Group Assignment: Applied Deep Learning – Dr. Philippe Blaettchen

Assignment overview

When using machine learning for classification, things are easiest if classes are "balanced" – that is, when the number of observations belonging to each of the classes are of the same order of magnitude. Unfortunately, this is often not the case. In this assignment, you will work with a dataset of car-insurance claims and try to classify claims into fraudulent (1) and non-fraudulent (0). There are more than 10,000 claims in the dataset, but only around 100 are fraudulent. Nevertheless, we want to create a model that helps the insurance provider target its investigation efforts.

Learning objectives

- Understand how to evaluate machine learning models in the context of unbalanced datasets and based on the specific application.
- Evaluate models with different degrees of complexity.
- Learn about the broad spectrum of uses for autoencoders and how they can be applied to detect anomalies.
- Gain intuition on how to train neural networks.

Task description

1. Carefully read the case "To Catch a Thief", in the Week 5 materials on Moodle. Shift's fraud detection algorithm flags claims that seem suspicious, which are then investigated by human claim handlers. A key metric of success of the algorithm is the "hit rate" (or conversion rate). However, there is a trade-off between the hit rate and the detection rate. In under 300 words, describe the two metrics, the trade-off between them, and how you would evaluate a model with hit rate *h* and detection rate *d* in monetary terms.

2. Load the dataset "Insurance_claims.csv" (the data and data lexicon are also available in the Week 5 materials on Moodle). Pre-process the data set appropriately for the use case.

3. Consider first a simple model: a claim is suspicious if the incident occurs within *t* days of the subscription date. What are the hit rate and detection rate of this model for different values of *t*? How do you choose a sensible value for *t*? What logic underlies this model?

4. Consider next an intermediate model: in particular, implement a decision tree classifier to identify cases of fraud. What are the hit rate and detection rate of this model? Be sure that you evaluate the model on a separate test set!

5. Consider next a complex model: in particular, implement and tune a (deep) neural network in TensorFlow. What are the hit rate and detection rate of this model? Be sure that you evaluate the model on a separate test set!

6. Because we have only limited data, we might want to use anomaly detection rather than predicting fraud labels.
   a) Create a training set that contains only non-fraudulent claims, as well as validation and test sets that contain non-fraudulent and fraudulent claims. Make sure to spread fraudulent claims evenly (but randomly) across validation and test sets.
   b) Using TensorFlow, create an autoencoder. Because your training set only contains non-fraudulent claims, the autoencoder should learn to represent those well. A good metric is the mean squared error over all features between the true data and the reconstructed data.
   c) Plot the errors between prediction and true data for your validation set in a histogram – in a good model, this error should be much higher for fraudulent claims than non-fraudulent ones. Be sure to mark the different types of claims in your histogram.
   d) If the distinction is not clear in your histogram, you may want to fine-tune your model. Be careful in your choice of metrics.
   e) Once you have identified a final model, set a threshold on the error between actual and predicted data. That is, if an observation has a bigger error than the threshold, you predict it is a fraudulent claim. If the error is below the threshold, you predict it is non-fraudulent.
   f) Using your model and your threshold, find the hit rate and detection rate. Be sure that you evaluate the model on the test set!

7. Discuss the different approaches you have attempted, in terms of hit rate and detection rate. Also, consider the transparency behind each approach. Why can it be problematic when there is no clear explanation why the tool flagged a certain incident as possibly fraudulent? Use less than 300 words for your answer.

8. What other approaches dealing with imbalanced data can you envision to improve Shift's prediction of insurance fraud? Use less than 150 words for your answer.

Hints

- Don't spend too much time perfecting the model in part 5. What is most important at this point is to understand the issues faced by training a model with such an uneven dataset and to explore a range of models. The performance of your model will not matter for evaluation as long as you are able to beat the decision tree classifier (meaning you obtain a higher hit rate for the same or a higher detection rate).
- In your autoencoder (part 6), make sure that your network can recreate your inputs. For example, if your output layer uses tanh activation, the outputs would be between -1 and 1. Remember to scale your inputs appropriately.

- For part 6, start by creating a minimum viable product, keeping in mind how we defined a useful model. Only once everything runs should you go back to it and see how to improve your model. The performance of your model will matter for evaluation, but not as much as having a complete answer.
- The same goes for cleaning the dataset: combine rare categories, don't overengineer new features, and instead start with only a subset of the existing ones. Then, once everything works smoothly, you can think about enhancing your dataset.

Materials to submit

- A Jupyter notebook that allows recreating your solutions. If you have tried and abandoned different approaches, clear those out but make sure to describe your process.
- The trained models you develop in parts 5 and 6 as .h5-files
- Your written answers, either within the Jupyter notebook or in a separate .pdf-file. Make sure to create numbered sections within your notebook and separate .pdf-file corresponding to the task at hand (1-8).
- A PowerPoint slide deck of up to 3 slides comparing the different approaches you have developed and discussing the main challenges you faced. Several groups will be selected randomly to present their work in the next class. The selected groups will be informed by 8 pm on the day of submission. Please keep your presentation to under 5 minutes.

Assessment

Your submission will be evaluated against four criteria:

- appropriate use of concepts and frameworks discussed in class;
- effectiveness of the proposed answer/solution;
- originality and creativity of the proposed answer/solution;
- organization and clarity of submitted materials.