Quiz:

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

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

(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.

(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.

(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.

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

(f) All in the lecture mentioned measures for correlation and dependencies are 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.

Quiz:

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

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

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

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

(e) Name the advantages of ALE over PDP.

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

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

(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 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, e.g., 28 m^2 to 32 m^2.

   b) ... how the model predictions changes on average when flats with 28 m^2 to 32 m^2 have 32 m^2 vs. 28 m^2.

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

Exercise 1:

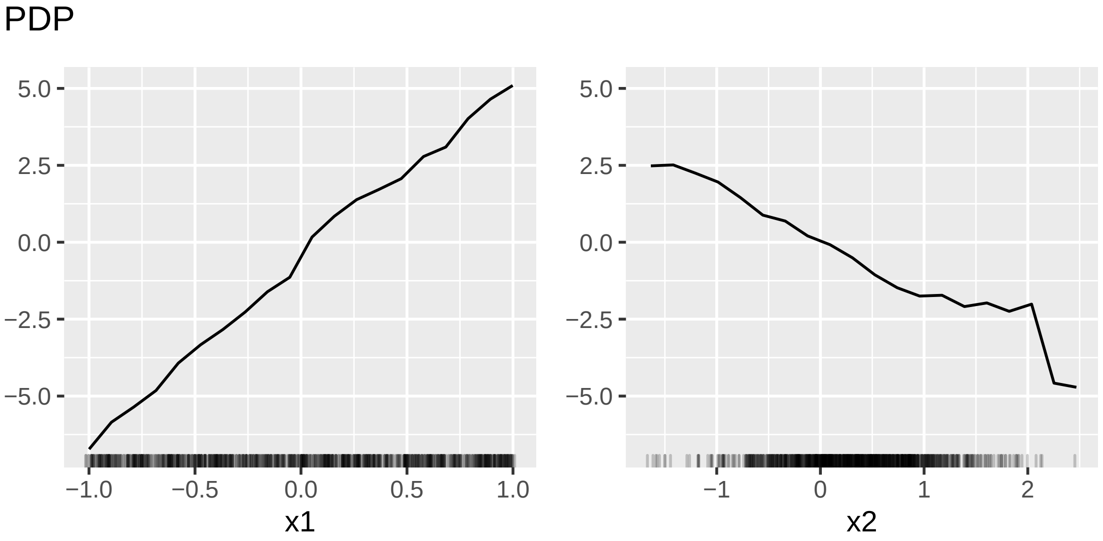
You receive a dataset with 1000 data points from a data generating process with $X_1 \sim \mathcal{U}(-1, 1)$, $X_2 = X_1^2 + \delta$, $\delta \sim \mathcal{N}(0, 0.04)$ and $Y = 5X_1 - 2X_2 + \epsilon, \epsilon \sim \mathcal{N}(0, 1)$.

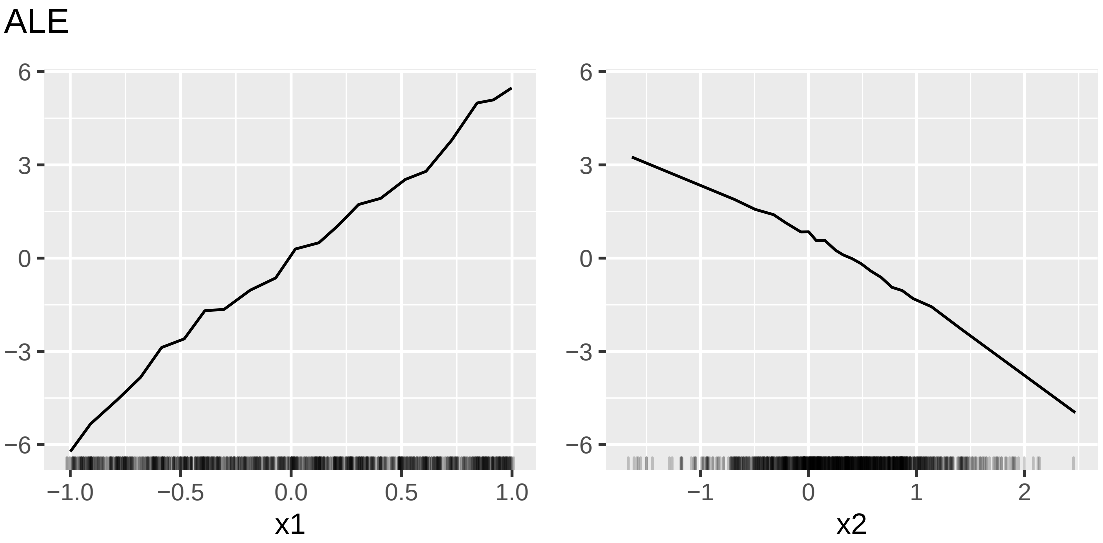
The fitted linear model has the following form: $\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + \hat{\beta}_2\mathbf{x}_2 + \hat{\beta}_3\mathbf{x}_1\mathbf{x}_2$.

Below, the PDP (first row) and ALE (second row) for x_1 and x_2 are shown.

(a) Interprete the plots with respect to the feature effect of x_1 and x_2.

(b) Would you rather trust the PDP or ALE plot? Give reasons for your decision.


PDP
x1
x2


ALE
x1
x2