Solution Quiz:

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

(a) Interpretation methods are *only* used ot explain the global behavior of a model.

⇒ **Wrong**, there are several needs for interpretability. (Gain global and local insights how the IML model works, better control, improve and debug the IML model, justify decisions)

(b) If a model-agnostic and a model-specific interpretation method are applied on the same ML model, the output of the two methods will always be the same.

⇒ **Wrong**, as the methods work different they will probably give a divergent output.

(c) While feature effects methods show the influence of a feature on the target, feature importance methods focus on a feature's impact on the model performance.

⇒ **Correct**.

(d) In IML we distinguish between global IML methods, which explain the behavior of the model over the entire feature space, and local IML methods, which only explain the prediction of individual observations.

⇒ **Correct**.

(e) Technically, Pearson correlation is a measure of *linear* statistical dependence.

⇒ **Correct**.

(f) All in the lecture mentioned measures for correlation and dependencies are limited to continuous random variables.

⇒ **Wrong**, mutual information is not limited to continuous random variables.

(g) A feature interaction between two features x_j and x_k is apparent if a change in x_j influences the impact of x_k on the target.

⇒ **Correct**.

Solution Quiz:

(a) What is the problem of PDP when interactions between features are present? How about extrapolation?

⇒ When interactions are apparent, the effects may in average cancel each other out, i.e. PDP shows no effect.

⇒ Extrapolation can cause issues in regions with few observations or if features are correlated, i.e. resulting data point may be unrealistic or very unlikely.

(b) How do PDPs and ICE curves correspond with each other?

⇒ The value of the PDP at a point x_j, corresponds to the point-wise average of the values of the ICE curves at this point.

(c) Which problem do we need to keep in mind when using centered ICE/PDP for categorical features?

⇒ If we center the ICE/PDPs for categorical features, the expected changes always refer to a selected reference category.

(d) M-Plots handle correlated data well and do not suffer from extrapolation. Which disadvantage does this method have?

⇒ M-plots suffer from omitted variable bias.

(e) Name the advantages of ALE over PDP.

⇒ Computationally faster (measurable when they are based on the same grid); less to no extrapolation.

(f) Can you think of a situation in which ALE equals PDP?

⇒ If features are uncorrelated, ALE plots are equal to PDPs.

(g) How does the interpretation between M-Plots and ALE differ?

⇒ In the M-Plot one can not infer, if the effect is due to the feature of interest or due to correlated features. ALE only shows the effect of the feature of interest.

(h) You fitted a model that should predict the value of a property depending on the number of rooms and square meters. You want to compute feature effects using the following methods: PDP, M-plots and (uncentered) ALE plots. Which of the following strategies reflect which method?
The feature effect for a 30 m^2 corresponds to...

a) ... what the model predicts on average for flats that also have around 30 m^2, for example, 28 m^2 to 32 m^2. ⇒ **M-plot**

b) ... how the model's predictions change on average when flats with 28 m^2 to 32 m^2 have 32 m^2 vs. 28 m^2. ⇒ **uncentered ALE**

c) ... what the model predicts on average if all properties in the dataset have 30 m^2. ⇒ **PDP**

Solution 1:

(a) Both PDP and ALE plots show a strong linear effect of x_1, where higher values of x_1 lead to higher values of predicted value. The PDP and ALE plot of x_2 show a strong decreasing effect of x_2 on the prediction. The PDP of x_2 shows a steep jump for large values of x_2, while the ALE plot shows a strong linear effect over the whole value range of x_2. Interpretation at $x_1 = 0.5$:

- PDP: the model predicts on average a value of around 2.3 for y if for all data instances $x_1 = 0.5$.
- ALE: the model predicts on average an increase of around 2.5 of y for data instances with $x_1 = 0.5$ compared to the average prediction.

(b) PDPs assume that features are uncorrelated. We know from the GAM output above - as well as the scatter plot - that x_1 and x_2 are highly correlated. Since PDPs extrapolate over predictions of artificial points that are out of distribution, the interpretations might be misleading - especially in areas with low data density (high values of x_2) and if the model contains interactions. ALE on the other hand, does not predict in regions that are far away from the training data and therefore do not suffer from the extrapolation issue of PDPs.