INFO370 Problem Set 7: Confusion Matrix, Cross-Validation

November 22, 2020

Instructions

This PS has the following goals:

1. Learn confusion matrix, A, P, R and F scores, and what are the advantages/disadvantages of these measures.
2. Learn to predict using *sklearn* and logistic regression.
3. Learn to test/select your model through cross-validation. These two steps form a major part of predictive modeling.

1 Confusion Matrix (45pt)

This question does not expect you to do any programming. Most of the computations you can do manually, maybe you want to use calculator. Write your answers as valid markdown text, including confusion matrices as markdown tables!

You are working as an ML expert at the paleonthology department for professor Smith. Your task is to devise a model that can distinguish between bones of *simolestes vorax* (S) and other (O) marine dinosaurs (see Figure 1). You have some test data and you develop two models, *M1* and *M2*. The model performance is in Table 1, left panel.

1. (4pt) Show the confusion matrix for M1 and M2.
2. (5pt) Compute accuracy, precision, recall, and F-score for both models.

Figure 1: Reconstruction of *Simolestes vorax*. Dmitry Bogdanov, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

3. (6pt) Prof Smith is going to publish her paper and wants you to give the final results tomorrow. Which of these two models, M1 or M2 will you recommend her? Explain to her.

   Note: prof Smith is a paleontologist who knows nothing about data science and such. But she wants to hear your suggestions and reasoning!

Next, you get a job at the mycology department. You are designing an image recognition algorithm for prof. Joffe to distinguish between poisoneus (P) and edible (E) mushrooms. You design two models: M3 and M4. The performance on test data is in Table 1, middle panel.

4. (4pt) Show the confusion matrices for M3 and M4.

5. (5pt) Compute the accuracy, precision, recall for both models.

6. (6pt) Prof. Joffe wants to get the app out tomorrow. Which model, M3 or M4 will you recommend him to use? Explain your reasoning!

Afterwards you get an internship in King County Superior Court. Sheriff Johanknecht wants you to develop an ML algorithm that classifies the defendants to guilty (G) or innocent (I) based on accessible evidence. Although the final decision is done by the judge, your results will weight heavily in that decision. You devise two models, M5 and M6. The test run results are in Table 1, right panel.

Table 1: True value, and model predictions on the test data

S. vorax			Mushrooms			Defendants		
True	M1	M2	True	M3	M4	True	M5	M6
S	S	S	P	P	P	G	G	I
O	O	S	E	P	E	I	G	I
S	O	S	E	E	E	I	I	I
S	S	S	P	P	E	G	G	G
O	S	S	P	P	P	I	I	I
S	S	S	E	E	E	G	G	I
O	O	O	E	P	E	G	I	G
O	O	O	E	E	E	I	G	I
O	S	S	E	P	E	I	I	I
S	S	S	E	E	E	G	G	I

7. (4pt) Show the confusion matrices for M5 and M6.
8. (5pt) Compute the accuracy, precision, recall for both models.
9. (6pt) Johanknecht wants to commission the AI system tomorrow. Which model would you recommend her to use? Explain your reasoning!

2 Cross-validate to the best model (55pt).

Your second task is to create the best model to predict cancer based on Wisconsin Breast Cancer data. The data includes the diagnosis ("B" = no cancer, "M" = cancer), and 30 different cell measures, most of which are incomprehensible for the uninitiated ☺

1. (2pt) Load the data. Ensure you understand the variables well enough (but no need to learn their biological meaning). We recommend to consult the uploaded doc file.

   Remove the *id* variable.

Now it is time to create a predictive model. Let us limit our work here just to models that only contain 3 features (cell measures) to predict the outcome. We are also going to use *sklearn* library.

2. (2pt) Create your outcome vector $\mathbf{y}$. While `LogisticRegression` can easily handle string labels "B" and "M", the `f1_score` cannot. So we recommend you to convert the diagnosis into a numeric 1/0 label.

3. (4pt) Create the design matrix X that contains three arbitrary columns from among your features. I recommend to use `.iloc`, e.g. `X = data.iloc[:,[1,5,24]].values` to make a design matrix from these three columns.

4. (6pt) Fit a logistic regression model predicting $\mathbf{y}$ using this design matrix X.

5. (9pt) Compute the predicted outcomes, and display the confusion matrix and F-score.

   Hint: read https://otoomet.bitbucket.io/machinelearning-py.html#82_Predicting_with_Logistic_Regression for logistic regression with sklearn, https://otoomet.bitbucket.io/machinelearning-py.html#83_Confusion_Matrix%E2%80%93Based_Model_Goodness_Measures for confusion matrix and related measures.

6. (9pt) What do you think, what is the most appropriate measure here? Why is F-score a good way to test the models?

7. (9pt) Cross-validate the model goodness using F-score as the outcome. (use 5-fold or more)

Now it is time to check all possible 3-feature models. You can do three nested loops that run over the columns, create X out of those columns, and repeat the above.

8. (9pt) Loop over all 3-feature combinations and:

   (a) fit the corresponding logistic regression model
   (b) cross-validate the F-score

   Note: this is slow (5-10min). For debugging purposes, you can loop over just a small subset of combinations (say, 10), and run the final results just when done with everything else. If you find your computer too slow, then just put some sort of limit how many combinations you run through.

9. (5pt) Finally, tell what is your best model and its F-score and accuracy. Which three features did you include?

How much time did you spend on this PS?