Group 6: Timm Behner, Philipp Bruckschen, Patrick Kaster, Markus Schwalb
MA-INF 4111 - Intelligent Learning and Analysis Systems: Machine Learning
Exercise Sheet 4

1. Confidence Intervals

(a) given: $\text{error}_S(h) = \frac{10}{65}, n = 65$

$$\text{error}_D^{95\%}(h) \leq \text{error}_S(h) + z_N \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}$$

$$= \frac{10}{65} + 1.64 \sqrt{\frac{\frac{10}{65}(1 - \frac{10}{65})}{65}}$$

$$\approx 0.23$$

(b)

$$\text{error}_D^{90\%}(h) \leq \frac{10}{65} + 1.28 \sqrt{\frac{\frac{10}{65}(1 - \frac{10}{65})}{65}}$$

$$\approx 0.21$$

∎

2. Sample Size

given: $\text{error}_D(h) \in [0.2, 0.6] \Rightarrow \sigma = 0.2$
wanted: bound on error $E < 0.1$

$$E = z_N \frac{\sigma}{\sqrt{n}} \Rightarrow$$

$$n = \left(z_N \frac{\sigma}{E}\right)^2$$

$$n = \left(1.96 \frac{0.2}{0.1}\right)^2$$

$$= \lceil 15.3664 \rceil = 16$$

∎

3. confusion matrix

The needed functionality is implemented in `confusion_matrix.py` which is used in `leaveoneout.py`. `leaveoneout.py` was only slightly modified to use a confusion matrix and to print all values stated by the exercise. ∎

4. AUC

(a)

$$\text{Accuracy}_S(h) = \frac{\#\text{ Correctly Classified}}{\#\text{Test Examples}}$$
$$= \frac{990}{100} = 0.99$$

∎

5. AUC

Use probability based ranking as described in the lecture. Assume that for every sample $x \in S$ there is a teacher output classifying x to be a positive or negative sample. Define H^+ to be the half-space that contains the most positive samples and H^- to be the other half-space. If both sides have an equal number of positive samples then choose one side to be H^+ at random. Now the probability that a sample $x \in S$ is positive is $p^+(x) = \frac{^+n_S^{H^i}}{n_S^{H^i}}$ where $i \in +, -$, i is the half-space where sample x is located in, $^+n_S^{H^i}$ is the number of all samples in H^i that are ranked positive and $n_S^{H^i}$ is the number of all samples in H^i. Obviously this would only lead to 2 ranks and would be quite a poor ranking.

A better ranking would be to consider the distance of x to the hyperplane H. The further away a point x in H^+ is away from H, the better it is ranked (one hopes that the further away a point is from the border the more sure one can be that is in the correct half-space). On the other side a point x in H^- is ranked higher the closer it is to H (here one hopes that a misclassified sample will be close to the border rather than being far away on the wrong side). Consider the distances between points x and the hyperplane H to be between x and a point on the hyperplane that yields the smallest distance in the euclidean distance measure. In general this will lead to more ranks than 2. (unless all points are located equally far away from either side of H). ∎