M4L1 Solutions

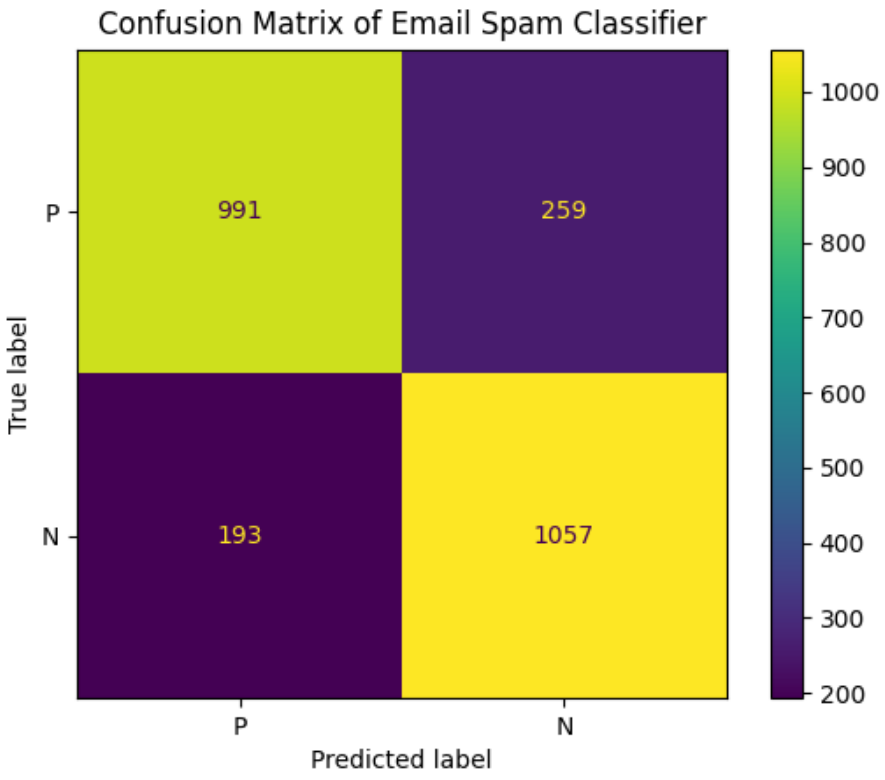
January 2024

Exercise A

Following is the confusion matrix of a hypothetical machine learning model classifying spam emails where the default assumption is that emails are not spam. The "P" stands for "Positive" samples (spam), and the "N" stands for "Negative" samples (not spam.)

[1]:
```
# display image
from IPython import display
display.Image("../prototype/emailspam.png")
```

[1]:



How many type II errors does this model have?

a) There are 991 type II errors.

b) There are 259 type II errors.
c) There are 193 type II errors.
d) There are 1057 type II errors.

Solution

Type I & II Errors

In statistical hypothesis testing, a type I error is the rejection of an actually true null hypothesis, while a type II error is the failure to reject a null hypothesis that is actually false.

1. Type I errors are the same as false positives. We are in the presence of a false positive if valid email is marked as spam. Type I errors are the rejection of a true null hypothesis.
2. Type II errors are the same as false negatives. We are in the presence of a false negative if spam message is passed as a valid email. This is a type II error because we accept the conclusion of the email being good, even though it is incorrect. Type II errors are the acceptance of a false null hypothesis.

Therefore, there are a total of 259 type II errors in this hypothetical model.

Exercise B

What are Distance Metrics? Discuss two popular distance metics used in Machine Learning.

Solution

Euclidean Distance

Euclidean Distance is one of the most commonly used distance metric. It represents the shortest distance between two vectors and is the square root of the sum of squares of differences between corresponding elements. Most machine learning algorithms, including K-Means use this distance metric to measure the similarity between observations. This is L2-norm.

Manhattan Distance

Manhattan Distance is the sum of absolute differences between entries in the vectors. This is preferred when dealing with data in high dimensions. This is L1-norm.

References

- Python Resources

* * *