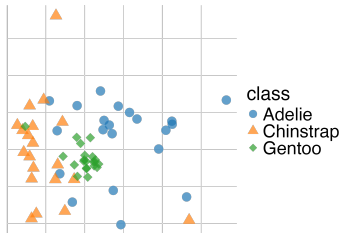
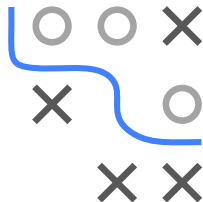


Introduction to Machine Learning

Random Forest Proximities



Learning goals

- Understand how RF can be used to define proximities of observations
- Know how proximities can be used for visualization, outlier detection and imputation

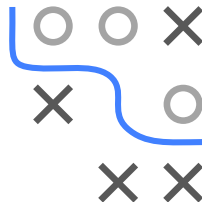
PROXIMITIES

RFs have built-in similarity measure for pairs of observations:

ID	Color	Form	Length
1	yellow	oblong	14



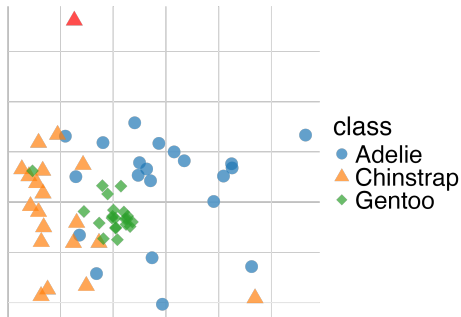
ID	Color	Form	Length
2	brown	oblong	10



- After training, push all observations through each tree
- To calculate $\text{prox}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$: Percentage of how often both points are placed in **same terminal node of a tree**
- Here: $\text{prox}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = 2/3$
- All proximities are arranged in symmetric $n \times n$ matrix

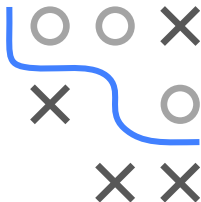
OUTLIER DETECTION

- Can also be used to **locate outliers**
- Or mislabeled points, especially in manually labeled data sets



IMPUTING MISSING DATA

ID	Color	Form	Origin	Length
1	yellow	round	domestic	14
2	brown	oblong	imported	???
3	brown	oblong	imported	19
4	???	round	domestic	14

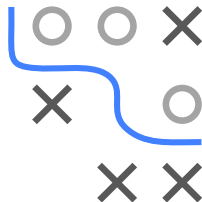


- 1 Replace missings per feature by median (of available values)
- 2 Compute proximities (NB: data has changed)
- 3 Replace missings in $\mathbf{x}^{(i)}$ by weighted average of non-missings; weights proportional to proximities

Steps 2 and 3 are iterated a few times.

IMPUTING MISSING DATA

ID	Color	Form	Origin	Length
1	yellow	round	domestic	14
2	brown	oblong	imported	14
3	brown	oblong	imported	19
4	brown	round	domestic	14



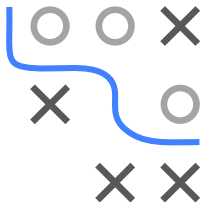
- 1 Replace missings per feature by median (of available values)
- 2 Compute proximities (NB: data has changed)
- 3 Replace missings in $\mathbf{x}^{(i)}$ by weighted average of non-missings; weights proportional to proximities

Steps 2 and 3 are iterated a few times.

IMPUTING MISSING DATA

ID	Color	Form	Origin	Length
1	yellow	round	domestic	14
2	brown	oblong	imported	17
3	brown	oblong	imported	19
4	brown	round	domestic	14

weighted average
using proximities



- 1 Replace missings per feature by median (of available values)
- 2 Compute proximities (NB: data has changed)
- 3 Replace missings in $\mathbf{x}^{(i)}$ by weighted average of non-missings; weights proportional to proximities

Steps 2 and 3 are iterated a few times.