What is the Difference Between Data Science and Machine Learning?

Geoffrey E. Schneider

June 21, 2017

1 Introduction

Data science is the practice of extracting useful information from data. Machine learning is a set of programming techniques in which the programmer sets up a model with variable parameters and a method by which these parameters will be ”learned” from data. There. Question answered. In practice, however, when searching for a job, it isn't so clear. Because of the large overlap between the two, a job listed as a data science job may really be more of a machine learning job, and vice versa. And say you want to pick up a few new, relevant skills, but you are primarily interested in one of the two fields. Can we disentangle which skills are more relevant to your chosen field?

The goal of this project is to use textual analysis to distinguish between these two fields. Code is written in the new and exciting programming language Julia. This document was produced using LaTeX cite these.

2 Data

The data for this project is 150 posted job descriptions. They are divided into three parts, the machine learning descriptions (50), the data science descriptions (50), and the control descriptions (50). These were gathered by doing searches for the relevant words (”machine learning job”/”data science job”/”job”), and then divided into sequences of words using regular expressions. In Julia this is implemented as:

```
function texttowords(text)
  matchall(
    r"(\w+)",
    lowercase(replace(text, r"([\r\n,\.\(\)!;:\?/]|\ufeff)", s" "))
    )
end
```

The sequences of words were then divided into three ”bags of words” (i.e. word counts) for each of the search terms.

```
function wordcount(wordvec)
  worddict = Dict{String,Int64}()
  for w in wordvec
    worddict[w]=get(worddict,w,0) + 1
  end
  worddict
end
```

3 WordNorm

To compare two bags of words, we will use two variations on what we will call "WordNorm" (non-technical folks can skip to the tldr at the end if this section, if you want). For word counts α and β (e.g. $\alpha =$ data science and $\beta =$ control) and word w, the ideal version of WordNorm

$$\mathrm{WordNorm}_{\alpha\beta}(w) = \frac{\alpha(w) - \beta(w)}{\alpha(w) + \beta(w)}. \tag{1}$$

We will write $\mathrm{WordNorm}(w)$ when the choice of α and β are clear or irrelevant. Note that

$$\mathrm{WordNorm}(w) = \begin{cases} 1 & w \in \alpha \setminus \beta \\ -1 & w \in \beta \setminus \alpha \end{cases}$$

and that 1 and -1 are the maximum and minimum values of $\mathrm{WordNorm}$. So, if w is near 1, it is much more closely associated to α, and if w is near -1, it is much more closely associated to β. In an ideal setting, with a very large amount of data (so that each word count has at least one of each word), this would work well, however in a practical setting there is an issue. For example, suppose we're interested in comparing the word "dog" to the word "cat", and α("dog") $= 5$, α("cat") $= 1$, but β("dog") $= \beta$("cat") $= 0$. $\mathrm{WordNorm}$ does not distinguish between "dog" and "cat"! $\mathrm{WordNorm}$("dog") $= \mathrm{WordNorm}$("cat") $= 1$, despite the fact that it is clear that α is probably more strongly associated with "dog" than it is with "cat". To deal with this problem, we will use the two modifications that follow.

When comparing a sample α (e.g. the counts of words in machine learning job descriptions) to a control sample (which we write as c) we use

$$\mathrm{WordNormCtrl}_{\alpha}(w) = \frac{\alpha(w) - (c(w) + 1)}{\alpha(w) + (c(w) + 1)} \tag{2}$$

where we've written the parentheses to emphasize the interpretation that $\mathrm{WordNormCtrl}$ is just $\mathrm{WordNorm}$ where we've added one of each word to the control sample. Now using β in the example above as a control, we get $\mathrm{WordNormCtrl}$("dog") $= \frac{2}{3}$ and $\mathrm{WordNormCtrl}$("cat") $= 0$. For our setting $\mathrm{WordNormCtrl}$ has the nice feature that common words ("is", "the", etc.) and general job words ("applicant", "responsibilities") get a score close to 0, and so when we are looking at the top scores, these are automatically ignored. It loses the (anti)symmetry of $\mathrm{WordNorm}$, but this is acceptable, because the sample and control are playing different roles here.

When comparing two samples, we use

$$\mathrm{WordNormComp}_{\alpha\beta}(w) = \frac{\alpha(w) - \beta(w)}{\alpha(w) + \beta(w) + 1}. \tag{3}$$

This is equivalent to adding half a count of each word to both samples. In a sense, we are saying that we expect that if $\alpha(w) = 0$, the "real value" of $\alpha(w)$ is between 0 and 1, but we need a bigger sample to determine the value, so we guess that it is 0.5, but then to be fair to the other words where $\alpha(w) > 0$, we also add 0.5 to their counts. Here the (anti)symmetry is restored, as it should be since both samples play interchangeable roles.

[*tldr: There are a few variations of WordNorm, but in all cases, it assigns a value between 1 and -1 to each word. 1 means the word is much more closely associated to the first bag of words, and -1 means the word is much more closely associated to the second. When one of these is the control, that will always be the second.*]

4 Results and Discussion

In Figure 1 we show the results of applying $\mathrm{WordNormCtrl}$ to data science (vs. control) and to machine learning (vs. control). We see that the two give broadly similar scores validating the assertion that there is much overlap between the two. Languages such as R, Python and Spark, as well as the distributed system framework Hadoop appear in the top 10 for both.

For "visualization" we see a higher data science score, confirming the stronger importance of communicating findings to allow companies to make informed business decisions, whereas the higher machine learning score for "ai" indicates the importance of building automated systems as a product in themselves. The word "academy" appears at first to be strongly associated with data science and strongly dissassociated with machine learning, however, it is an aberration which should be ignored, as it is on this list only because it appeared a very large number of times in one particular data science job description, and not at all in the control and machine learning data.

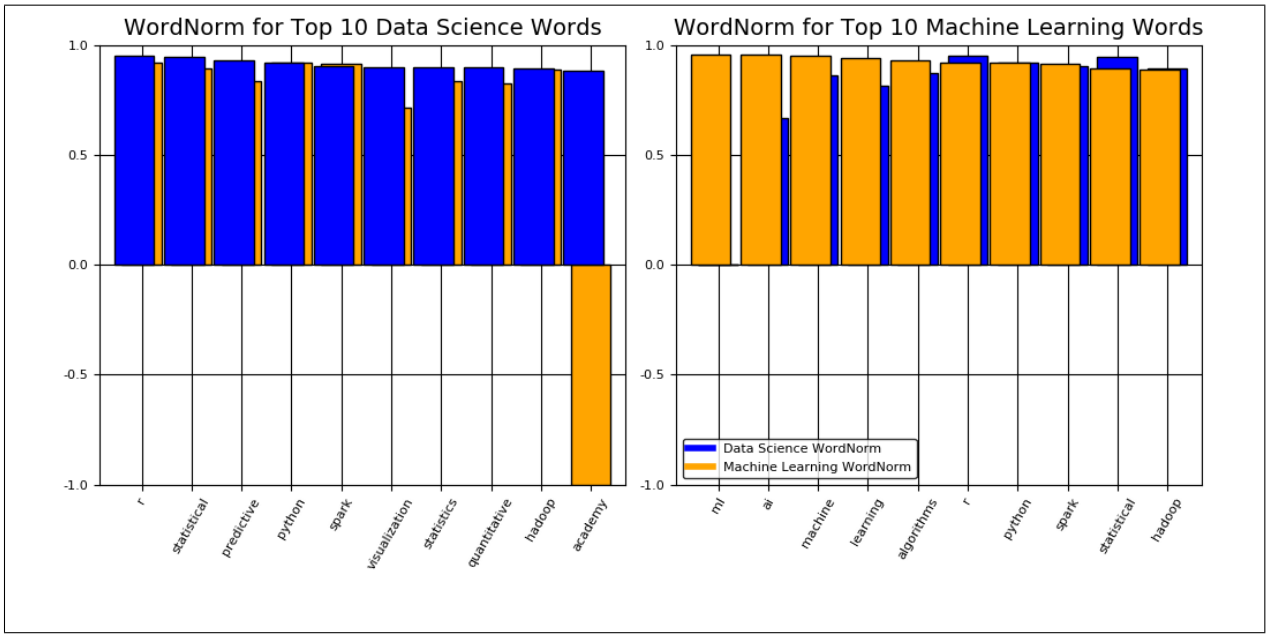

Figure 1: The words with the highest $\mathrm{WordNormCtrl}$ scores for both data science and machine learning.

Comparing the data science and machine learning data to the control tells us a lot about what

is important to each of them, but due to their large overlap, it does not allow us to *distinguish* them. To do this, we compare them directly using WordNormComp. Figure 2 shows the result of applying $\text{WordNormComp}_{ml,ds}$ to directly compare machine learning to data science.

Machine learning frameworks TensorFlow and Caffe are both represented with a stronger association with machine learning whereas words like "consumers" and "audiences" are associated with data science. We also see the names of companies that are associated with machine learning or data science. These company names may be due to the inclusion of jobs advertised by those companies in the dataset, though this is not entirely clear in the case of software companies, where the inclusion may be due to requirements that applicants have knowledge of their software.

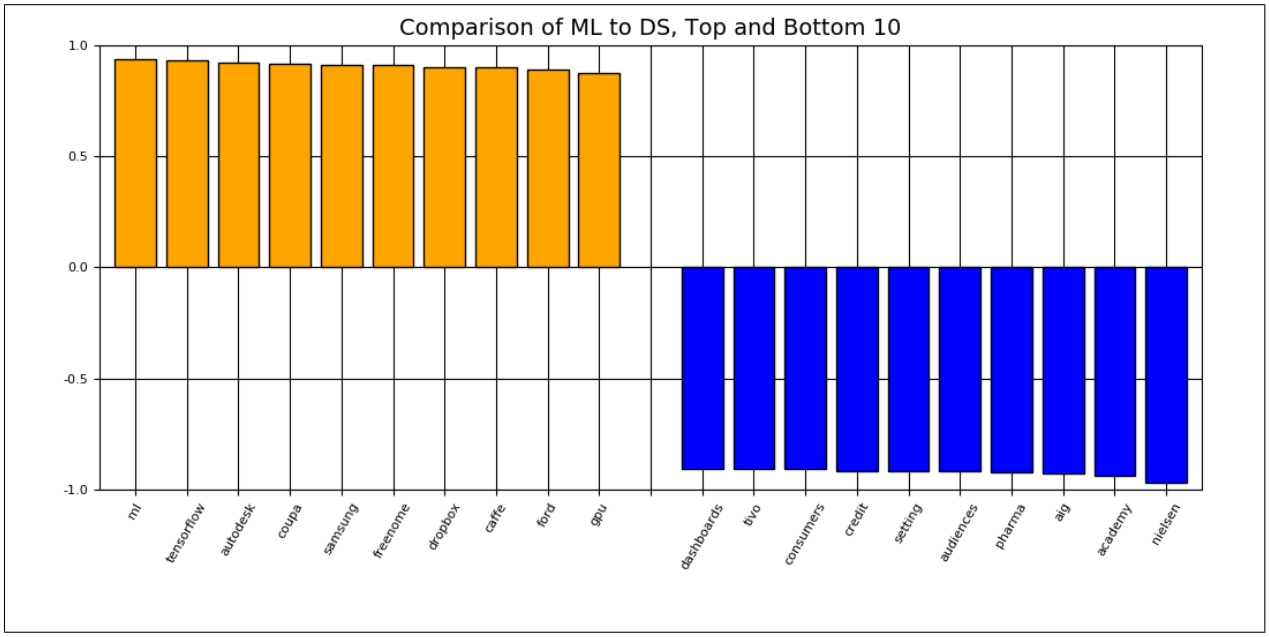

Figure 2: The result of comparing machine learning to data science with WordNormComp. Positive scores indicate stronger association with machine learning, and negative scores indicate stronger association with data science.

Using the WordNormComp score, we can score any text in terms of whether it is more closely associated with data science or machine learning. For example, the Insight Data Science Fellows Program has a white paper with a section titled "What is a Data Scientist?". We can score this text, word by word, and then average to get a score of -0.13 indicating a closer association with data science than with machine learning.

Figure 3 shows the first 40 words of this document, each colored according to their WordNormComp score.

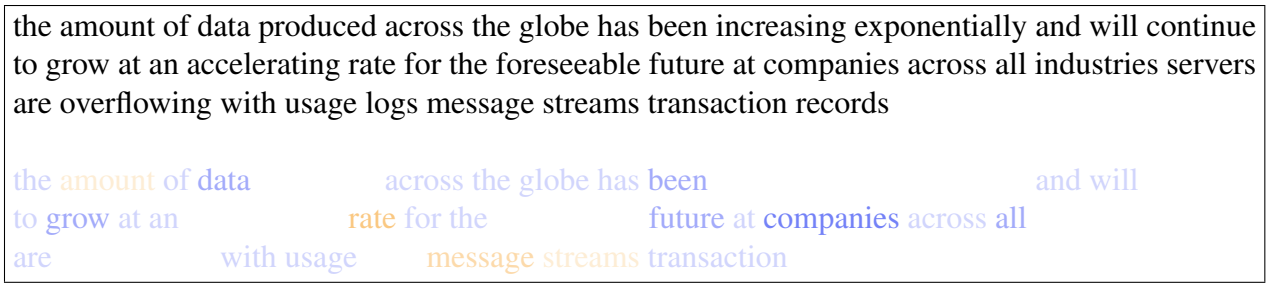

Figure 3: The first 40 words of "What is a Data Scientist?" from the Insight Data Science Fellows white paper, first uncolored, then colored according to its WordNormComp score. Data science associated words are blue, machine learning associated words are in orange, and lighter means closer to 0.