**Exercise 1:**

(a) Which of the following statement(s) is/are correct?

(i) A single ICE curve is a local explanation method. Correct

(ii) Robust local explanation methods should return similar explanations for similar observations. Correct

(iii) In ordinary Gower's distance all feature receive different weight. Not correct, all receive a weight of 1.

(b) Which of the following statement(s) about local surrogate models is/are correct?

(i) Surrogate models produced by LIME should have the same prediction as the model to be explained for the whole training dataset. Not correct, they should be faithful in the neighborhood of the point of interest, the closer a point is to the point of interest, the closer the prediction of the local surrogate model should be to the original prediction.

(ii) The choice of the sampling process and the definition of locality are important hyperparameters of LIME that have a large impact on the behavior of the method. Correct

(iii) LIME does not require any adaptions to be applicable to deep learning models for image data. Not correct, adaption to distance function is necessary

(iv) LIME requires the surrogate model to use all available features - a selection of features is not allowed. Not correct, L0-regularized/LASSO model possible

(v) If the kernel width for the exponential kernel is set to infinity, all observations receive a proximity measure/weight of 1 independent of their distance to $\mathbf{x}$. Correct

**Exercise 2:**

a) Fill out table:

| | pension | age | job type | marital status | $\hat{f}$ | $d(\mathbf{x}, \mathbf{z}.)$ | $\phi_{\sigma=0.15}(\mathbf{z}.)$ | $\phi_{\sigma=0.5}(\mathbf{z}.)$ |
|---|---|---|---|---|---|---|---|---|
| $\mathbf{x}$ | 1800 | 21 | sedentary | single | 30.6 | - | - | - |
| $\mathbf{z}_1$ | 1600 | 21 | sedentary | married | 25.8 | 0.25 | 0.06 | 0.78 |
| $\mathbf{z}_3$ | 2200 | 32 | sedentary | married | 85.2 | 0.32 | 0.01 | 0.66 |
| $\mathbf{z}_2$ | 1200 | 23 | physically | single | 74.9 | 0.49 | 0.00 | 0.38 |

- The smaller the kernel width $\sigma$ the smaller the proximity measure, the smaller the weight for the sampled data points
- If the kernel is set too small, many or all sampled observations receive a weight close to 0.
- Since there are not many datapoints used to fit the surrogate model, the model might be unstable and not faithful to the original model.

b)

$$\begin{aligned} L(\hat{f}, g_1, \phi_{\mathbf{x}}) &= \sum_{\mathbf{z} \in Z} \phi_{\mathbf{x}}(\mathbf{z}) L(\hat{f}(\mathbf{z}), g(\mathbf{z})) \\ &= 0.06 \cdot (28 - 25.8)^2 + 0.01 \cdot (105 - 85.2)^2 + 0 \\ &= 4.21 \end{aligned}$$

$$
\begin{aligned}
L(\hat{f}, g_2, \phi_{\mathbf{x}}) &= \sum_{\mathbf{z} \in Z} \phi_{\mathbf{x}}(\mathbf{z}) L(\hat{f}(\mathbf{z}), g(\mathbf{z})) \\
&= 0.06 \cdot (26.1 - 25.8)^2 + 0.01 \cdot (92.7 - 85.2)^2 + 0 \\
&= 0.57
\end{aligned}
$$

According to the faithfulness, $g_2$ should be preferred because it has a lower weighted loss.

c) No, because a random forest is by far less interpretable than a linear model with three features.

d) Yes, because there is a high probability that the random forest overfitted on the sampled data. With a new sampled dataset the faithfulness might be lower for the random forest.