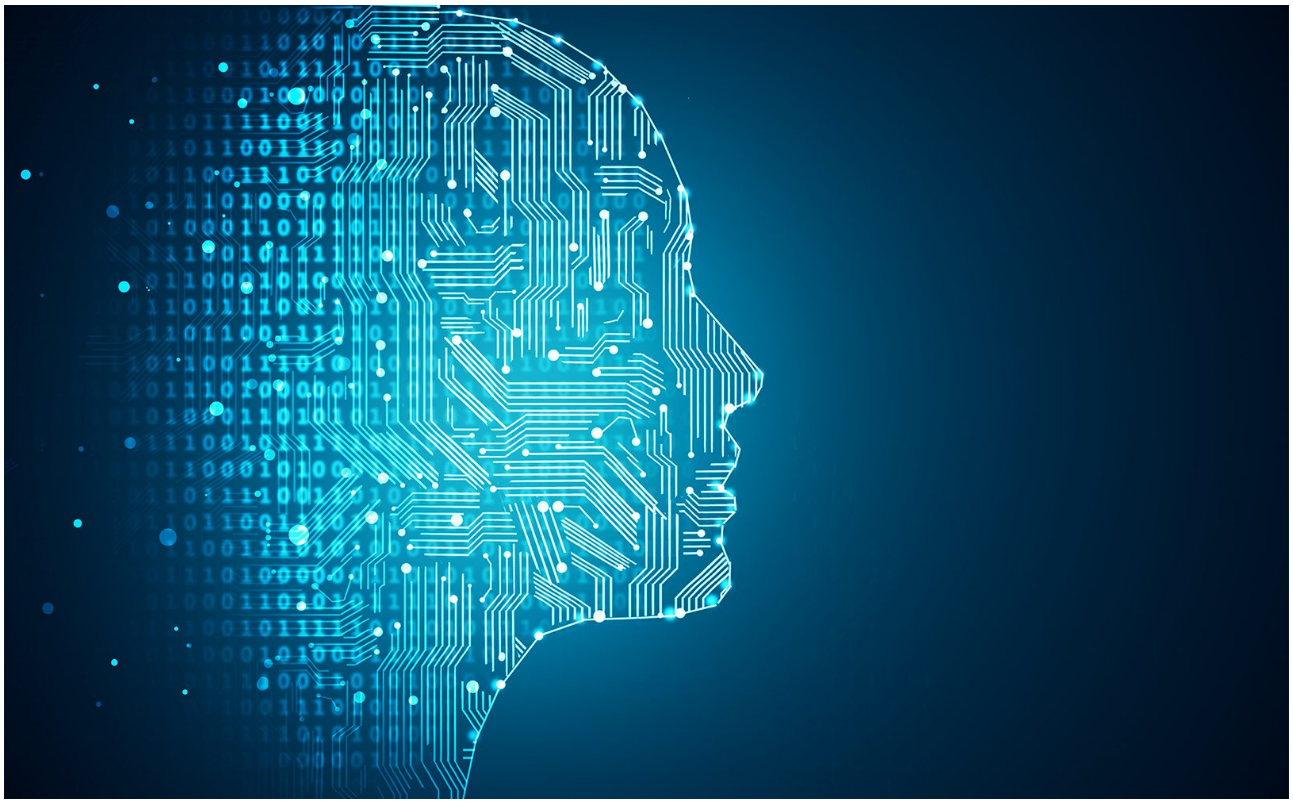
ECEN 4532 - Lab 6: Machine Learning

Andrew Teta

May 1, 2019



Contents

1 Introduction

Researchers in many different fields often collect large quantities of data, with the hope of finding it useful to develop predictions of various phenomena. One very common example in today's age would be advertisement targeting. As users browse the web, certain details about the demographic of people viewing and purchasing products can be collected and stored as a large data set. Machine learning may be a good way to predict what a random user will purchase based on their purchase history or demographic relationship. Many other problems similar to this can be solved using tools like machine learning. Often, they are problems of classification. By classifying users based on their gender, age, region of residency, or other factors, a program or algorithm can predict an outcome for a given activity. The classification involved may be binary or higher order, where a statistic could fall into one of many classification "bins". Other problems may involve continuous predictions.

When discussing the concept of machine learning, it is helpful to have some motivation and notation. Considering the first example of this lab, we have a domain of objects (passengers on the Titanic) which we would like to classify into categories ("survived or "perished"). Often, the objects are too complex to classify directly and so instead, an array of characteristics for each object are defined as *features*. These features are grouped into a *feature vector*. Each feature vector is referred to as an *instance* and the set of all feature vectors, the *instance space*.

We will then take a "training set" of data, which correctly maps the instances to its classification. The training set is incomplete and only contains a sample of all the possible instances, so doing our best, we can "learn" an approximation to the real function mapping features to classifications.

In this lab, we will investigate binary classification using Support Vector Machines, linear regression for predicting a continuous quantity, and multi-class classification using neural networks. The approach here is in the context of data analysis.

2 Background

Three data sets will be used for three different classification techniques, respectively. Specifically, we consider:

- Support Vector Machine (SVM) to predict who survived the sinking of the Titanic. This is an example of a binary classification problem (i.e. "survived" or "perished").
- Linear regression to predict the water resistance of a sailing yacht based on a selection of design parameters. The predicted value will be a real number here, so this is an example of a continuous classification problem.
- Neural networking to recognize hand-written digits. This is a multi-class classification problem as there are 10 classes or "bins".

For all cases, we will utilize a set of labeled outcomes. As an example, for hand-written digit recognition, we have 70,000 images of numerals 0 through 9, correctly labeled. Thus, we are practicing "supervised" machine learning by fitting an appropriate function to a data set based on a large number of input/output pairs.

2.1 Support Vector Machines

A Support Vector Machine (SVM) is a method of binary classification through supervised learning. A model is built based on a training set which can assign a new data point to one classification or the other. It is easiest to understand the process and goal of fitting this model with a graphical representation. Figure 1 shows that an SVM aims to find the optimal "line", called a *hyperplane* between data points falling into either classification.

Calling the hyperplane a "line" is inaccurate because really it forms a plane of separation between the feature vectors, represented by red squares and blue circles in fig. 1. This plane lives in multidimensional space. If each support vector has dimension p, then the hyperplane is of dimension $p-1$.

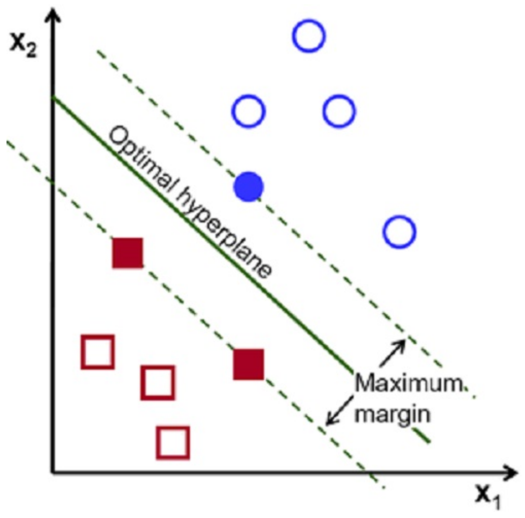


Figure 1: Graphical representation of an SVM hyperplane.
Credit: https://www.aitrends.com/ai-insider/support-vector-machines-svm-ai-self-driving-cars/

As the graphic suggests, another characteristic of SVMs is the margin of the hyperplane. An optimal hyperplane will maximize the margin, effectively finding the largest separation between classifications. Thus, a new data point will fall more decisively on one side of the hyperplane. Furthermore, the vectors defining the margin are called support vectors.

2.2 Linear Regression

Fundamentally, the problem of linear regression seeks to find a "best fit" line passing through a data set, to find a model to which predicted data can be fit. The concept of multiple linear regression deals with more than one independent variable. Linear regression is useful when a data set is "overdetermined"–often occurring in scientific or statistical measurement. This would be the case where a solution is desired for only one or two dependent variables, but hundreds or thousands of independent variable samples are collected. Then, rather than a single solution to what has become a linear algebra problem, we now have multiple solutions. Through the following proof, we can show that the "best" solution can be found. We begin with the assumption that we have a set of N measurements, x_i, collected together into a vector,

$$\vec{x} = (x_1, x_2, \ldots, x_N)$$

Recall, the length of $\vec{x}^2$ is its "energy". Additionally, we will refer to a vector of constants, all equal to μ as $\vec{\mu}$. Then, we can define the average energy per vector component of $\vec{x}$ as the error power and call this the *variance*

$$\sigma^2 = \frac{1}{N}|\vec{x} = \vec{\mu}|^2$$

where the vertical bars imply vector length and $\vec{\mu}$ is the mean vector of $\vec{x}$. Finding *standard deviation*, σ, we have:

$$\sigma = \frac{1}{\sqrt{N}}|\vec{x} - \vec{\mu}|$$

Now, considering the case of data point vectors, $\vec{x}$ and $\vec{y}$ (independent/independent variable pairs) we can begin to find a linear regression. First, we remap each vector into unit vector form:

$$\vec{w} = \frac{1}{\sigma_x}(\vec{x} - \vec{\mu_x})$$

$$\vec{z} = \frac{1}{\sigma_y}(\vec{y} - \vec{\mu_y})$$

We now seek a line, defined by the scalar constants a and b, which minimizes the cost (or error power) and write:

$$C(a,b) = |\vec{z} - (a\vec{w} + \vec{b})|^2$$

which can be written as a dot product:

$$\begin{aligned} C(a,b) &= (\vec{z} = (a\vec{w} + \vec{b})) \cdot (\vec{z} - (a\vec{w} + \vec{b})) \\ &= N - 2a\vec{w} \cdot \vec{z} + a^2N + Nb^2. \end{aligned}$$

Then, it is possible to show that $N - 2a\vec{w} \cdot \vec{z} + a^2N$ is a parabola in terms of a and has its minimum at

$$\begin{aligned} a &= \frac{\vec{w}\vec{z}}{N} \\ &= \frac{\vec{w}}{\sqrt{N}} \cdot \frac{\vec{z}}{\sqrt{N}} \\ &= cos\theta \end{aligned}$$

Thus, the value a is the correlation coefficient and is the cosine of the angle between the vectors $\vec{w}$ and $\vec{z}$. For notation consistency with other documentation, we will redefine $a = \rho$ and substitute to arrive at:

$$y = \beta x + (\mu_x - \beta\mu_x), \quad \text{where } \beta = \rho\frac{\sigma_y}{\sigma_x}$$

2.3 Neural Networks

The basic building blocks of a neural network are "perceptrons" (fig. 2) which compute the dot product of its input vector $\vec{x}$, with its weights, $\vec{w}$. The result is passed through a threshold with respect to a bias, b (fig. 3).

The process of training a neural network optimizes the parameters of the perceptrons in the network. The hidden layers shown in figure 2 can be many layers deep and define the complexity of the network, but eventually narrow the input into a single classification at the output layer.

3 Procedure

The procedure for this lab is divided into three parts.

1. Binary classification using an SVM
2. Continuous classification using linear regression
3. Multiple classification using a neural network

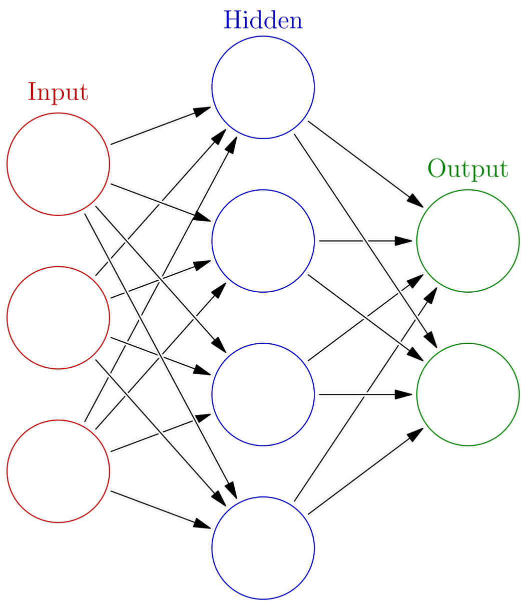


Figure 2: Perceptrons shown on the left.

3.1 SVM

For this portion of the lab, we will be using a training set of data, organized in multiple column fields of a `.csv` file saved to the disk.

The data set we will be using is a collection of information about passengers who were on the Titanic when it sank. The binary classification we wish to predict is whether each passenger "survived" or "perished". For each passenger, the file contains the following columns:

- Class
- Survived
- Name
- Sex
- Age
- Siblings and Parents
- Parents and Children
- Ticket number
- Fare
- Cabin
- Location Embarked
- Boat
- Body Weight
- Destination

Utilizing the Python module, `csv`, we can import this table and convert the fields into a `numpy` matrix of support vectors and a vector of labels. Some entries of this `.csv` table are in text format, others in integer, however all entries are initially imported as character arrays and so must be parsed correctly to prepare the data for processing in an SVM. Thus the data import procedure is:

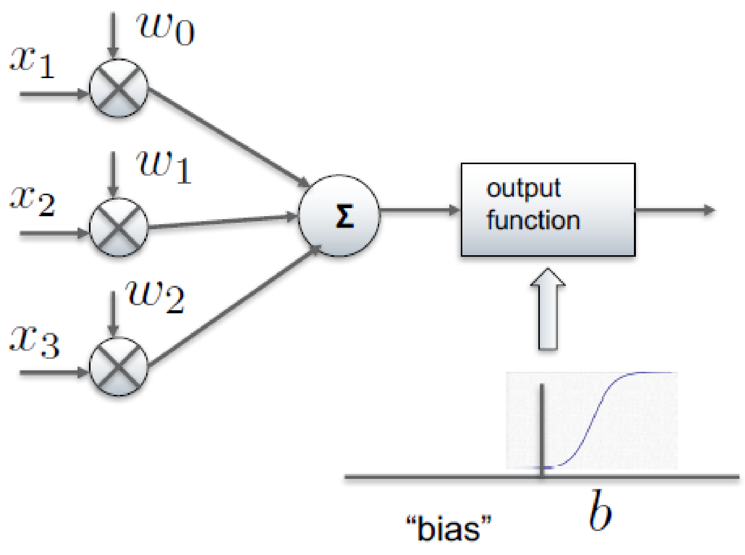


Figure 3: Inside of a perceptron.

1. Create `csv` reader object.
2. Iterate through `.csv` file, element by element.
3. Parse individual data fields and build support vectors for each instance
4. Build support vector matrix and label vector

We will not import all columns into our data matrix, but instead use only:

- Class
- Sex
- Age
- Siblings and Parents
- Parents and Children
- Fare
- Location Embarked

and use the "survived" column as our label vector.

Next, we want to analyze this data using the techniques of an SVM. Rather than writing the complex algorithms from scratch, we will utilize a Python module, `SVM` from the package `sklearn`. Then, we feed our support vector matrix and label vector to the `fit()` function to "train" our SVM. Finally, we can run some predictions using the `predict()` function and analyze the performance of our machine. Again, listing the procedure:

1. Declare a `sklearn` SVM object.
2. `fit()` the SVM to our data set, passing the label vector in.
3. `predict()` an outcome for each instance and measure performance.

3.1.1 Results

We started by training the SVM on all features. While the outcome was not bad, the process took a very long time, so we considered limiting the training data set to only three features. Defining the percentage of correctly predicted outcomes as p, we found:

$$p = 80.51\%$$

To find the optimal features to use, while still achieving similar prediction performance, we ran the SVM training and prediction process on each feature individually. The results for each feature are shown below.

Feature	Outcome Success (%)
Fare	66.00
Class	67.68
Sex	77.99
Age	66.08
Siblings and Parents	62.33
Parents and Children	64.32
Location Embarked	64.17

Having obtained some insight into which features are most correlated with the outcome, we re-combined multiple features into our training set. Selecting the three features with best prediction accuracy, we found highest overall accuracy based on age, sex, and class to be:

$$p = 78.99\%$$

which is very close to the performance we had with all features considered. This is a much better solution, because the processing time was much less.

3.2 Linear Regression

Again, using `sklearn`, this time we will use a module from the `linear_model` package, called `LinearRegression`. And, you guessed it, this module will help us perform linear regression.

Overall, the procedure is very similar to what we did for the SVM.

- Import training data from a `.csv`
  - Organize independent variables in a matrix and labels into a vector
- `fit()` a model to the data, using the label vector
- `predict` the outcome based on the input
- Statistically measure the performance

For this portion of the lab, we will be using linear regression to fit a model to a set of data containing a list of parameters used in the design of a yacht hull. These design parameters are suspected to contribute to the resulting resistance seen by the yacht against the water. The purpose of this model will then be to accurately measure which parameters contribute most heavily to yacht resistance and fit a model which could then be used to optimize a design.

The design parameters considered are:

- Longitudinal Position Center of Buoyancy

- Prismatic Coefficient
- Length-Displacement Ratio
- Beam-Draught Ratio
- Length-Beam Ratio
- Froude Number

And, more formally, the value we wish to predict is the "Residuary resistance per unit weight of displacement."

3.2.1 Measuring Performance

It will be important to compare the performance of two regression model predictors and so we will define the two terms,

- Total Sum of Squares (TSS)
- Residual Sum of Squares (RSS)

Where,

$$TSS = \sum_{i=1}^{M} (y_i - \mu_y)^2$$

and

$$RSS = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2.$$

Then, the "Fraction of Variance Unexplained" is defined as

$$FVU = \frac{RSS}{TSS}$$

and the "Coefficient of Determination", R^2 as

$$R^2 = 1 - FVU = 1 - \frac{RSS}{TSS}.$$

The value R^2 is essentially the fraction of variance of y (the outcome) that is "explained" by the prediction. As follows, the quantity, $1 - R^2$ is the fraction of variance that is "unexplained" by the prediction. Thus, we should attempt to maximize the value of R^2 to find the best prediction model.

3.2.2 Results

To begin, we `fit()` a linear regression model to all six design parameters over a range of over 300 sample data points for each. Running the prediction over the training set, we found the value of $R^2 = 0.6575$. Figure 4 shows graphically what this outcome looks like. This R^2 value is not great, but there are a couple things to consider. First, not all the design parameters are guaranteed to have a significant effect on the outcome, and so could be introducing unnecessary model parameters. Secondly, and also more importantly, we have just trained this regression model with the assumption of a linear relationship between each of the design parameters and outcome.

Considering first, that some parameters may not be contributing much, we ran the same procedure of fitting, predicting, and calculating R^2 for each parameter set alone. The results are shown in table 3.2.2.

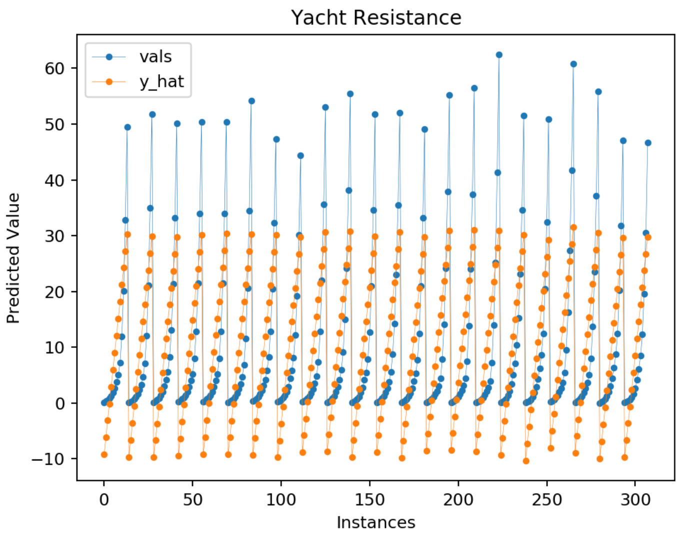


Figure 4: Comparison of labeled outcomes and predicted outcomes based on a linear regression learning method.

Design Parameter	R^2
Longitudinal Position Center of Buoyancy	0.000372
Prismatic Coefficient	0.000816
Length-Displacement Ratio	$8.8e^{-6}$
Beam-Draught Ratio	0.000154
Length-Beam Ratio	$1.05e^{-6}$
Froude Number	0.656

It is easy to see that most parameters do not contribute much to the variance of the outcome, with the highest contenders being:

1. Froude Number
2. Prismatic Coefficient
3. Longitudinal Position Center of Buoyancy

Taking this into consideration, we ran the procedure again, using only these three parameters. As may have been expected, however, we found $R^2 = 0.6574$, with the remaining significant figures attributed to the other parameters. Since this obviously did not improve our prediction results, we need to look for another cause.

To gain some insight into the relationship between each parameter and the outcome, we plotted all data points for each, as shown in figure 5. It can be seen than the Prismatic coefficient and Longitudinal positional center of buoyancy do not hold a strong relationship. However, the Froude number appears to fit an exponential relationship.

It would be logical to assume a form $y = e^x$ for the Froude number. Taking $y = e^{fn}$, we found $R^2 = 0.6934$. After some trial and error, a quickly deduced "optimal" solution was found for $y = e^{20 \cdot fn}$, resulting in $R^2 = 0.9856$. Then, for ease of calculation, we combined the Prismatic coefficient, Longitudinal positional center of buoyancy, and Froude number vectors into a single matrix, M and calculated

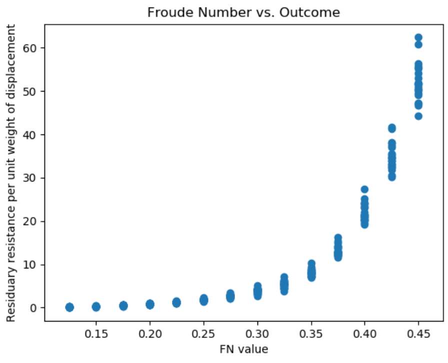
$$R^2 = 0.9870$$

for

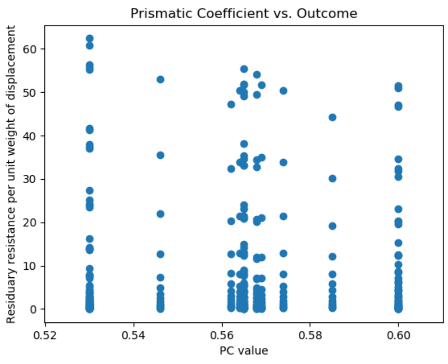
$$e^{20 \cdot M}.$$

Additionally, the labels and prediction values were plotted together and are shown in figure 6

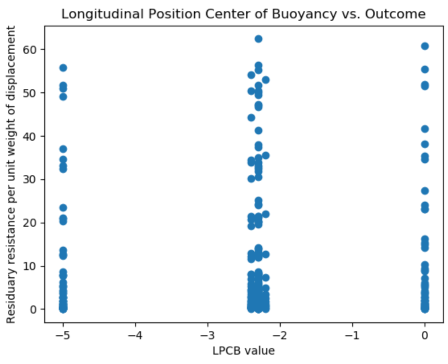
3.3 Neural Network



(a)



(b)



(c)

Figure 5: The relationship between input and output of the three highest scoring parameters.

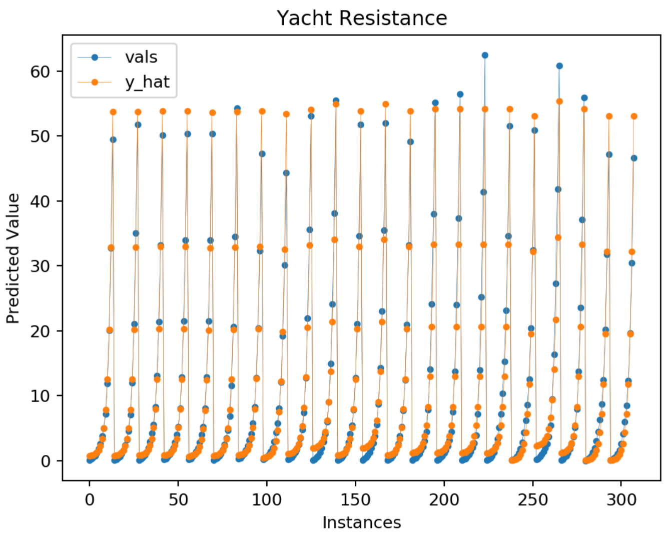


Figure 6: Residuary resistance per unit weight of displacement (labeled vs. predicted).