Name: _____

5 Model Evaluation

5. (15 points) Triage is done in emergency rooms when limited medical resources must be allocated to maximize the number of survivors. Some versions of the triage system involve a color-coding scheme using, in decreasing order of treatment urgency, red (immediate), yellow (observation), green (wait), and white (dismiss) tags. In normal days at the BT Hospital (BTH), the proportion of tags given to patients is 5% red, 25% yellow, 60% green, and 10% white. You can assume that these tags are completely accurate.

   The BTH staff is training classifiers based on physiological signals and other patients' features in order to identify the level of urgency of its patients more quickly; classifications will help inform final triage decisions. Assume that BTH stores anonymized records of 10,000 patients.

   (a) For this question assume for some reason that you can only use one held-out set to evaluate a classifier, so you divide the data into two parts: part 1 has 70% of the data and part 2 has 30%.

   i. If you want to train a classifier that is more likely to work well on new data, how would you use the parts?

   ◯ part 1 (70%) for training, part 2 (30%) for testing

   ◯ part 2 (30%) for training, part 1 (70%) for testing

   ii. If you want to be more sure about your classifier's error estimate, how would you use the parts?

   ◯ part 1 (70%) for training, part 2 (30%) for testing

   ◯ part 2 (30%) for training, part 1 (70%) for testing

   iii. If we want to train a good classifier *and* predict the model's performance highly accurately, would it be a good idea to instead use all data for both training and testing? Why or why not?

   (b) Aiming to predict its model's performance for future patients, BTH uses accuracy score to evaluate classifiers. We define `accuracy score`, where N is the number of predictions, as

$$\frac{\sum_{i=1}^{N} \text{equals}(y_i, \hat{y}_i)}{N}$$

   where $\text{equals}(y_i, \hat{y}_i) := 1$ when y_i (actual tag) equals $\hat{y}_i$ (predicted tag), and 0 otherwise.

   i. If the model assigns random tags so that each tag is equally probable, what would the average accuracy score be? ◯ 5% ◯ 25% ◯ 33.3% ◯ 50% ◯ 60%

   ii. If the model always predicts the most likely tag, what would the average accuracy score be? ◯ 5% ◯ 25% ◯ 33.3% ◯ 50% ◯ 60%

(c) Notice that a model could reach a 95% accuracy score while mis-identifying all of the patients who need help most urgently (the red color tags). How can we change our evaluation metric so that it treats the prediction accuracy on each class equally? Define a new scoring formula with this property.

(d) Accuracy score also assumes that all classification errors have equal cost. But, for example, it is very costly to predict *white* when the answer should have been *red*. Assume you have a function $C(g, a)$ where g is your predicted tag color and a is the actual one, which quantifies how bad it is to predict g when the correct answer is a.

Provide a scoring formula in terms of C that takes these error costs into account.

(e) Evelyn thinks the C function approach is too complicated and suggests simply assigning integer values *red* $= 4$, *yellow* $= 3$, *green* $= 2$, *white* $= 1$ and letting $C(g, a) = (g - a)^2$. Either (1) explain why Evelyn's approach is as expressive as using a general C function, or (2) give a concrete example in this domain of a preference that Evelyn's method cannot express.