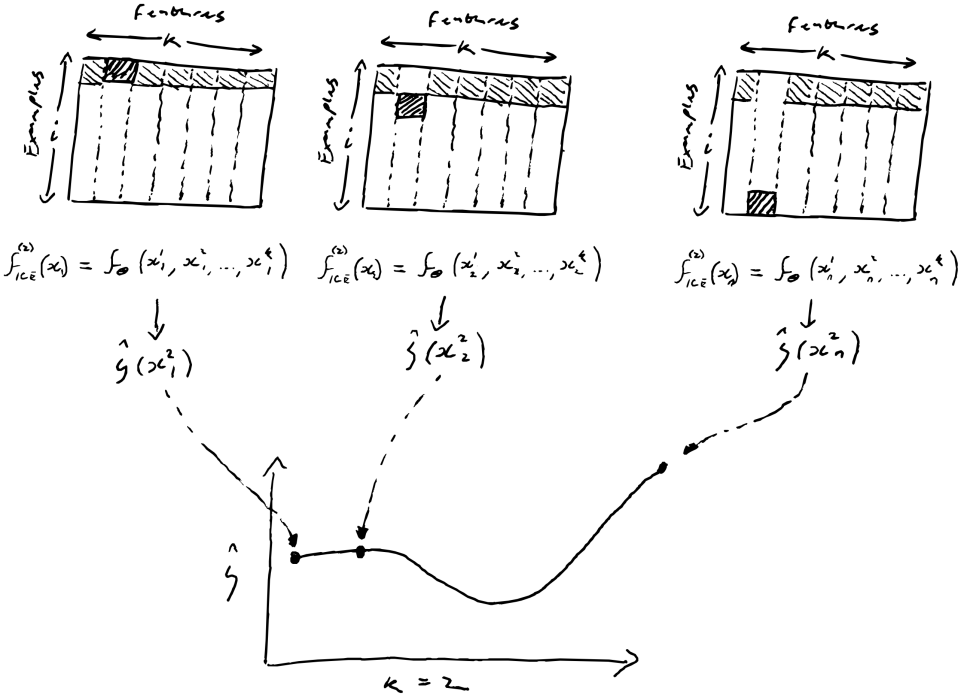
Individual Conditional Expectation Plots (ICEs)

Information is always lost when taking an average. Not only do PDPs average across features, but they also average over the entire dataset of examples. ICEs attempt to retain more information by moving feature p under investigation across its range of values while keeping the reaming variables fixed for each example. In this way ICE carves out n curves by calculating the summand of the PDP



$$f^2_{PDP}(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{f(x_i^1, x^2, \cdots, x_i^k)}_{\text{ICE}},$$

Thus retaining finer detail. Unlike PDP, ICE is a local explanation because it provides an explanation for single data instances. ICE still suffers from the same assumptions of variable independence and sparsity. The plot can often become overcrowded and messy, so it is good practice to plot only a subset of ICE curves and center the curves by subtracting a baseline. Oftentimes this baseline is the model prediction value for the instance fixed at the feature's lowest value. Selecting this baseline results in all curves starting at zero. It is important to note that although the plots become easier to interpret, we lose the information about absolute value.