
W3WI DS304.1 Applied Machine Learning Fundamentals

Exercise Sheet # 8 - Evaluation of Machine Learning Models

Question 1 EX 2020, modified (Confusion matrix and evaluation metrics)

The evaluation of a neural network on a separate test dataset has produced the confusion matrix depicted in table 1:

Conf. mat.		gold			
		\mathcal{C}_1	\mathcal{C}_2	\mathcal{C}_3	Σ
predicted	\mathcal{C}_1	25	4	9	38
	\mathcal{C}_2	8	31	0	39
	\mathcal{C}_3	2	3	18	23
	Σ	35	38	27	100

Table 1: Confusion matrix produced on a test set.

Please answer the following questions:

1. What is the accuracy of the model?
2. Compute precision, recall, and the F_1 -score separately for all classes \mathcal{C}_k ($1 \leq k \leq 3$).
3. What is the micro average precision/recall? What do you observe?
4. What is the macro average precision/recall?

Question 2 (Micro and macro averages)

When should you use micro average and when macro average?

Question 3 (F_1 -score and the harmonic mean)

The F_1 -score is defined as the harmonic mean of precision and recall. Can you imagine why the harmonic mean is used? (Why not the arithmetic mean?)

Question 4 EX 2023, modified (Confusion matrix and evaluation metrics)

You have to evaluate a binary classifier on a test set. Table 2 lists the predictions of the model along with the correct labels. \oplus represents the positive class, and \ominus the negative class.

Data point	1	2	3	4	5	6	7	8	9	10
Prediction	\ominus	\oplus	\oplus	\ominus	\oplus	\oplus	\oplus	\ominus	\oplus	\ominus
True label	\oplus	\oplus	\ominus	\ominus	\oplus	\oplus	\ominus	\oplus	\ominus	\ominus

Table 2: Results on the test dataset.

1. Create the confusion matrix.
2. Compute precision, recall, and the F_1 -score of your model.

Question 5 EX 2020

Imagine you have trained a classification model to classify skin tissue samples as either cancerous or healthy. The model should avoid false negatives at all costs. Which evaluation metric (precision or recall) would you prefer? (*Please explain your answer.*)

Question 6 EX 2023 (Drawback of accuracy)

What advantage does the F_1 -score have over accuracy?

Question 7 EX 2021, modified (AUC and ROC)

For ten test instances a logistic regression classifier has output the probabilities (of the positive class \oplus) shown in table 3. Draw the ROC curve (*receiver operating characteristic*) and compute the AUC (*area under the curve*). Do not forget to label the axis! How do you rate the performance of the model?

Data point	Gold Label	Probability
1	\oplus	0.95
2	\oplus	0.30
3	\ominus	0.35
4	\ominus	0.10
5	\oplus	0.80
6	\oplus	0.55
7	\ominus	0.25
8	\ominus	0.75
9	\ominus	0.05
10	\oplus	0.20

Table 3: Probabilities output by a logistic regression classifier for ten test instances.

Question 8 EX 2021 (Occam's razor)

What is meant by the term *Occam's razor*?

Question 9 EX 2021 (Bias and variance)

Which statement concerning bias and variance is correct?

- ☐ Models suffering from high bias tend to overfitting.
- ☐ A high variance can be mitigated by adding more training examples.
- ☐ A decision stump has a low bias.
- ☐ The terms bias and variance are not related to overfitting and underfitting.
- ☐ None of the above is correct.

Question 10 EX 2021 (Overfitting and underfitting)

Complete figure 1 below using the following terms:

Underfitting, Overfitting, good Model, Train Error, Test Error, and Model Complexity.

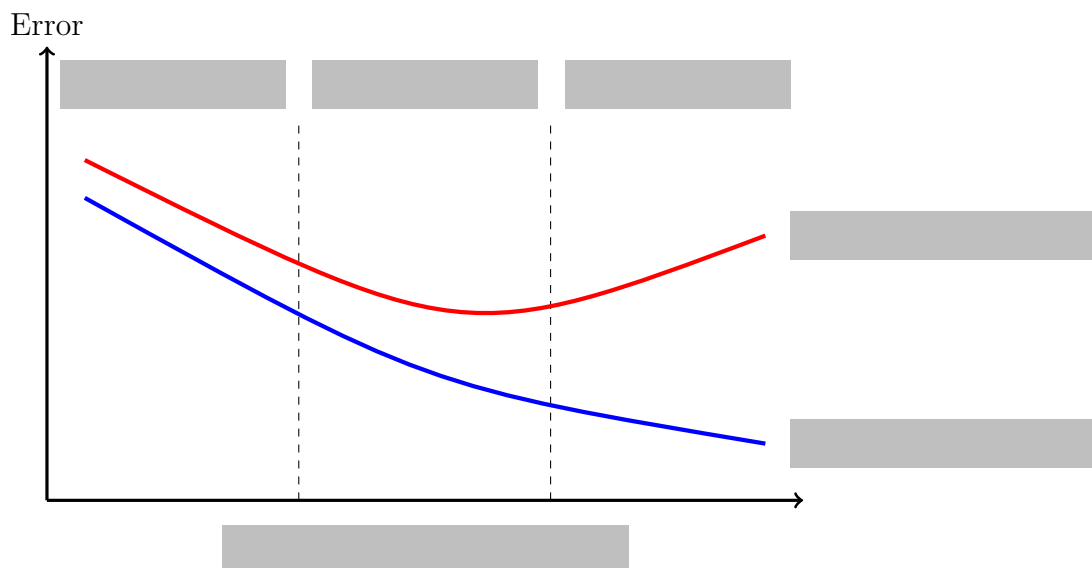


Figure 1: Complete the figure!

Question 11 EX 2021 (Early stopping)

What is *early stopping*?