1. Theoretical Tasks

1.1 Loss Functions

The following formulas are useful for doing the exercise, where n denotes the length of both the prediction vector y and the ground truth vector g.

Cross-Entropy Loss (or Logistic Loss)

$$H(y,g) = -\sum_{i}^{n} g_i log(y_i)$$

Mean Squared Error-Loss

$$MSE(y,g) = \frac{1}{n}\sum_{i}^{n}(y_i - g_i)^2$$

Hinge Loss (or SVM Loss)

$$SVM(y,j) = \sum_{i \neq j}^{n} max(0, y_i - y_{i+1})$$

where j is the index of the correct label for the sample.

Task

Consider the following two vectors:
g = [0, 1, 0]
y = [0.25, 0.6, 0.15]
Using the formula defined in the previous section, calculate the following values between the two vectors:

- Cross-Entropy Loss

ANSWER:
$H(y,g) = -0 * log(0.25) - 1 * log(0.6) - 0 * log(0.15) = 0.2218$

- Mean Squared Error Loss

ANSWER:
$MSE(y,g) = \frac{1}{3} * ((0.25-0)^2 + (0.6-1)^2 + (0.15-0)^2) = \frac{1}{3} * 0.245 = 0.0817$

- Hinge Loss

ANSWER:
$SVM(y,j) = max(0, 0.25 - 0.6) + max(0, 0.6 - 0.15) = 0.45$

Resources:

What's an intuitive way to think of cross entropy?
https://www.quora.com/Whats-an-intuitive-way-to-think-of-cross-entropy/answer/Vlad-Vurdalak?srid=Oh7g
Section 3.13 from the Deep Learning Book
https://www.deeplearningbook.org/contents/prob.html#pf15
Notes from CS231n
https://cs231n.github.io/linear-classify/#softmax

(Optional) Evaluation Metrics

Measuring the performance of a system is a common yet very important task. Of particular interest is doing it in a way such that the results of two different system are comparable.
For this reason there are metrics which are computed in a very specific way, s.t. their values for two systems can be compared directly.
However, in practice, things are slightly more complicated.

Task A: Theoretical Foundations

Typically people refer to accuracy as THE evaluation metric, but there are a lot of evaluation metrics which can be better suited than accuracy depending on the task/dataset.

1. *In which situation using accuracy is not necessarily a good idea?*

ANSWER:
For example when dealing with imbalanced data sets accuracy can be misleading. Let's imagine we're developping a network that classifies images between two classes cats and dogs.
If our data sets is composed at 99% of cats, and our model classifies all of our images as dogs our accuracy will be of 0%, which is really really low. But it is not representative of our model. It is the same otherwise, if our network had classify all the images in the data set as cats the accuracy would be 98%, but it wouldn't be representative of our model.
For accuracy to be representative of the model it is applied to, we need to be certain of the data set imbalencement and its coherence, as it should be for our class in a classification problem.

2. *What part of the formula for computing the accuracy makes it less desirable than the Jaccard Index (Intersection Over Union) in a multi-class setting ?*

ANSWER:
The Jaccard index measure the overlap (similarity) between the predicted and segmentation masks. It also allows for different classes to be weighted differently, which is useful in cases where some classes are more important than others. Plus, Jaccard index for each class can then be averaged to give an overall measure of performance.
Whereas accuracy in a multi-task setting assumes that each class is equally import, which can be problematic if classes are imbalanced or if some classes have more importance than others.

3. *What is the difference between Jaccard Index (Intersection Over Union) and F1-Measure? Which one is more suited to measure performance of NNs ?*

ANSWER:
Both the Jaccard Index and F1-Measure are evaluation metrics, within machine learning. However, they have different approach regarding their measurement.
In one hand we have the Jaccard Index that measure the similarity between two sets, which is basically the size of the intersection between the sets divided by the size of the union of these sets.
In the other hand we have the F1-Measure, it is a measure of the balance between precision and recall.
Precision is the ratio of true positives to the total number of positive predictions. And recall is the ratio of true positives to the total number of actual positives.
F1-Measure combines these two metrics into a single score by taking their harmonic mean.
The choice of metrics depends on the specific problem being addressed. Indeed, the Jaccard Index is more suited to measuring the similarity between sets, while the F1-Measure is more suited to measuring the performance of binary classifiers.

Task B: Practice

In this part of the exercise we want to compute some common alternatives which can be used instead of accuracy. We'll take an example from a real case scenario of layout analysis at pixels level of historical documents.
Given the following prediction and ground truth (note: this is a multi-class and multi-label scenario!), where B stays for background, T for text, D for decoration and C for comment.

	1	2	3	4	5	6	7	8
GT	B	B	B	B	B	TD	TD	TD	TD
P	B	T	TD	BD	BC	TC	T	TD

Compute the class frequencies and the following metrics per class:

- Class Frequencies

ANSWER:
B: 4
T: 2
D: 2
C: 1

- Jaccard Index

ANSWER:
B: 0.5
T: 0
D: 0
C: 0

- Precision

ANSWER:
B: 1
T: 1
D: 0.7
C: 0.5

- Recall

ANSWER:
B: 2/(2+2)=0.5
T: 1
D: 0.5
C: 1

- F1-measure

ANSWER:
B: 0.7
T: 1
D: 0.6
C: 0.7

1. Intuitively the Exact Match would be the strictest metric possible. However,it might not be the lowest number: how come?

ANSWER: The exacr metric would be 0 since there are instances for which the model did not predict the correct label.

2. Can you compute the global accuracy on this example? Justify your answer.

ANSWER: Here is the global accuracy: 4/8=0.5
We can calculate it since 4 out of 8 scenarion have been predicted correctly.

3. What is the main issue going from multi-class to multi-label setting?

ANSWER: The main issue is that, in the multi-class setting, each instance is assigned to only one class, while in the multi-label setting, each instance can be assigned to multiple labels.

Resources

Jaccard Index (or Intersection over Union):
https://en.wikipedia.org/wiki/Jaccard_index
Exact Match (and others metrics):
https://en.wikipedia.org/wiki/Multi-label_classification
Precision and Recall:
https://en.wikipedia.org/wiki/Precision_and_recall