Implementing of RankNet

Ryan Greenup

February 21, 2021

Contents

CONTENTS

#+TODO: TODO IN-PROGRESS WAITING DONE

1   Introduction

Ranknet is an approach to *Machine-Learned Ranking (often refered to as "*/Learning to Rank" [27]) that began development at Microsoft from 2004 onwards [7], although previous work in this area had already been undertaken as early as the 90s (see generally [12, 11, 12, 13, 52]) these earlier models didn't perform well compared to more modern machine learning techniques [30, §15.5].

Information retrieval is an area that demands effective ranking of queries, although straight-forward tools such as grep, relational databases (e.g. sqlite, MariaDB, Post-greSql) or NoSQL (e.g. CouchDB, MongoDB) can be used to retrieve documents with matching characters and words, these methods do not perform well in real word tasks across large collections of documents because they do not provide any logic to rank results (see generally [48]).

2   Motivation

Search Engines implement more sophisticated techniques to rank results, one such example being TF-IDF weighting [31] , well established search engines such as *Apache Lucene* [2] and *Xapian* [19], however, are implementing Machine-Learned Ranking in order to improve results.

This paper hopes to serve as a general introduction to the implementation of the Ranknet technique to facilitate developers of search engines in more modern languages (i.e. Go and Rust) in implementing it. This is important because these more modern languages are more accessible [18] and memory safe [38] than C/C++ respectfully, without significantly impeding performance; this will encourage contributors from more diverse backgrounds and hence improve the quality of profession-specific tooling.

For a non-comprehensive list of actively maintained search engines, see §9.1 of the appendix.

3   Implementation

Neural Networks

Ranking/ is the process of applying machine learning algorithms to ranking problems, it .

This implementation will first apply the approach to a simple data set so as to clearly demonstrate that the approach works, following that the model will be extended to support wider and more complex data types before finally being implemented on a corpus of documents.

3.1   Neural Networks

The Ranknet method is typically implemented using a Neural Networks [1], although other machine learning techniques can also be used [7, p. 1], Neural Networks are essentially a collection of different regression models and classifiers that are fed into one another to create a non-linear classifier, a loss function is used to measure the performance of the model with respect to the parameters (e.g. RMSE [2] or BCE [3]) and the parameters are adjusted so as to reduce this error by using the *Gradient Descent Technique* (although there are other optimisation algorithms such as RMSProp and AdaGrad [32] that can be shown to perform better, see [4]). The specifics of Neural Networks are beyond the scope of this paper (see [16] or more generally [41]).

3.1.1   The Ranknet Method

The Ranknet method is concerned with a value p_{ij} that measures the probability that an observation i is ranked higher than an observation j.

A Neural Network (n) is trained to return a value s_k from a feature vector $\mathbf{X}_k$:

$$n(\mathbf{X}_i) = s_i \quad \exists k$$

So as to minimise the error of:

$$p_{ij} = \text{sig}\left(\sigma, (s_i - s_j)\right) \quad \exists \sigma \in \mathbb{R} \tag{1}$$
$$\text{where:}$$
$$\text{sig}\left(\sigma, x\right) = \frac{1}{1 + e^{\sigma \cdot x}} \tag{2}$$

Version Control   The implementation in this paper corresponds to the `walkthrough` branch of the `git` repository used in production of this work, id values (e.g. `:08db5b0:`) will be appended to titles to denote specific changes made in that section. See §9.4 for more specific guidance.

Code listings   The code listings provided will use a standard set of import statements (see §9.2) and so they will be omitted from the listings, for more comprehensive guidance on implementing this code refer to the documentation page[4] that accompanies the `git` repo.

---

[1] An early goal of this research was to evaluate the performance of different machine learning algorithms to implement the Ranknet method, as well as contrasting this with simple classification approaches, this research however is still ongoing, see §5.3

[2] **RMSE** *Root Mean Square Error*

[3] **BCE** *Binary Cross Entropy*

[4] crmds.github.io/CRMDS-HDR-Training-2020/

3.2   Creating Data   CF9AB26

The first step is to create a simple data set and design a neural network that can classify that data set, the data set generated should have two classes of data (this could be interpreted as relevant and irrelevant documents given the features or principle components of a data set), this can be implemented using sci kit learn as shown below and visualized[5] in figure 1.[6]

In order to fit a Neural Network the *PyTorch* package can be used [37], this will allow the gradients of the neural network to be calculated numerically without needing to solve for the partial derivatives, hence the data will need to be in the form of tensors.

```python
def make_data(create_plot=False, n=1000, dtype=torch.float, dev="cpu", export=""):
    X, y = datasets.make_blobs(n, 2, 2, random_state=7)
    # X, y = datasets.make_moons(n_samples=n, noise=0.1, random_state=0) # Moons Data
    ↪  for later

    # Save the data somewhere if necessary
    if export != "":
        export_data(X, y, export)

    # Reshape the data to be consistent
    y = np.reshape(y, (len(y), 1))  # Make y vertical n x 1 matrix.

    # -- Split data into Training and Test Sets --------------------
    data = train_test_split(X, y, test_size=0.4)

    if(create_plot):
        # Create the Scatter Plot
        plt.scatter(X[:, 0], X[:, 1], c=y)
        plt.title("Sample Data")
        plt.show()

    # Make sure we're working with tensors not mere numpy arrays
    torch_data = [None]*len(data)
    for i in range(len(data)):
        torch_data[i] = torch.tensor(data[i], dtype=dtype, requires_grad=False)

    return torch_data

# Set Torch Parameters
dtype = torch.float
dev = test_cuda()

# Generate the Data
X_train, X_test, y_train, y_test = make_data(
    n=int(300/0.4), create_plot=True, dtype=dtype, dev=dev, export =
    ↪  "/tmp/simData.csv")
```

---

[5]See §9.3 for the specific method definition used to export the data to a `csv`.

[6]Visualisations for this Report were implemented using `org-babel` [10] inside *Emacs* [44] to call **R** [43] with *GGPlot2* [50] (and *Tidyverse* [51] generally), the source code for this is avaliable in the report manuscript available in the `git` repository available at github.com/RyanGreenup/ranknet/blob/main/Report/Report.org
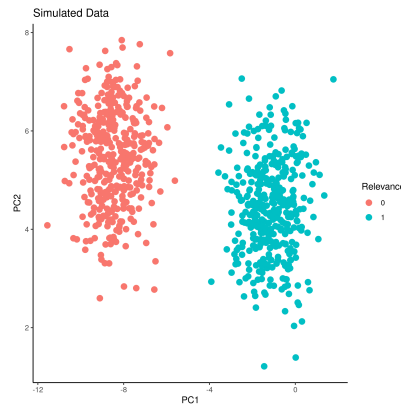
Figure 1: Generated data, output classes denote document relevance and the axis features or principle components

3.3 Creating a Neural Network   7291112

A Neural Network model can be designed as a class, here a 2-layer model using Sigmoid functions has been described, this design was chosen for it's relative simplicity:

```python
class three_layer_classification_network(nn.Module):
    def __init__(self, input_size, hidden_size, output_size, dtype=torch.float,
    ↪ dev="cpu"):
        super(three_layer_ranknet_network, self).__init__()
        self.wi = torch.randn(input_size, hidden_size,
                              dtype=dtype,
                              requires_grad=True,
                              device=dev)
        self.wo = torch.randn(hidden_size, output_size,
                              dtype=dtype,
                              requires_grad=True,
                              device=dev)

        self.bi = torch.randn(hidden_size,
                              dtype=dtype,
                              requires_grad=True,
                              device=dev)
        self.bo = torch.randn(output_size,
                              dtype=dtype,
                              requires_grad=True,
                              device=dev)

        self.σ = torch.randn(1, dtype=dtype, requires_grad=True, device=dev)

        self.losses = []        # List of running loss values
        self.trainedQ = False  # Has the model been trained yet?

    def forward(self, x):
        x = torch.matmul(x, self.wi).add(self.bi)
        x = torch.sigmoid(x)
        x = torch.matmul(x, self.wo).add(self.bo)
        x = torch.sigmoid(x)
        return x

    def loss_fn(self, x, y):
```

```
35            y_pred = self.forward(x)
36            return torch.mean(torch.pow((y-y_pred), 2))
37
38        def misclassification_rate(self, x, y):
39            y_pred = (self.forward(x) > 0.5)
40            return np.average(y != y_pred)
```

A model can then be instantiated, a 2-3-1 model has, arbitrarily, been implemented in this case:[7]

```
1  # Set Seeds
2  torch.manual_seed(1)
3  np.random.seed(1)
4
5  # Set Torch Parameters
6  dtype = torch.float
7  dev = test_cuda()
8
9  # Make the Data
10 X_train, X_test, y_train, y_test = make_data(
11     n=100, create_plot=True, dtype=dtype, dev=dev)
12
13 # Create a model object
14 model = three_layer_classification_network(
15     input_size=X_train.shape[1], hidden_size=2, output_size=1, dtype=dtype, dev=dev)
16
17 # Send some data through the model
18 print("\nThe Network input is:\n---\n")
19 print(X_train[7,:], "\n")
20 print("The Network Output is:\n---\n")
21 print(model.forward(X_train[7,:]).item(), "\n")
```

```
The Network input is:
---

tensor([-1.5129,  2.9332])

The Network Output is:
---

0.22973690927028656
```

3.4   Train the Model with Gradient Descent

Now that the model has been fit, a method to train the model can be implmented [8]:

```
1  class three_layer_classification_network(nn.Module):
2      # __init__ method goes here, see above
3      # ...
```

---

[7]note that the model has not yet been trained, the weights are random and the model output is not related to the data at all.

[8]This class definition is incomplete and serves only to show the method definition corresponding to the original class shown in §3.3

```
4        # ...
5
6        def train(self, x, target, η=30, iterations=2e4):
7            bar = Bar('Processing', max=iterations) # progress bar
8            for t in range(int(iterations)):
9
10               # Calculate y, forward pass
11               y_pred = self.forward(x)
12
13               # Measure the loss
14               loss = self.loss_fn(x, target)
15
16               # print(loss.item())
17               self.losses.append(loss.item())
18
19               # Calculate the Gradients with Autograd
20               loss.backward()
21
22               with torch.no_grad():
23                   # Update the Weights with Gradient Descent
24                   self.wi -= η * self.wi.grad; self.wi.grad = None
25                   self.bi -= η * self.bi.grad; self.bi.grad = None
26                   self.wo -= η * self.wo.grad; self.wo.grad = None
27                   self.bo -= η * self.bo.grad; self.bo.grad = None
28                   self.σ  -= η * self.σ.grad;  self.σ.grad = None
29               bar.next()
30           bar.finish()
31               # ; Zero out the gradients, they've been used
32
33       # Rest of the Class Definition Below ...VVV...
```

With this definition the model can hence be trained in order to produce meaningful classifications, as shown below, this model classifies the points perfectly, even on the testing data, the training error over time is shown in figure 2.

```
1   # Make the Data
2   X_train, X_test, y_train, y_test = make_data(
3       n=100, create_plot=True, dtype=dtype, dev=dev)
4
5   # Create a model object
6   model = three_layer_classification_network(
7       input_size=X_train.shape[1], hidden_size=2, output_size=1, dtype=dtype, dev=dev)
8
9   # Train the Model
10  model.train(X_train, y_train, η=1e-2, iterations=10000)
11
12  # Plot the losses
13  plt.plot(model.losses)
14  plt.title("Losses at each training iteration")
15  plt.show()
16
17  print("The testing misclassification rate is:\n")
18  print(model.misclassification_rate(X_test, y_test))
```

3.5   Implement Ranknet   F25F376:05DF04F

Now that the model can classify the data, the implementation will be modified to:

- Measure loss using a BCE function which is reported to perform better in the
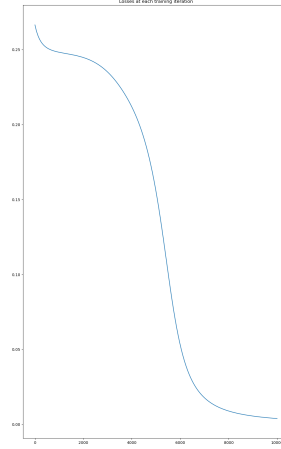
Figure 2: Training error, given by $l\left(x\right) = \sum_{i=1}^{n}\left[\left(x_i - f\left(x_i\right)\right)^2\right]$, at each iteration of training

literature [7, 6]

- Modify the model so that it operates pairwise, such that:

  1. Two points are identified, sent through the neural network and two values returned:

  $$s_i = n(\mathbf{X}_i) \tag{3}$$
  $$s_j = n(\mathbf{X}_j) \tag{4}$$

  The network previously created can be adapted for this and hence the method will be renamed to `forward_single` and this will represent function $n()$ implemented in (3) and (4)

  2. These values will be combined to give a single value which is intended to measure the model confidence:[9]

  $$\hat{P}_{ij} = \text{sig}\left(\sigma, \left(s_i - s_j\right)\right), \quad \exists \sigma \in \mathbb{R} \tag{5}$$
  $$= \frac{1}{1 + e^{\sigma \cdot (s_i - s_j)}} \tag{6}$$

  3. The range of (6) is the interval $\hat{P}_{ij} = [0, 1]$, let $\bar{P}_{ij}$ be the known probability[10] that $\mathbf{X}_i \triangleright \mathbf{X}_j$, the simulated data has a boolean range of $\bar{P}_{ij} \in \{0, 1\}$, this can be recast to $\{-1, 0, 1\}$ and then linearly scaled to $[0, 1]$ like so:

---

[9]This value is a measurement of the models "confidence" but could be extended to represent the "measured probability" of one item being ranked higher than an other (e.g. the probability that a person would rank one type of wine as better than the other in a random sample).

[10]Note the convention that the symbols $\triangleleft$, $\triangleright$ have been adopted to denote the ranking of two observations, analogous to $<$, $>$

$$\bar{P}_{ij} \leftarrow p_i - p_j \tag{7}$$

$$\bar{P}_{ij} \leftarrow \frac{1 + \bar{P}_{ij}}{2} \tag{8}$$

These modifications only need to be made to the neural network class like so:

```python
class three_layer_ranknet_network(nn.Module):
    # __init__ method
    # ...
    # ...

    def forward(self, xi, xj):
        si = self.forward_single(xi)
        sj = self.forward_single(xj)
        out = 1 / (1 + torch.exp(-self.σ * (si - sj)))
        return out

    def forward_single(self, x):
        x = torch.matmul(x, self.wi).add(self.bi)
        x = torch.sigmoid(x)
        x = torch.matmul(x, self.wo).add(self.bo)
        x = torch.sigmoid(x)

        return x

    def loss_fn(self, xi, xj, y):
        y_pred = self.forward(xi, xj)
        loss = torch.mean(-y * torch.log(y_pred) -
                          (1 - y) * torch.log(1 - y_pred))
        return loss

def pairwise(iterable):
    "pairwise([1,2,3,4]) --> [(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)]"
    s = list(iterable)
    pair_iter = chain.from_iterable(combinations(s, r) for r in [2])
    return pair_iter
```

The training method must be adapted to interact with these changes like so:[11]

```python
class three_layer_ranknet_network(nn.Module):
    # __init__ method
    # ...
    # ...
    def train(self, x, target, η=1e-2, iterations=4e2):
        self.trainedQ = True
        # Create a progress bar
        bar = Bar('Processing', max=iterations)
        # Train for a number of iterations
        for t in range(int(iterations)):
            sublosses = []
            # Loop over every pair of values
            for pair in pairwise(range(len(x) - 1)):
                xi, yi = x[pair[0], ], target[pair[0]]
```

---

[11]Note the definition of the pairwise function, this was incorrectly implemented initially (f25f376) and rectified shortly after (05df04f). see 9.4.1

```
15                    xj, yj = x[pair[1], ], target[pair[1]]
16
17                    # encode from {0, 1} to {-1, 0, 1}
18                    y = yi - yj
19
20                    # Scale between {0,1}
21                    y = 1 / 2 * (1 + y)
22
23                    # Calculate y, forward pass
24                    y_pred = self.forward(xi, xj)
25
26                    # Measure the loss
27                    loss = self.loss_fn(xi, xj, y)
28                    sublosses.append(loss.item())
29
30                    # Calculate the Gradients with Autograd
31                    loss.backward()
32
33                    # Update the Weights with Gradient Descent
34                    # ; Zero out the gradients, they've been used
35                    with torch.no_grad():
36                        self.wi -= η * self.wi.grad; self.wi.grad = None
37                        self.bi -= η * self.bi.grad; self.bi.grad = None
38                        self.wo -= η * self.wo.grad; self.wo.grad = None
39                        self.bo -= η * self.bo.grad; self.bo.grad = None
40                        self.σ  -= η * self.σ.grad ; self.σ.grad  = None
41
42                self.losses.append(np.average(sublosses))
43                bar.next()
44            bar.finish()
45            self.threshold_train(x, target, plot=False)
```

This can then be implemented as before, the loss function is provided at figure 3.

```
1   # Make the Data
2   X_train, X_test, y_train, y_test = make_data(
3       n=30, create_plot=True, dtype=dtype, dev=dev)
4
5   # Create a model object
6   model = three_layer_ranknet_network(
7       input_size=X_train.shape[1], hidden_size=2, output_size=1, dtype=dtype, dev=dev)
8
9   # Train the Model
10  model.train(X_train, y_train, η=1e-1, iterations=1e2)
11
12  # Save the Losses
13  np.savetxt(fname="/tmp/losses.csv", X=model.losses, delimiter=',')
```

3.6 Implement sorting   99B390A:7D46636

One of the difficulties in implementing this, however, is that it is not simple to determine whether or not the model has classified the data well,[12] In order to address this the model can be implemented to sort the data by ranked values and then visualised. To

---

[12]A naive misclassification method was implemented (f25f376), but it was not very insightful and so was omitted from this report.
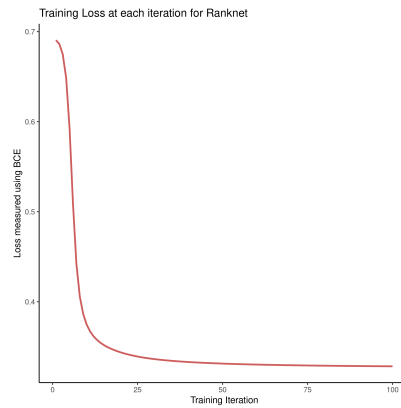
Figure 3: BCE training loss at each iteration for the Ranknet method.

implement this a derivative of the *quicksort* algorithm was chosen as a sorting function [17], this was implemented by adapting code already available in the literature [23, §4.10] and online [42]:

```python
def split(values, left, right, data, model):
    # Define the leftmost value
    l = (left-1)
    # Set the right value as the pivot
    pivot = values[right]  # TODO The pivot should be random

    for q in range(left, right):
        # Only move smaller values left
        if leq(values[q], pivot, data, model):
            # +1 next left element
            l = l+1
            # Swap the current element onto the left
            values[l], values[q] = values[q], values[l]

    # Swap the pivot value into the left position from the right
    values[l+1], values[right] = values[right], values[l+1]
    return (l+1)


def qsort(values, left, right, data, model):
    if len(values) == 1:
        return values
    if right > left:
        # pi is the index of where the pivot was moved to
        # It's position is now correct
        pi = split(values, left, right, data, model)

        # Do this again for the left and right parts
        qsort(values, left, pi-1, data, model)
        qsort(values, pi+1, right, data, model)

import random

def leq(a, b, data, model):
    score = model.forward(data[a, :], data[b, :])
    if score <= 0.5:
        return True
```

```
38          if score > 0.5:
39              return False
40
41          if (a < b):
42              return True
43          else:
44              return False
45
46
47  if DEBUG:
48      for i in range(3):
49          import random
50          values = random.sample(range(9), 7)
51          n = len(values)
52          print(values)
53          qsort(values, 0, n-1, data, model)
54          print("==>", values)
```

The data can then be plotted, as in figure and exported like so:[13]

```
1   # Main Function
2   def main():
3       # Make the Data
4       X_train, X_test, y_train, y_test = make_data(n=100,
5                                                    create_plot=True,
6                                                    dtype=dtype,
7                                                    dev=dev)
8
9       # Create a model object
10      model = three_layer_ranknet_network(input_size=X_train.shape[1],
11                                          hidden_size=2,
12                                          output_size=1,
13                                          dtype=dtype,
14                                          dev=dev)
15
16      # Train the Model
17      model.train(X_train, y_train, η=1e-1, iterations=1e2)
18
19      # Visualise the Training Error
20      plot_losses(model)
21
22      # Misclassification won't work for ranked data
23      # Instead Visualise the ranking
24      plot_ranked_data(X_test, y_test, model)
25
26
27  def plot_losses(model):
28      plt.plot(model.losses)
29      plt.title("Cost / Loss Function for Iteration of Training")
30      plt.show()
31
32
33  def plot_ranked_data(X, y, model):
34      # Create a list of values
35      n = X.shape[0]
36      order = [i for i in range(n)]
37      # Arrange that list of values based on the model
```

---

[13]In this case the plot has been generated by *GGPlot2* [6]

```
38        quicksort(values=order, left=0, right=(n - 1), data=X, model=model)
39        print(order)
40
41        ordered_data = X[order, :]
42        y_ordered = y[order]
43
44        np.savetxt("/tmp/ordered_data.csv", X=ordered_data.numpy(), delimiter=',')
45
46        p = plt.figure()
47        for i in range(len(ordered_data)):
48            plt.text(ordered_data[i, 0], ordered_data[i, 1], i)
49        plt.scatter(ordered_data[:, 0], ordered_data[:, 1], c=y_ordered)
50        plt.title("Testing Data, with ranks")
51        plt.show()
52
53
54   if __name__ == "__main__":
55       main()
```
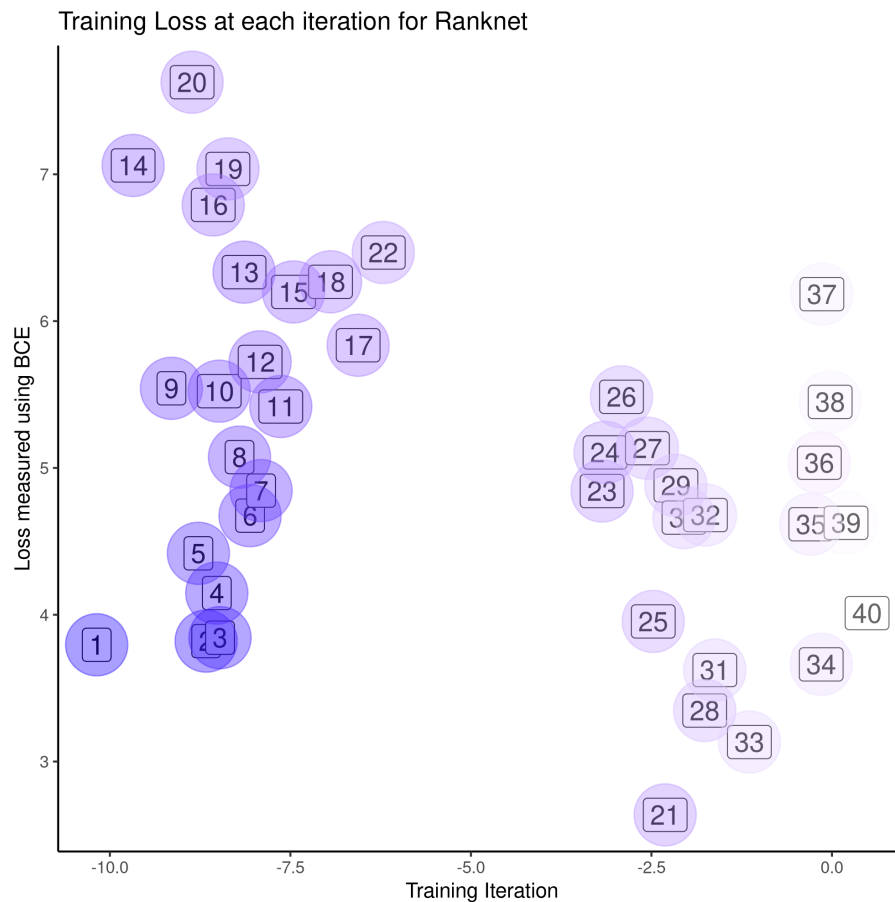

Training Loss at each iteration for Ranknet
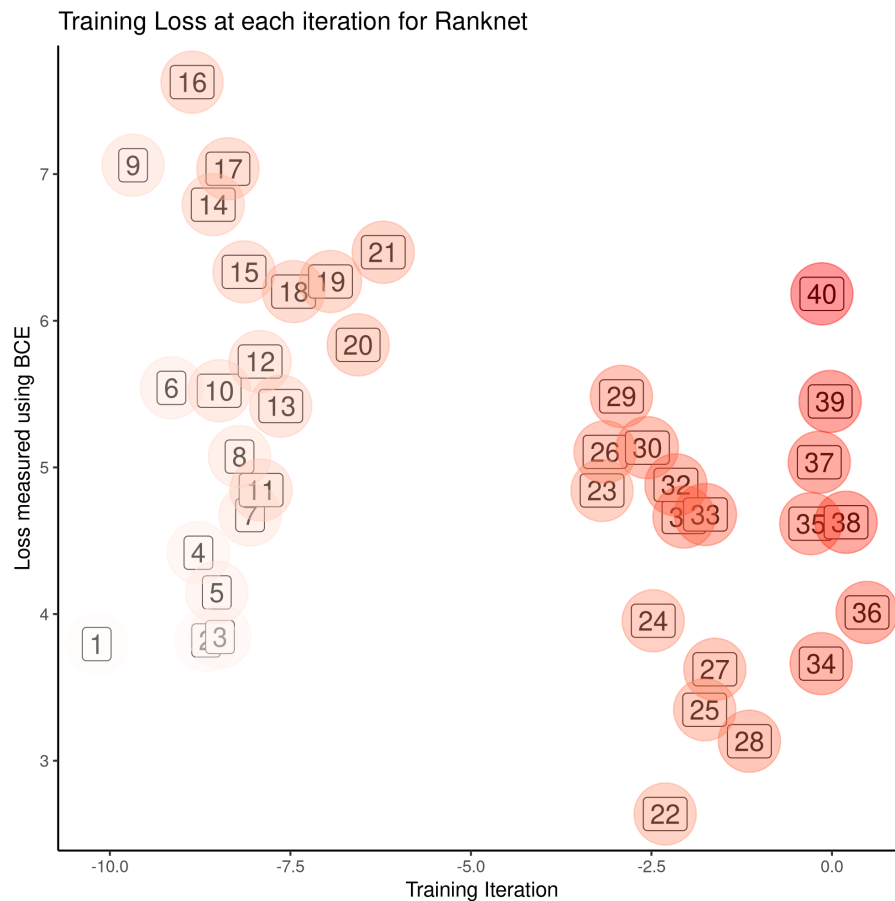
3.6.1   Visualising Model Performance

Although the results in figure appear very encouraging at first site, if the same approach is for an untrained model (i.e. a model with random weights that), a similar

pattern emerges nonetheless, this is shown[14] in figure

Training Loss at each iteration for Ranknet

So instead of ranking, sort the values, this produces the output.

but this is the problem, did it work? it's not clear, because even if the model was not trained we get the following (put them side by side).

So this is definitely one of the hard issues.

what would be better would be to classify data with a rating (i.e. wine scores), only show the model whether the wine is good/bad and compare the output order with the input order, that would be an effective way to see that it works. This was not yet effectively implemented.

3.7   Moons

3.8   Optimisers

3.9   Batches

3.10   Wine

---

[14]For Clarity sake the colours have been inverted.

3.11   Rank Wiki Articles

4   Difficulties

- Don't use torch

  - Do it by hand first because it can be hard to see if the correct weights are being updated sensibly, making debugging very difficult.
  - R or Julia would be easier because counting from 0 get's pretty confusing when dealing with {1, 0}, {-1, 0, 1}.

- Don't use misclassification rate to measure whether the ranking

  - In hindsight this is obvious, but at the time misclassification was a tempting metric because of it's interpretability

  was correct

  Very difficult to see if the model is working

- A continuous function will still produce an ordered pattern in the ranking of results, even if the model hasn't been trained, so visualising isn't helpful either.

- Implement it on a data set that already has order, obfuscate the order and then contrast the results

  - or use a measurement

- Plot the loss function of the training data live, the model is slow to train and waiting for it to develop was a massive time drain.

5   Further Research

5.1   Practical Improvements

- Apply this to documents to get a sorted list, like the wine data

- The "Quicksort" algorithm likely needs a random pivot to be efficient [45]

5.2   Evaluate performance improvements

It is still not clear how the performance of Ranknet compares to traditional approaches implemented by search engines (see §9.1), further study would ideally:

- Write a program to query a corpus of documents using an existing search engine.

  - Or possibly just implement TF-IDF weighting in order to remove variables.

- Extend the program to implement machine learned ranking

- Measure and contrast the performance of the two models to see whether there are any significant improvements.

This could be implemented with TREC datasets [47] using a cummulated-gain cost function [21] as demonstrated in previous work [48].

5.3   Evaluate alternative machine learning models

i.e. can SVM's or trees be used instead of neural networks?

6   Conclusion

7   Text and References

Fractals are complex shapes that often occur from natural processes, in this report we hope to investigate the emergence of patterns and complex structures from natural phenomena. We begin with an investigation into fractals and the concept of dimension and then discuss links between fractal patterns and natural processes.

This is a Reference [46] and another [34] and yet another [6].

8   Fractals

Images are shown in figure .

9   Appendix

9.1   Search Engines

There are many open source search engines available , a cursory review found the following popular projects:

- Zettair (C) [20]

- Apache lucene/Solr (Java) [2]

  - Implemented by DocFetcher [9]

- Sphinx (C++) [53]

- Xapian (C++) [36]

  - Implemented by Recoll [22]

More Modern Search engines include:

- LunrJS (JS) [35]

- Bleve Search (Go) [31]

- Riot (Go) [49]

- Tantivy (Rust) [8]

- SimSearch (Rust) [28]

9.1.1   Fuzzy String Match

Somewhat related are programs that rank string similarity, such programs don't tend to perform well on documents however (so for example these would be effective to filter document titles but would not be useful for querying documents):

- `fzf` [5]

- `fzy` [15]

- `peco` [26]

- Skim [54]

- `go-fuzzyfinder` [25]

- Swiper [24]

9.2   Import Statements

The following import statements were included, where used, [15] separate scripts were used to make the model as modular as possible, such corresponding inputs have also been listed:

```python
1   # Import Packages
2   from itertools import chain
3   from itertools import combinations
4   from itertools import tee
5   from progress.bar import Bar
6   import math as m
7   import matplotlib.pyplot as plt
8   import numpy as np
9   import random
10  import sys
11  import sys
12  import torch
13  import torch
14  from torch import nn
15
16  # Sepereate Scripts lcated below main
17  from ranknet.test_cuda import test_cuda
18  from ranknet.make_data import make_data
19  from ranknet.neural_network import three_layer_ranknet_network
20  from ranknet.quicksort import quicksort
```

9.3   Export Data Method

The data was exported by printing the values to a text file like so:

---

[15]Including `import` statements where they are not used is fine, other than complaints from a *linter* following *PEP* [33] (e.g. autopep [14]) the code will function just fine.

```
1    def export_data(X, y, export):
2        try:
3            os.remove(export)
4            print("Warning, given file was over-written")
5        except:
6            pass
7
8        with open(export, "a") as f:
9            line = "x1, x2, y \n"
10           f.write(line)
11           for i in (range(X.shape[0])):
12               line = str(X[i][0]) + ", " + str(X[i][1]) + ", " + str(y[i]) + "\n"
13               f.write(line)
14       print("Data Exported")
```

9.4   Version Control Repository

The `git` repository used in production of this code is currently available on *GitHub* at github.com/CRMDS/CRMDS-HDR-Training-2020, in order to get a local copy, execute the following commands (`bash`):

```
1    # Clone the repository
2    git clone https://github.com/CRMDS/CRMDS-HDR-Training-2020
3
4    # Change to the subdirectory
5    cd CRMDS-HDR-Training-2020/ranknet
6
7    # Checkout the Walkthrough branch
8    git checkout walkkthrough
9
10   # List the changes
11   git log
```

Consider the use of tools like magit [29] and git-timemachine [40] (or GitLens [1] and git-temporal [3] in VsCode) in order to effectively preview the changes at each step, alternatively a pager like bat [39] can also be used with something like fzf [5] like so:

```
1    git log | grep '^commit' | sed 's/^commit\ //' |\
2        fzf --preview 'git diff {}^! |\
3         bat --color always'
```

9.4.1   Version Control Log for Walkthrough

Commit ID	Message
ed5f4cf	*Initial Commit*
075acf9	*Walkthrough Initial Commit*
cf9ab26	*Generate data to use for classification*
7291112	*Create a Neural Network Model*
7d46636	*Implement gradient descent to train neural network*
f25f376	*Adapt Neural Network to perform Ranking*
42509ab	*Implement sorting algorithm to visualise ranking order*
05df04f	*Adapt Neural Network to perform Ranking*
99b390a	*Implement sorting algorithm to visualise ranking order*
473dce3	*Implement optimizer to replace mere gradient descent*
4141e92	*Train Model using Batches not entire dataset*
a2671a6	*Format code to make it more readable*
d11e607	*plot and only train on different ranked pairs*

References

[1]   Eric Amodio. *Eamodio/Vscode-Gitlens*. Nov. 6, 2016. URL: https://github.com/
        eamodio/vscode-gitlens (visited on 02/20/2021) (cit. on p. 19).

[2]   Apache Software Foundation. *Learning To Rank | Apache Solr Reference Guide 6.6*.
        Solr Reference Guide. 2017. URL: https://lucene.apache.org/solr/guide/
        6_6/learning-to-rank.html (visited on 02/20/2021) (cit. on pp. 3, 17).

[3]   Bee Wilkerson. *Git-Temporal/Git-Temporal: The Mono Repo for Components, Data
        Layer, and Interfaces of Git-Temporal*. Sept. 12, 2018. URL: https://github.com/
        git-temporal/git-temporal (visited on 02/20/2021) (cit. on p. 19).

[4]   Vitaly Bushaev. *Understanding RMSprop — Faster Neural Network Learning*. Medium.
        URL: https://towardsdatascience.com/understanding-rmsprop-faster-
        neural-network-learning-62e116fcf29a (visited on 02/16/2021) (cit. on p. 4).

[5]   Junegunn Choi. *Junegunn/Fzf*. Feb. 20, 2021. URL: https://github.com/junegunn/
        fzf (visited on 02/20/2021) (cit. on pp. 18, 19).

[6]   Christopher Burges. *From RankNet to LambdaRank to LambdaMART: An Overview
        (MSR-TR-2010-82*. Jan. 1, 2010. URL: https://www.microsoft.com/en-us/
        research/uploads/prod/2016/02/MSR-TR-2010-82.pdf (cit. on pp. 9, 17).

[7]   Christopher Burges. *RankNet: A Ranking Retrospective*. Microsoft Research. July 7,
        2015. URL: https://www.microsoft.com/en-us/research/blog/ranknet-a-
        ranking-retrospective/ (visited on 02/15/2021) (cit. on pp. 3, 4, 9).

[8]   Clement Renault, Marin, and Quentin de Quelen. *Meilisearch/MeiliSearch*. MeiliSearch,
        Feb. 20, 2021. URL: https://github.com/meilisearch/MeiliSearch (visited
        on 02/20/2021) (cit. on p. 17).

[9]   Docfetcher Development Team. *DocFetcher - Fast Document Search*. URL: http:
        //docfetcher.sourceforge.net/en/index.html (visited on 02/15/2021) (cit.
        on p. 17).

[10]   Carsten Dominik. *Org Mode Reference Manual*. Place of publication not identi-
        fied: 12TH MEDIA SERVICES, 2018. ISBN: 978-1-68092-165-6 (cit. on p. 5).

[11]   Norbert Fuhr. "Optimum Polynomial Retrieval Functions Based on the Probabil-
        ity Ranking Principle". In: *ACM Transactions on Information Systems* 7.3 (July 1,
        1989), pp. 183–204. ISSN: 1046-8188. DOI: 10.1145/65943.65944. URL: https:
        //doi.org/10.1145/65943.65944 (visited on 02/15/2021) (cit. on p. 3).

[12]   Norbert Fuhr. "Probabilistic Models in Information Retrieval". In: *The Computer
        Journal* 35.3 (June 1, 1992), pp. 243–255. ISSN: 0010-4620. DOI: 10.1093/comjnl/
        35.3.243. URL: https://doi.org/10.1093/comjnl/35.3.243 (visited on
        02/15/2021) (cit. on p. 3).

[13]   Fredric C. Gey. "Inferring Probability of Relevance Using the Method of Logistic
        Regression". In: *Proceedings of the 17th Annual International ACM SIGIR Confer-
        ence on Research and Development in Information Retrieval*. SIGIR '94. Berlin,
        Heidelberg: Springer-Verlag, Aug. 1, 1994, pp. 222–231. ISBN: 978-0-387-19889-7
        (cit. on p. 3).

[14]   Hideo Hattori. *Autopep8: A Tool That Automatically Formats Python Code to Conform to the PEP 8 Style Guide.* Version 1.5.5. URL: https://github.com/hhatto/autopep8 (visited on 02/20/2021) (cit. on p. 18).

[15]   John Hawthorn. *Jhawthorn/Fzy.* Feb. 20, 2021. URL: https://github.com/jhawthorn/fzy (visited on 02/20/2021) (cit. on p. 18).

[16]   HMKCode. *Backpropagation Step by Step.* URL: https://hmkcode.com/ai/backpropagation-step-by-step/ (visited on 02/16/2021) (cit. on p. 4).

[17]   C. A. R. Hoare. "Algorithm 64: Quicksort". In: *Communications of the ACM* 4.7 (July 1, 1961), p. 321. ISSN: 0001-0782. DOI: 10.1145/366622.366644. URL: http://doi.org/10.1145/366622.366644 (visited on 02/21/2021) (cit. on p. 12).

[18]   Daniel Hunt. "A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science". In: (), p. 30 (cit. on p. 3).

[19]   James Aylett. *GSoCProjectIdeas/LearningtoRankStabilisation − Xapian.* Oct. 20, 2019. URL: https://trac.xapian.org/wiki/GSoCProjectIdeas/LearningtoRankStabilisation (visited on 02/20/2021) (cit. on p. 3).

[20]   Dr Rolf Jansen. *Cyclaero/Zettair.* Dec. 22, 2020. URL: https://github.com/cyclaero/zettair (visited on 02/20/2021) (cit. on p. 17).

[21]   Kalervo Järvelin and Jaana Kekäläinen. "Cumulated Gain-Based Evaluation of IR Techniques". In: *ACM Transactions on Information Systems* 20.4 (Oct. 1, 2002), pp. 422–446. ISSN: 1046-8188. DOI: 10.1145/582415.582418. URL: http://doi.org/10.1145/582415.582418 (visited on 02/20/2021) (cit. on p. 17).

[22]   Jean-Francois Dockes. *Recoll User Manual.* URL: https://www.lesbonscomptes.com/recoll/usermanual/usermanual.html (visited on 02/15/2021) (cit. on p. 17).

[23]   Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language.* 2nd ed. Englewood Cliffs, N.J: Prentice Hall, 1988. 272 pp. ISBN: 978-0-13-110370-2 978-0-13-110362-7 (cit. on p. 12).

[24]   Oleh Krehel. *Abo-Abo/Swiper.* Feb. 20, 2021. URL: https://github.com/abo-abo/swiper (visited on 02/20/2021) (cit. on p. 18).

[25]   ktr. *Ktr0731/Go-Fuzzyfinder.* Feb. 16, 2021. URL: https://github.com/ktr0731/go-fuzzyfinder (visited on 02/20/2021) (cit. on p. 18).

[26]   lestrrat. *Peco/Peco.* peco, Feb. 18, 2021. URL: https://github.com/peco/peco (visited on 02/20/2021) (cit. on p. 18).

[27]   Tie-Yan Liu. "Learning to Rank for Information Retrieval". In: *Foundations and Trends® in Information Retrieval* 3.3 (June 26, 2009), pp. 225–331. ISSN: 1554-0669, 1554-0677. DOI: 10.1561/1500000016. URL: https://www.nowpublishers.com/article/Details/INR-016 (visited on 02/15/2021) (cit. on p. 3).

[28]   Andy Lok. *Andylokandy/Simsearch-Rs.* Feb. 15, 2021. URL: https://github.com/andylokandy/simsearch-rs (visited on 02/20/2021) (cit. on p. 17).

[29]   *Magit/Magit.* Magit, Oct. 19, 2008. URL: https://github.com/magit/magit (visited on 02/20/2021) (cit. on p. 19).

[30] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008. 482 pp. ISBN: 978-0-521-86571-5 (cit. on p. 3).

[31] Marty Schoch and Abhinav Dangeti. *Bleve Search Documentation*. URL: http://blevesearch.com/ (visited on 02/20/2021) (cit. on pp. 3, 17).

[32] Mahesh Chandra Mukkamala and Matthias Hein. "Variants of RMSProp and Adagrad with Logarithmic Regret Bounds". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, Aug. 6, 2017, pp. 2545–2553 (cit. on p. 4).

[33] Nick Coghlan. *PEP 8 – Style Guide for Python Code*. Python.org. Aug. 1, 2001. URL: https://www.python.org/dev/peps/pep-0008/ (visited on 02/20/2021) (cit. on p. 18).

[34] Olympia Nicodemi, Melissa A. Sutherland, and Gary W. Towsley. *An Introduction to Abstract Algebra with Notes to the Future Teacher*. Upper Saddle River, NJ: Pearson Prentice Hall, 2007. 436 pp. ISBN: 978-0-13-101963-8 (cit. on p. 17).

[35] Oliver Nightingale. *Olivernn/Lunr.Js*. Feb. 20, 2021. URL: https://github.com/olivernn/lunr.js (visited on 02/20/2021) (cit. on p. 17).

[36] Olly Betts and Richard Boulton. *Xapian/Xapian*. Xapian, Feb. 19, 2021. URL: https://github.com/xapian/xapian (visited on 02/20/2021) (cit. on p. 17).

[37] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (cit. on p. 5).

[38] Jeffrey M. Perkel. "Why Scientists Are Turning to Rust". In: *Nature* 588.7836 (7836 Dec. 1, 2020), pp. 185–186. DOI: 10.1038/d41586-020-03382-2. URL: https://www.nature.com/articles/d41586-020-03382-2 (visited on 02/20/2021) (cit. on p. 3).

[39] David Peter. *Sharkdp/Bat*. Apr. 21, 2018. URL: https://github.com/sharkdp/bat (visited on 02/20/2021) (cit. on p. 19).

[40] Peter Stiernström and Basti. *Emacsmirror/Git-Timemachine*. Emacsmirror, June 29, 2014. URL: https://github.com/emacsmirror/git-timemachine (visited on 02/20/2021) (cit. on p. 19).

[41] Philip Picton. *Neural Networks*. Basingstoke, Hampshire ; New York: Palgrave, 1994. 195 pp. ISBN: 978-0-333-94899-6 (cit. on p. 4).

[42] *Python Program for QuickSort*. GeeksforGeeks. Jan. 7, 2014. URL: https://www.geeksforgeeks.org/python-program-for-quicksort/ (visited on 02/15/2021) (cit. on p. 12).

[43] R Core Team. *R: A Language and Environment for Statistical Computing*. manual. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/ (cit. on p. 5).

[44] Richard Stallman. *GNU Emacs Manual.* 15.ed., updated for Emacs version 21.2. Boston, Mass: Free Software Foundation, 2002. 602 pp. ISBN: 978-1-882114-85-6 (cit. on p. 5).

[45] Tim Roughgarden, director. *Quicksort Overview.* Jan. 28, 2017. URL: https://www.youtube.com/watch?v=ETo1cpLN7kk&list=PLEAYkSg4uSQ37A6_NrUnTHEKp6EkAxTMa&index=25 (visited on 02/15/2021) (cit. on p. 16).

[46] Enmei Tu et al. *A Graph-Based Semi-Supervised k Nearest-Neighbor Method for Nonlinear Manifold Distributed Data Classification.* June 3, 2016. arXiv: 1606.00985 [cs, stat]. URL: http://arxiv.org/abs/1606.00985 (visited on 12/02/2020) (cit. on p. 17).

[47] US National Institute of Standards and Technology. *Text REtrieval Conference (TREC) Home Page.* URL: https://trec.nist.gov/ (visited on 02/20/2021) (cit. on p. 17).

[48] Vik Singh. *A Comparison of Open Source Search Engines.* Vik's Blog. July 6, 2009. URL: https://partyondata.com/2009/07/06/a-comparison-of-open-source-search-engines-and-indexing-twitter/ (visited on 02/20/2021) (cit. on pp. 3, 17).

[49] vz. *Go-Ego/Riot.* ego, Feb. 20, 2021. URL: https://github.com/go-ego/riot (visited on 02/20/2021) (cit. on p. 17).

[50] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org (cit. on p. 5).

[51] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686 (cit. on p. 5).

[52] S. K.M. Wong and Y. Y. Yao. "Linear Structure in Information Retrieval". In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '88. New York, NY, USA: Association for Computing Machinery, May 1, 1988, pp. 219–232. ISBN: 978-2-7061-0309-4. DOI: 10.1145/62437.62452. URL: https://doi.org/10.1145/62437.62452 (visited on 02/14/2021) (cit. on p. 3).

[53] Yuri Schapov, Andrew Aksyonoff, and Ilya Kuznetsov. *Sphinxsearch/Sphinx.* Sphinx Technologies Inc, Feb. 18, 2021. URL: https://github.com/sphinxsearch/sphinx (visited on 02/20/2021) (cit. on p. 17).

[54] Jinzhou Zhang. *Lotabout/Skim.* Feb. 19, 2021. URL: https://github.com/Lotabout/skim (visited on 02/20/2021) (cit. on p. 18).