1 Exercise 1

1.1 Cross Entropy

$$H(y, g) = -(0 * log(0.25) + 1 * log(0.6) + 0 * log(0.15)) = 0.222$$

1.2 Mean Squared Error Loss

$$MSE(y, g) = \frac{1}{3}(0.25^2 + (-0.4)^2 + 0.15^2) = 0.0817$$

1.3 Hinge Loss

$$SVM(y, j) = max(0, 0.25 - 0.6 + 1) + max(0, 0.15 - 0.6 + 1) = 1.2$$

2 Task A

2.1 Accuracy

$$ACC = \frac{TP + TN}{P + N}$$

, high accuracy of predicting predominant class is easily achieved and misleading. If we have highly unbalanced data the problem of accuracy as a metric is quite apparent: Let's consider an example with medical data: normally only a few patients have a certain disease. If our dataset consists of 200 patients, of which only 8 have cancer, than a classifier which predicts that no patient has cancer would yield 96 percent accuracy; this prediction however would be deadly for the patients. If we also consider false negatives, than we see that in every case of a cancer patient, the classifier predicted no cancer and thus false negatives are 100 percent.The ability to capture misclassifications in a satisfying way is what is lacking in the accuracy metric.

In a multi-class setting, exact matching does not distinguish between partially correct and completely incorrect labelling. The denominator, **P+N** which is equivalent to number of samples(N) is constant. It is thus sensitive to only how many labels correctly identified with respect to N

In a multiclass setting, the evaluation is averaged over the individual class instances. This therefore gives the following formulae for F1 and jaccard Index:

$$F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2tp}{2tp + fp + fn}$$

$$Jacc = \frac{1}{n}\sum_{i=1}^{n}\frac{tp}{tp + fp + fn}$$

$$Jacc_c \leq F1_c$$

Jaccard's metric penalizes bad classification more than the F score

"https://stats.stackexchange.com/questions/273537/f1-dice-score-vs-iou" As F1 automatically weights more TP than Jaccard, given the same set of predictions, F1 will give the notion of high performance than Jaccard for any single instance. Averaged over number of classes, F1 will easily achieve high evaluation values. Thus between F1 and Jaccard in NN in multi-label settings, Jaccard will pressure the network to give more TP's keeping FP and FN low, to achieve high evaluation scores, which is what one would expect of a NN.

2.2 Task B

GT Class	Freq
B	4
T	4
D	4
C	0

$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{tp}{tp+fp}$ $Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{tp}{tp+fn}$ $Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{tp}{tp+tn+fp+fn}$
$J_{acc} = \frac{1}{n}\sum_{i=1}^{n}\frac{tp}{tp+fp+fn}$

Confusion Matrix

B $\begin{array}{c|c} 2 & 3 \\ \hline 1 & 2 \end{array}$ T $\begin{array}{c|c} 3 & 2 \\ \hline 2 & 1 \end{array}$ D $\begin{array}{c|c} 1 & 2 \\ \hline 2 & 3 \end{array}$ C $\begin{array}{c|c} 0 & 6 \\ \hline 2 & 0 \end{array}$

	Accuracy	Precision	Recall	Jacc	F1
B	$\frac{5}{8} = 0.63$	$\frac{2}{3} = 0.67$	$\frac{2}{4} = 0.5$	$\frac{2}{5} = 0.4$	$\frac{4}{7} = 0.57$
T	$\frac{5}{8} = 0.63$	$\frac{3}{5} = 0.6$	$\frac{3}{4} = 0.75$	$\frac{3}{6} = 0.5$	$\frac{2}{3} = 0.67$
D	$\frac{1}{8} = 0.13$	$\frac{1}{3} = 0.33$	$\frac{1}{4} = 0.25$	$\frac{1}{6} = 0.17$	$\frac{2}{7} = 0.29$
C	$\frac{0}{8} = 0$	$\frac{0}{2} = 0$	$\frac{0}{2} = 0$	$\frac{0}{2} = 0$	$2\frac{0}{0} = 0$
AvgC	0.35	0.4	0.38	0.27	0.38
AvgF	0.46	0.53	0.5	0.36	0.51

Mean Computations
class No:

$$Jacc = \frac{1}{4}(0.4 + 0.5 + 0.17 + 0) = 0.27$$

$$Precision = \frac{1}{4}(0.67 + 0.6 + 0.33 + 0) = 0.4$$

$$Recall = \frac{1}{4}(0.5 + 0.75 + 0.25 + 0) = 0.38$$

$$F1 = \frac{1}{4}(0.57 + 0.67 + 0.29 + 0) = 0.38$$

class frequency:

$$Jacc = \frac{1}{12}(0.4 * 4 + 0.5 * 4 + 0.17 * 4 + 0)$$

$$Jacc = \frac{1}{12}(0.4 * 4 + 0.5 * 4 + 0.17 * 4)$$

$$Jacc = \frac{1}{12} \times 4\,(0.4 + 0.5 + 0.17)$$

$$Jacc = \frac{1}{3}(0.4 + 0.5 + 0.17) = 0.36$$

$$Precision = \frac{1}{3}(0.67 + 0.6 + 0.33) = 0.53$$

$$Recall = 0.5 + 0.75 + 0.25/3 = 0.5$$

$$F1 = 0.57 + 0.67 + 0.29/3 = 0.51$$

exact match:

$$ExMatch = \frac{\#correctInstances}{\#instances} = \frac{2}{8} = 0.25$$

Questions

1. Intuitively the Exact Match would be the strictest metric possible. However,it is not be the lowest number: how come?
In this example, the exact match had the lowest value.

2. Can you compute the global accuracy on this example? Justify your answer.
The global accuracy can be calculated by taking the average of the individual class instance accuracies and taking into account the class imablances. By averaging accuracies with respect to class frequencies (micro averaging) produces the global accuracy of the system. Calculated above as 0.46

3. What is the main issue going from multi-class to multi-label setting?
In a multi-class setting a prediction is either correct of not. In a multi-label setting we can have partially correct labels. Under these conditions we need different definition for the metrics.