COVID-19 DIAGNOSES VIA SALIVA TEST WITH MACHINE LEARNING MODELS

Zhanyang Sun

Wake Forest University

Abstract

COVID-19 is a highly infectious disease that caused a global pandemic. During the pandemic, there was a need to test individuals on whether or not they were infected. However, the standard tests involved an uncomfortable nose swab. We study a test involving only saliva that could predict the illness as well as the nose-swab test. Dan Lesky [4] published a mass spectroscopy data set on proteins in saliva that could potentially be used to classify if a person has COVID-19. We use machine learning approaches including logistic regression, support vector machines, neural networks, and autoencoder to model this task, and compare and contrast the precision, recall, and F1-scores of the corresponding saliva test with the nose-swab test. The models we developed reached the FDA requirements for recall. Nevertheless, we have too many false positive results in our current models.

Fig. 1: Annoying nose-swab test [3]

Data and Preprocessing

The data consists of 3,257 patient records, each with 2,715 variables. The variables are protein measurements of saliva. The response variable is whether the patient has covid-19 using a PCR test.

We convert the response variables to be 0 and 1, where 0 means a negative result and 1 signifies a positive result (the person is sick). We use min-max scaling to transform the protein data. Finally, we split the data into training, validation, and testing sets with a proportion of 70:15:15.

All computer codes are done using Python. [5, 6, 7]

References

[1] Renesh Bedre. *Support Vector Machine (SVM) basics and implementation in python*. Apr. 2021. URL: https://www.reneshbedre.com/blog/support-vector-machine.html.

[2] Steven Flores. *Variational autoencoders are beautiful*. Apr. 2019. URL: https://www.compthree.com/blog/autoencoder/.

[3] Rebecca Ann Hughes. *Nose injuries and trauma: Should Italy's schoolchildren undergo regular COVID swab tests?* Oct. 2020. URL: https://www.forbes.com/sites/rebeccahughes/2020/10/01/nose-injuries-and-trauma-should-schoolchildren-undergo-regular-covid-swab-tests/?sh=1f1d88e177c8.

[4] Dan Lesky. *Saliva Testing Dataset*. https://www.kaggle.com.

[5] Wes McKinney et al. "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56.

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

[7] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Techniques

Logistic regression is a classification method that yields a probability for the response. The user sets a threshold of when to label a probabilistic response as positive (patient sick) or negative (patient not sick).

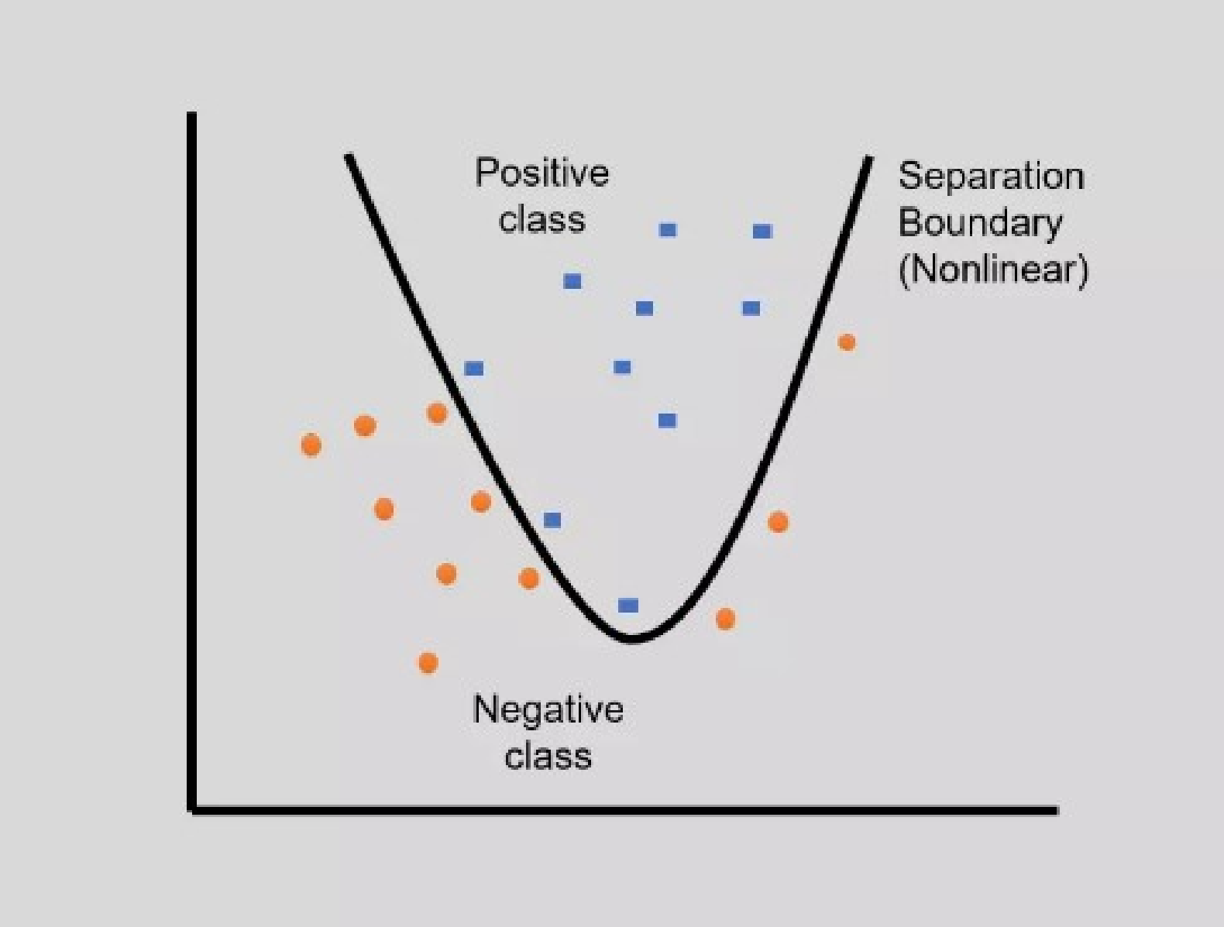

Fig. 2: Support Vector Machine[1]

Support vector machines (SVM) classify data points using decision boundaries with parameters c, gamma, and a chosen kernel function. The parameter gamma determines the curvature of the decision boundaries. The parameter c controls the error. The kernel function is used to transform and regularize the data. These parameters are optimized using grid search. For our data, the optimal parameters are c = 10, gamma = 0.1, and kernel rbf.

Neural networks use multiple hidden layers to classify the response variables. The number of hidden layers we chose is 3, with 100, 100, and 200 hidden neurons, respectively. In addition, Rectified Linear Unit (ReLu) is used as the activation function.

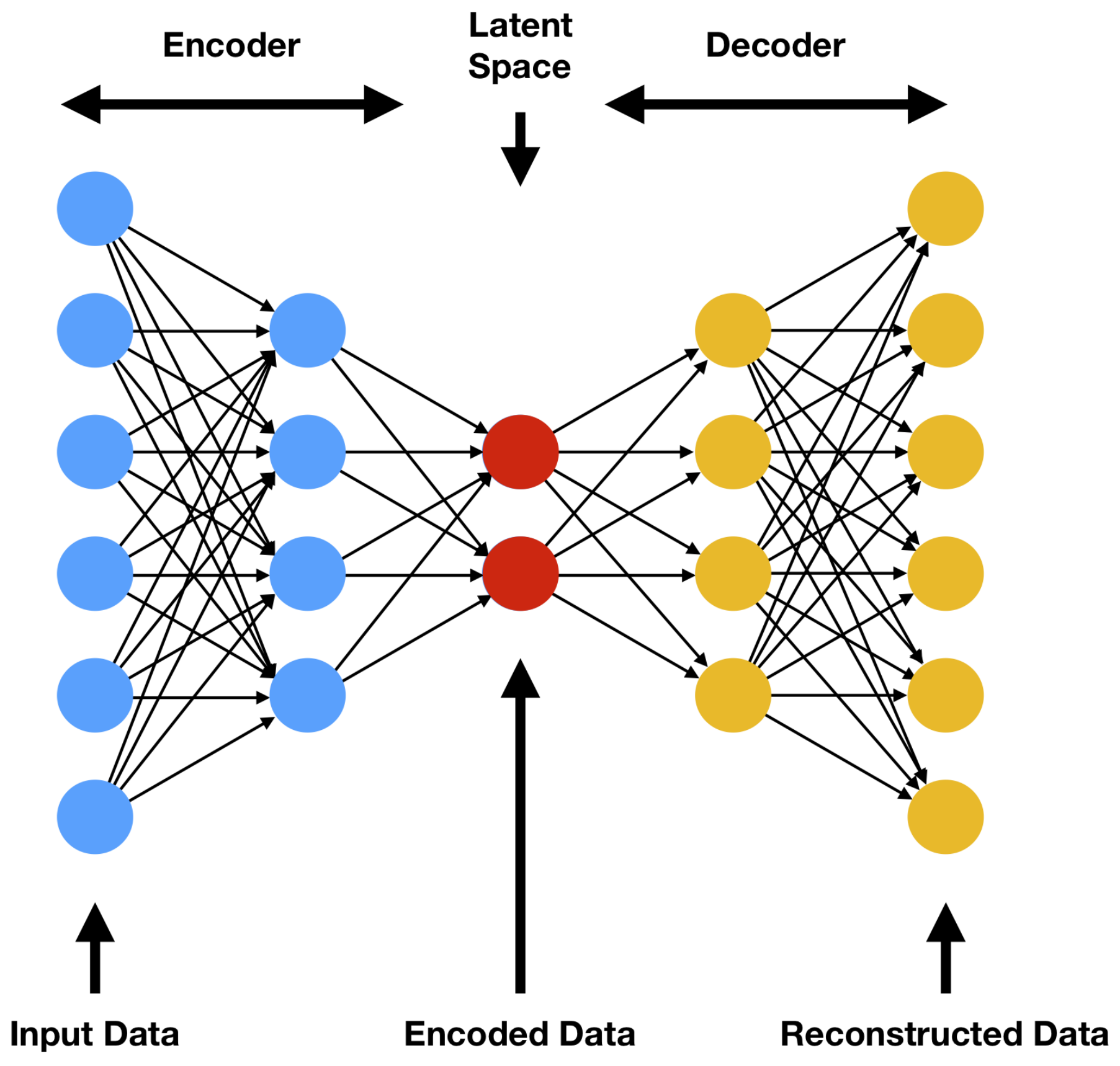

Fig. 3: Autoecoder structure[2]

The above figure demonstrates the basic idea of **autoencoder**. Autoencoder includes both dimension reduction and a neural network. The data is encoded into a smaller dimension, which is then reconstructed. If there is a small loss between the original data and the reconstructed data, the encoded data is a good abstraction of the original. We then apply a neural network to the encoded data.

We applied the logistic regression, svm, neural network, and autoencoder methods to the validation data set to identify which model yielded the best performance. On the validation data set, logistic regression and svm performed similarly. We chose svm because its confusion matrix was slightly better.

Methods

First, Lasso regression is applied to the training set to reduce the number of variables from 2,715 to 400.

Second, with this smaller set of variables, we build logistic regression, support vector machine, and neural network models on the training data.

Third, we used autoencoder, a method that encompasses both a reduction in dimension of the data and a deep neural net structure to construct an additional model on the training data.

Comparison of models on the validation set

The logistic regression model has its best performance with a threshold of 0.54, yielding the following accuracy table.

	precision	recall	f1-score	support
0	0.89	0.83	0.86	116
1	0.92	0.95	0.94	245
accuracy			0.91	361

For SVM with $c = 10$, $gamma = 0.1$, and kernel rbf, the accuracy table is shown below

	precision	recall	f1-score	support
0	0.91	0.81	0.86	116
1	0.91	0.96	0.94	245
accuracy			0.91	361

The accuracy table for the neural network is

	precision	recall	f1-score	support
0	0.88	0.81	0.84	116
1	0.91	0.95	0.93	245
accuracy			0.90	361

Autoencoder performed poorly compared to the others. The best result occurred with a threshold 0.62.

	precision	recall	f1-score	support
0	0.48	0.22	0.31	116
1	0.71	0.89	0.79	245
accuracy			0.67	361

Results

The SVM model applied to the testing data set yielded an accuracy of 0.93 (surprisingly better than its performance on the validation data set).

	precision	recall	f1-score	support
0	0.93	0.86	0.89	121
1	0.93	0.97	0.95	240
accuracy			0.93	361

Conclusion

The models for the saliva test has a false positive rate of 0.07. The nose-swab test, on the other hand, has a false positive rate of 0.05, which is slightly better than our model. The nose-swab test is more accurate compared to our saliva test.