CS5228 – Tutorial 5

Classification & Regression I (Evaluation)

1. **Basic classification metrics.** Assume that you have trained a binary classifier that aims to predict if a bank customer will default on his/her credit. Your test data contained 4,000 samples and your final model yields the following confusion matrix:

		actual label: **1** (default)	actual label: **0** (no default)
prediction	**1** (default)	**260**	**610**
prediction	**0** (no default)	**10**	**3120**

   (a) Calculate the *Accuracy*, *Specificity*, *Sensitivity*, *Recall*, *Precision*, and *F1 score*!

   (b) For each metric, provide a verbal interpretation of the resulting value in the context of predicting a customer's likelihood to default on his or her credit! Discuss if we can be happy with the result, or what result might cause problems in practice!

2. **Imbalanced datasets.** One reason why we have different metrics to evaluate the quality of a classifier is because of imbalanced datasets where a majority class contains most of the samples whereas a minority class contains only a fraction of samples (assuming a binary classification task).

   (a) List 5 example applications where you would expect a very imbalanced dataset.

   (b) While not covered in the lecture, what do you think can be done to address the issue of imbalanced datasets (beyond picking the right metric)?

3. **Assessing classification errors.** In case of a binary classification, we can make 2 types of errors:

   - False Positives (FP), also called Type I Error
   - False Negatives (FN), also called Type II Error

   In the lecture, we mentioned that in many cases these two types of errors are not equally problematic.

(a) List 2 example applications where False Positives are more problematic than False Negatives, and vice versa. Provide a brief explanation!

(b) Regarding any difference between errors of Type I and Type II, how would you assess their relative importance for the following application use cases:

- Earthquake warning systems
- Automatic missile defense system

4. **How good is "good enough"?**

(a) Say you trained a binary sentiment classifier that classifies social media posts (e.g., tweets) into "negative" and "positive". Your datasets for training and testing were balanced and sufficiently large. Let's assume that the F1 score of your classifier is 0.85. How would you assess if this is a good result?

(b) Say you trained a classifier that identifies whether an image contains a Car, Boat, or Plane, and the F1-score is very high, say, 0.99. Your datasets for training and testing were balanced and sufficiently large. What might be a reason why the classifier would suddenly perform poorly in practice?