Machine Learning: algorithms, Feedback on reports

Nicky van Foreest

June 2, 2021

Contents

1 Goal of this feedback session **1**

2 Example 1, title, abstract, intro **1**

3 Example 2: data analysis **5**

4 Example 3: methods **7**

5 Example 4: Analysis **8**

6 Example 5: Conclusions **9**

7 My conclusions **11**

1 Goal of this feedback session

- A master learns the skill of assessing his/her own work.
- A master develops his/her own standards
- I'll present examples for each part of the report. I took 5 reports, at random, and assembled feedback on a specific section part of each report. So, for the intro I used one report, then I used another for the data analysis section, and so on.
- I'll give you some time to *write down* your opinion and suggestions for feedback.
- Then I'll give the points that I think can be improved.

2 Example 1, title, abstract, intro

2.1 Title: your opinion, suggestions for improvement?

'Prediction of Gender using Email-Addresses'

2.2 My feedback

- I like the title, nice and short. Clear.
- The title helps me to decide whether I have (would like) to read this report.

2.3 Abstract: opinion, suggestions for improvement

'In this report, we try to predict gender based on the email address of an individual. Several variables, such as name guesses and sentiment scores, are added based on the email addresses. Three different classification methods are implemented and fine-tuned, namely the Logistic Regression, Decision Trees and Random Forest. The best method turns out to be a Random Forest, which is able to achieve an accuracy of 71.36%.'

2.4 My points

The abstract does not tell me:

- Why do I want to know this? Where is the motivation?
- What is the data set used? Did the authors make it themselves? Otherwise?
- Three methods. Good. Why these? Are these methods specifically suited for this? Why not AdaBoost (for instance)?
- An accuracy of 4 digits? That is insane. Use a decent number of digits, perhaps 2. There are very little things we can estimate with a precision of 1%. Better yet, include the variance.

It's actually a good abstract for version 0.1: it's easy to change into a complete abstract. And, it is already short and clear.

2.5 Intro 1st para. opinion, suggestions for improvement

'The importance of gender equality has grown over the past decades. This increase shows, for example, importance in the economic environment. Improved gender equality results in increasing benefits on economic growth such as a higher productivity and employment rate (Maceira, 2017), but also on a smaller scale such as benefits in mental health and psychological well-being (Connell, 2003). Unfortunately, the goal of gender equality is far from reached. The problem occurs even on a daily base by mispronouncing. Mispronouncing is an act of misgendering people through the use of pronouns such as 'she' and 'he' (Ansara & Hegarty, 2014). Even though the problem seems small, the effect is large. As shown by Ansara and Hegarty, 2014, women that were given male pronouns during a job interview, felt less and less identified for the job along with a decrease in their motivation. An other interesting research in this field describes a gender pronoun reversal activity executed among students to encourage empathy for transgender people (MacNamara et al., 2017). In this exercise, the students were intentionally misgendered by using the pronouns of the opposite gender. Afterwards students reported experiencing a wide variety of emotions to undergoing and witnessing gender non-conformity during the exercise, such as embarrassment, amusement, bewilderment, confusion, dissonance, and guilt. ...'

2.6 My points

- Overall very good.
- Somewhat long, for me too long. I don't need 3 reasons to support the same claim, at least not in a report of this size and type.

(BTW, with my first name I have often been addressed as a woman, but I don't care about this. :-))

2.7 Intro: 2n para, opinion, suggestions for improvement

'Reynolds et al., 2006 show that people cannot break the stereotypical gender of a character, such as the gender of a surgeon, suggesting that misgendering still occurs very often at a individual level. This also suggests that the probability of misgendering exists in data bases. If this is the case, (automatic generated) emails and phone centers will, accidentally, misgender customers. If this happens frequently, the customer will probably lose its connection to the firm and henceforward leave the company resulting in high costs. This paper, therefore, tries to develop a machine learning approach to find misgendering in data bases in order to avoid large costs obtained from losing customers.'

2.8 My points

- Merge with the first para. The relation with data bases is nice, but it is bit too long for me.
- When readers start to get bored, they'll start skipping what you have written.
- When readers skip what you write, should you invest the effort into writing it?

2.9 Intro: 3rd para, opinion, suggestions for improvement

'One could argue whether this problem can be avoided by using gender neutral pronouns and salutations such as 'Dear reader'. However, this gives a signal that the firm is not interested in their customers/clients, which will not contribute to a positive and loyal connection between customers and the firm (Magids et al., 2015). Therefore, even though using gender neutral pronouns and salutations will avoid misgendering, more personal pronouns and salutations are preferred resulting in the problem mentioned before.'

2.10 My points

- Great first sentence. This answers exactly my implicit question.

2.11 Intro: 4th para, opinion, suggestions for improvement

'The goal of this research is to avoid misgendering, the methods discussed will be applied on the username of a customer's email address in order to predict his/her gender. Note, however, that this approach is not a substitution of the standard approach to obtain the gender of a customer, but rather an approach to check for mistakes made in the data base. For this reason, we will only consider the genders male and female and will leave out the option of

'other' or 'rather not say' that is nowadays a common option that is indicated on application forms. The suspected errors can be checked by sending this subset of clients a request to check and update their personal information in the system if necessary.'

2.12 My points

- I get the point.
- Why discuss how to filter the data in the intro?

2.13 Intro: 5th and last para, opinion, suggestions for improvement

'The report has the following structure. After having introduced the data set in Section 2, we describe in Section 3 the methods to analyze and validate the data. The main performance of our methods are reported in Section 4. Finally, in Section 5, the main findings of the analysis are summarized and potential extensions are discussed.'

2.14 My points

- OK. Nice overview.
- Why the shift in style from "we describe" to "are reported"?
- I dislike such sudden switches from active to passive writing.
- I prefer the active voice.

2.15 Overall, comments, suggestions?

2.16 My overall comments

- Why not discuss the used methods in the intro?
- There is no discussion at all of the followed methodology? I know that you will do it in Section 3, but (to put things very bluntly) I don't want to wait until section 3 to form an opinion of the report.
- There is no mention of how successful this approach is. (It's in the abstract, sure, but if the reader skipped that, then the intro does not help). Give some idea on the results the report delivers.
- I often tell me master thesis students that they can write as much as they like, but *I* only read what *I* like.
- The more they write, the more I skip.
- I also tell what parts I skipped.
- Repetition in text makes me (and the rest of the world too) impatient.
- Hence, I do what *you do when getting bored*: I stop reading, and spend my time on other things I like (more).

- So, all of you should also become really honest and clear about this: if you get bored, you stop.

Assessment:

- For version 0.1 it is already a real good intro. Even if this would be the final version, I would grade it with a 6 or 7.
- However, it's easy to improve to a 9 or 10. Summarize yet more: merge para's 1, 2 and 3 into just 1 paragraph. This gives room for a discussion of the methods used, and why you used these, and not others. This also gives room to lift a tip of the veil on the type of results you will deliver.

3 Example 2: data analysis

3.1 data analysis: 1st para, opinion, suggestion

'In order to be able to predict bankruptcy for a company, we will use a data set from the Taiwan Economic Journal. It is available via the UC Irvine Machine Learning Repository1, and the version we use was downloaded on April 29th, 2021. The data were collected for the years 1999 to 2009, and there are 95 features available, that cover a majority of the financial attributes. Besides, it contains a variable which equals 1 if the corresponding company went bankrupt, and 0 if not. The financial attributes in the data give insightful information about the performance of a company, and therefore we think that this data will be helpful in constructing a model that predicts bankruptcy.'

3.2 My points

- What's the plan of this section? Always spend a few sentences on what the reader can expect.
- "The ... insightful ... ": what if is was not insightful? Besides, it's the task of the algorithm to figure this out. As a general advice: remove text that's obvious, or irrelevant.

3.3 data analysis: 2nd para, opinion, suggestion

Before we modify the data such that it is suitable for the methods we want to use, we have to check whether there are features which contain data points that differ significantly from other observations. To inspect for such outliers, we make a boxplot of all 95 features that are in the data. This boxplot can be found in the appendix. We see that there are some features which contain data points that are located outside the whiskers of the boxplot. Some of these features even contain observations with very extreme values. Furthermore, we notice that some features do not have a maximum value of one, while they are ratios, which implies that they should be between zero and one. Since it is hard to determine whether or not an observation is an actual outlier or merely an extreme value, we decided to keep all observations.

3.4 My points

- Overall OK: the authors appear to be aware that data can contain errors. It is important to include this, as a simple check point. If authors do not show this awareness, why believe their conclusions?
- There is a decision about how to handle outliers. But, *what are the consequences of this decision?* It appears that the authors don't analyze this, thereby making their final conclusion less credible. In particular, as they say that they have seen extreme outliers, they should take out the suspected outliers, and redo the analysis.

3.5 data analysis: 4th para, opinion, suggestion

As one can see in Table (1), we have highly imbalanced data, where only 3% of the observations is an actual bankruptcy. Due to this imbalance, there are too few examples of the minority class in order for a model to properly learn the decision boundary. In Section 3 below, we will explain how we used the *Synthetic Minority Oversampling Technique* in order to account for this.

3.6 My points

- What would you ratio would you expect for bankruptcy? 30%? This is a bit high... That the data will be 'imbalanced' is what one expects, hence, the authors have to deal with this problem.
- '... too few examples, ...', Why? What is too few? And if the authors conclude this, then they should explicitly deal with this in the conclusions, because it must affect the credibility of the findinds. And if if does not affect the credibility, the data is not too few.
- The authors write that they will explain *how* they will use SMOT. But, they should first explain *Why* the use of SMOT is OK? Why is it suitable for this dataset? (If the *why* question is not covered, I don't care about the *how*)

3.7 Data analysis: 1 page table, opinion, suggestions?

- X1 0.505 0.061 0.000 0.477 0.503 0.536 1.000
- Total Asset Growth Rate 5.790000e+09 7.690000e+09 8.310000e+09 6.110000e+09 4.880000e+09

3.8 My points

- What is the meaning of the symbols?
- More importantly, why use (waste?) one full page on tables that (i think) nobody is going to read in detail?
- Use the space you get for more useful things.

4 Example 3: methods

4.1 Methods, 1st para

This section describes the methods that are used and compared to predict Alzheimer's disease. Since the dependent variable is a binary variable that can only lie in a finite set, we use classification methods. Classification methods come down to predicting the outcome variable by classifying the dependent variables into one out of the 2 categories (demented/nondemented) and determines the performance using a loss function. The most commonly used loss function for classification problems is Loss(y, $\hat{y}$) = 1($\hat{y}$ 6= y), where $\hat{y}$ is the predicted class and y is the real class. This loss function incurs a loss of 1 when the predicted class is not equal to the real class. The aim is to predict Alzheimer's disease as accurately as possible, and hence to minimize the expected loss function. The main classification method used for this research is K-nearest neighbor (K-NN) classification. This method will be described more extensively below. The classification results of the K-nearest neighbor algorithm will be compared with the following classification methods:

1. The logistic regression
2. The random forest method
3. Support vector machine

k-fold cross-validation is used to find the optimal hyper parameters of each model, and to test the performance of the different methods. The method with the highest prediction accuracy is considered as best performing.

4.2 My points

- Why no section number?
- Nice first sentence.
- Why one big paragraph? Typically, use one idea per paragraph. Split up important ideas.
- Why these three methods?
- The last sentence, what does it add? Do the authors also consider to advice to use a method with a lower accuracy?
- The file name is just Bayesian statistics.pdf. Get the file name right, i.e., according to the specs.

4.3 Methods: 3rd para, opinion, suggestions?

The K-nearest neighbor classification method has some limitations. For example, for large data sets this method could be slow since the calculations are repeated for each K. This makes this method computationally expensive. Besides, the K-NN suffers from the curse of dimensionality when a large number of features is used for prediction. However, the data set used contains 373 observations and 9 features and hence the computations are considered

doable. Moreover, the K-NN method is sensitive to outliers, since the nearest neighbors are chosen based on distance criteria. In addition, the K-NN method is not capable of dealing with missing values. However, imputation is used to resolve this issue.

4.4 My points

- The method is sensitive to outliers. OK. But, what are the consequences with respect to the findings? Do the authors address this. If not, why not? If yes, how, and *where*?

4.5 Methods 4th para

Besides, cross-validation is a useful method for assessing the effectiveness of a machine learning model. Hence, in addition to finding the optimal hyper parameters using cross-validation, we will also use it to test the performance of the model. k-fold cross-validation is a method that provides data for training and data for validation without wasting any data. Every data point is used once as validation data and $k-1$ times as training data. Since all the data is used multiple times for fitting the model this reduces the bias. Besides, since all the data is used for validation, the variances reduces. K-fold cross-validation is the preferred method since it is highly effective, but still computationally doable.

4.6 My points

- "... preferred...". Refer to a book/article to motivate why this method is preferred.

5 Example 4: Analysis

5.1 Analysis 1st para

The results obtained from the methods described in Section 3 are discussed in this section. All our findings are summarized in Figure 2, where we find the performance of the different prediction models divided over several classification (error) categories.

5.2 My points

- Nice overview, good start
- The authors use one figure to illustrate. I find this a very good idea (if this is possible.) It gives the section a good, consequent structure.
- Ensure that the figure is aligned/supports the main KPIs you want to address in the report. In other words, does the figure help answer the research question?

5.3 Figure

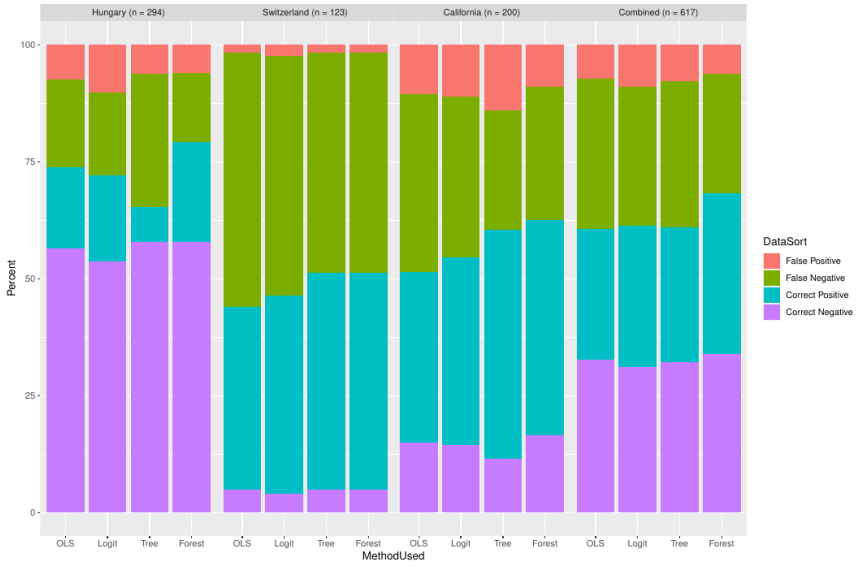

Figure 2: The results of the different methods on different (sub)sets of the validation data set

5.4 My points

- Quite some people read figures first, while browsing through the text. At first they simply skip the text, and use the figures to decide whether they will read (part of) the text. If they don't like the figure, or don't understand, they'll stop.
- Therefore, in figure captions, explain/summarize the 'story' of the figure and the main findings (What can I see in this figure?).
- It is unclear from the caption what the authors find the most important KPI. For instance, if false positives are the most important, the figure gives the answers right away, and, later, the figure can be used to look up things or in presentations. Hence, investing time in making good figures pays off.
- All in all, it's a great figure. It's not a 10 yet, but easy to get it at that level

6 Example 5: Conclusions

6.1 Conclusions, opinion, suggestions?

In this paper, we have applied different machine learning techniques on how to tackle the problem of predicting an infection with Hepatitis C. In doing so, we used a data set previously used by Hoffmann et al. (2018), which includes a number of different blood attributes of patients that may indicate whether a patient is infected with HCV or not. While a a naive prediction of always healthy would already have an overall accuracy of 85.9% for our data, we are able to rise this number to 95.72% by using Logistic Regression with a 3-order polynomial of the features and a Ridge penalty in the objective function.

6.2 My points

- Writing is a bit clumsy: "In this paper, we have applied different machine learning techniques on how to tackle the problem of predicting an infection with Hepatitis C." Alternative: We have applied different 4 (be specific) machine learning techniques that use 10(?) features in blood samples to predict whether a patient is infected with HCV.
- What is a naive prediction? (I think I know, but I am on the same page as the authors here?) Be specific.
- Rest is nice and clear.
- It turns out that the conclusion is spread over multiple pages, mixed with tables and figures. That makes a bit hard to find for later reference. But, what again is what people are looking for? ...
- Also, some people jump right away to the conclusions, and skip the rest at first. Ensure that it's easy to find the conclusions.

6.3 Conclusions, last para

In summary, we have shown that different machine learning techniques can help to detect or rule out an infection with Hepatitis C at a low cost. Instead of conducting complicated and costly tests for the disease, with machine learning models one only puts a set of blood measures from a patient into an estimated model and gets the prediction almost immediately. Still, we encourage future research to build larger data sets with a higher number of infected people in order to improve predictions for sick patients.

6.4 My points

- Good summary, good motivation.

6.5 Bibliography

- Geurts, P., A. Irrthum, and L. Wehenkel (2009). Supervised learning with decision tree-based methods in computational and systems biology. Molecular Biosystems 5(12), 1593–1605.
- Gunčar, Gregor, Matjaž Kukar, Mateja Notar, Miran Brvar, Peter Černelč, Manca Notar, and Marko Notar (2018). An application of machine learning to haematological diagnosis. Scientific

6.6 My points

- Inconsistent naming: some times first names, sometimes not.
- It's really easy to get the bibliography correct. Some people use it to check how serious the authors are. If the authors don't manage to get the trivial things right, why would they fare better with the harder things?
- Sloppiness decreases the credibility of the report.

7 My conclusions

- The level of the reports that I checked are already quite good.
- There are many simple points (low hanging fruit) in which the reports can be improved. It's not much work, but improves the appearance of the reports a lot.
- Learn to be critical wrt your own work, in other words, learn to read your work 'through the eyes of somebody else'.
- Be honest: you don't like to read reports of other people, hence other people will not like reading your work. Hence, keep readers involved.