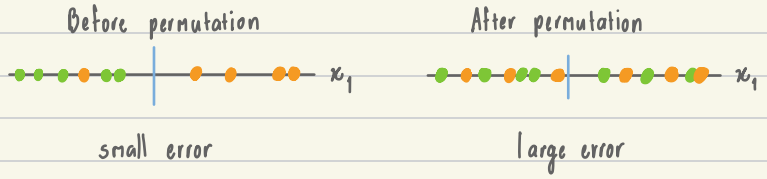
Global explainability

- general behaviour of a model
- often expressed as expected values based on the distribution of the data

Permutation Feature Importance :

- measure the increase in prediction error after permuting the feature
- the test dataset should be use



- x_1 is important since the different in error is large

Disadvantage

- adding a correlated feature can decrease the importance of the associated feature

Partial dependence profiles :

- marginal effect the feature have on the predicted outcome of the model

Disadvantage

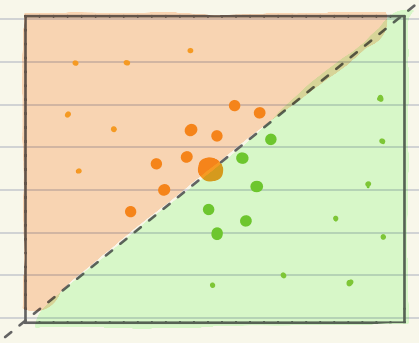
- independence assumption

Local explainability

- explain individual predictions

LIME :

- fit a linear regression in the local neighborhood



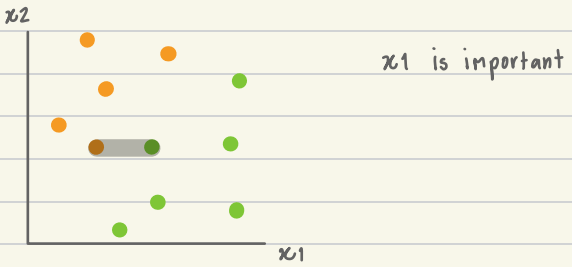
1) the point in the middle is the obs. we want to explain
2) sample the data around the obs. of interest
3) the closer the points to the obs. of interest the higher the weight
4) predicted the sample data using the original model
5) fit a linear model

Coefficient

	model intercept	x_1	x_2	important
case 1 :	0.5	1	0	x_1
case 2 :	0.5	0	1	x_2

Counterfactuals :

- find the closest observation that has a different class What are the changes needed to switch the class
- for the misclassified case, the counterfactual would be the true class



Shapley values :

KernelSHAP

1) Sample coalitions $z'_k \in \{0,1\}^M$, $k \in \{1, \dots, K\}$

Example -> x1 x2 x3

1 0 1

1 1 0

1 0 0

⋮

2) Get prediction for each coalitions by converted to the original feature space. Then apply the model

Example -> x1 x2 x3

actual average actual

actual actual average

actual average average

⋮

3) Compute the weight for each coalitions with SHAP kernel

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}\,|z'|\,(M-|z'|)},$$

where M is maximum coalition size

$|z'|$ is the number of present features in instance z'

4) Fit weighted linear model by optimizing

$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z')$ basically sum of squared errors

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

↑ shapley values

5) Return shapley values