Mariana Carini, Karl Hickel, Caesar Phan, Matthias Ronnau

Advanced Machine Learning-290

Professor Bojd

June 15, 2021

Final Project

Part 1

A.) In our current models, we are emphasizing the importance of measuring recall. The purpose for this is that our focus is to minimize the number of false negatives in our model. Our goal is to predict whether an individual has heart disease, and in this context a false negative would be an individual who we say does not have heart disease but in fact does. These individuals would miss out on needed care and treatment; this is a life-threatening problem that we want to minimize.

B.) A certain test that measures coronary artery calcium has a specificity of 87%.[1] This means that of all individuals classified as not having heart disease, 13% were incorrectly classified. We are interested in this number for the above reason, as an incorrect negative diagnosis can prove dangerous and deadly for these individuals.

C.)

Model	**Training Loss**	**Training Recall**	**Training Precision**	**Training True Positives**
Initial	0.665	0.376	0.659	6733
1	0.665	0.381	0.663	6817
2	0.664	0.384	0.665	6870
3	0.664	0.388	0.665	6950
4	0.664	0.391	0.663	7011
5	0.664	0.400	0.658	7168
6	0.665	0.386	0.657	6921
7	0.666	0.375	0.666	6711
8	0.666	0.375	0.661	6718
9	0.666	0.382	0.660	6851
10	0.666	0.376	0.664	6747

[1] Hanifehpour, Reza, Marzieh Motevalli, Hossein Ghanaati, Mona Shahriari, and Mounes Aliyari Ghasabeh. "Diagnostic Accuracy of Coronary Calcium Score Less than 100 in Excluding Coronary Artery Disease." Iranian journal of radiology : a quarterly journal published by the Iranian Radiological Society. Kowsar, March 20, 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5035795/.

Generally speaking, no, our model did not significantly improve with the addition of new layers with varying dense sizes. The model never saw a decrease in its loss lower than 0.664 when adding new layers. The recall also did not see much improvement as it only maintained its general trend of straying between 0.37 - 0.40. Precision sees some level of improvement topping off at around 0.666. However, our main goal again is to maximize our recall without the expense of precision.

D.)

Optimizer	Training Loss	Training Recall	Training Precision	Training True Positives
RMS	0.640	0.596	0.645	10673
SGD	0.666	0.364	0.664	6516
Adamax	0.664	0.404	0.659	7239

RMS prop appears to have had the biggest impact on recall. Benchmarking against our best baseline model (initial model 5), RMS prop increased the training recall from 0.400 to 0.596. On the other hand, it did come at the expense of precision. Using RMS prop, the updated model's precision of 0.645 actually decreased by 0.013 from the baseline model's precision of 0.658.

E.)

Epoch	Training Loss	Training Recall	Training Precision	Training True Positives
5	0.6309	0.6163	0.6561	11041
15	0.6187	0.6348	0.6736	11372
30	0.6068	0.6440	0.6942	11536

Looking at the table above, increasing the number of epochs from 5 to 30 increased our training recall by 0.0277 from 0.6163 to 0.6440 respectively. Precision saw an even greater increase moving from the 5-epoch model to the 30-epoch model by 0.0381.

F.)

Initializer	Training Loss	Training Recall	Training Precision	Training True Positives
He	0.643	0.604	0.641	10824
RandomNormal	0.626	0.625	0.666	11202
GlorotUniform	0.628	0.622	0.662	11150

Different types of initializations appear to improve our training performance. RandomNormal yielded the best results for both recall and precision. This is .225 higher than our baseline model (model 5) recall rate of 0.40 and 0.008 higher than the baseline precision of 0.658.

G.) Initial model 5 with relu activation, and RMSprop weight initializer appears to yield the most optimal training results. This model had a training loss of 0.6293, recall of 0.6214, precision of

0.6591 and correctly identified 11131 true positives. Against our baseline model 5, this model's loss is favorable by 0.035, recall is favorable by 0.2214, precision is favorable by 0.001, and correctly identified 3.9k more true positives.

Part 2

A.) With 30 epochs, our validation data has almost identical loss compared to the training data. The precision is slightly better on the validation set, with 76% of individuals classified as having heart disease actually having heart disease, whereas the percentage is 69% on the training set. However, the recall is much worse: 64% on the training and only 49% on the validation set. In this context, recall is important: this means that of all individuals that truly have heart disease, we are correctly identifying them only 49% of the time, which is worse than a random flip of a coin. We have an overfitting problem given that this number is much higher on the training set.

B.)

Model	L2 Parameter	Validation Loss	Validation Recall	Validation Precision	Validation True Positives
1	0.05	0.648	0.340	0.784	1070
2	0.01	0.617	0.535	0.715	1682
3	0.001	0.620	0.532	0.712	1672
4	0.1	0.659	0.626	0.691	1969
5	0.02	0.637	0.859	0.611	2702

Adding regularization parameters to our best model appears to have made our model's recall more favorable. Setting lambda to 0.02, our validation recall improves from 0.0.6214 to 0.859. However, using this lambda value yielded the lowest validation precision score of all the lambda values tested.

C.)

Model	Validation Loss	Validation Recall	Validation Precision	Validation True Positives
1	0.677	0.338	0.673	1064
2	0.669	0.351	0.667	1103
3	0.668	0.346	0.675	1087
4	0.665	0.351	0.680	1105
5	0.671	0.335	0.675	1055

Based on the above table we can see that different dropout rates in different layers (due to the multiple dropout regularizations we used we did not include these in the table, please find them specified in the notebook) did not affect the validation set much.

D.)

Model	Validation Loss	Validation Recall	Validation Precision	Validation True Positives
1	0.686	0.172	0.752	540
2	0.696	0.000	0.000	0
3	0.680	0.265	0.746	832
4	0.675	0.347	0.680	1090
5	0.657	0.365	0.682	1149

The second model performed very poorly, and it appears that it never predicted someone would have heart disease. Overall, the validation recalls for these models were much lower than on the training data, indicating significant overfitting.

E.)

Model	Validation Loss	Validation Recall	Validation Precision	Validation True Positives
1	0.866	0.974	0.523	3063
2	0.565	0.719	0.724	2261
3	0.560	0.692	0.738	2178
4	0.690	0.577	0.679	1815
5	0.602	0.510	0.810	1081

We observed an improvement in our model overall with the addition of batch normalization. Our highest recall rate was from a model with batch normalization implemented every two layers. In that case, the precision drops to 0.523. On the training side, we see a lower recall of 0.68 and a higher precision rate of 0.73. While it is not overfitting to the extent of the models from part d, there is still evidence of slight overfitting.

F.)

Model	Validation Loss	Validation Recall	Validation Precision	Validation True Positives
1	0.607	0.878	0.612	2786
2	0.650	0.916	0.592	2906
3	0.563	0.675	0.758	2140
4	0.632	0.610	0.691	1936
5	0.572	0.727	0.719	2305

The combination of batch normalization and dropout regularization did appear to improve our model, and it did not seem to be overfitting like other models above were. The combination of batch normalization and dropout appears to perform better than just dropout alone.

G.)

Model	Validation Loss	Validation Recall	Validation Precision	Validation True Positives
1	0.618	0.725	0.690	2298
2	0.696	0	0	0
3	0.645	0.696	0.702	2207
4	0.644	0.924	0.582	2931
5	0.651	0.176	0.860	560

Including a combination of batch normalization, dropout, and L2 regularization gave mixed results. Recall bounces around from 0 to 0.92 and 0 to 0.86 for precision. Overall, the results are inconsistent. The model with the best recall, 0.92, has a training recall rate of 0.65. From this we see evidence of underfitting in our training model. The precision for this model was relatively low.

H.) For our best model, we chose the last model in part F because it has the best F1 score. We used the F1 score as a way to balance recall and precision: high recall with low precision just means that many people without heart disease are being told they have heart disease. On the flip side, high precision and low recall means that few people are being told they have heart disease when they really do. This is very bad, as these people will be missing out on care and treatment they could otherwise be receiving. We believe that a balance between both is important so an F1 score makes the most sense.

Part 3

A.)

Dataset	Loss	Recall	Precision	True Positives
Training	0.578	0.673	0.725	11996
Validation	0.582	0.614	0.776	1948
Test	0.892	0.998	0.502	7007

When we run our best model on out test set, we observe that the recall in our test outperforms the recall in our training and validation sets. However, the loss and precision scores in our training and validation sets are better than in our testing set. As far as the fit goes, there is evidence of some overfitting in regards to the loss and underfitting with respect to the recall.