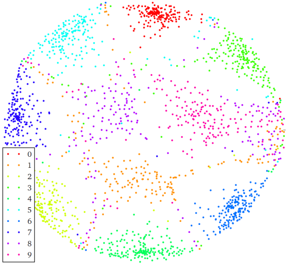
Introduction to Machine Learning

Random Forests: Proximities

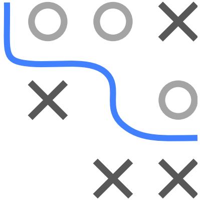


Proximity plot for a 10-class handwritten digit classification task.

Learning goals

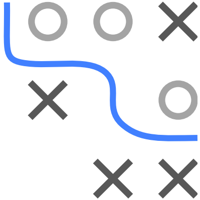
- Understand how a random forest can be used to define proximities of observations
- Know how proximities can be used for missing data, outliers, mislabeled data and a visualization of the forest

RANDOM FOREST PROXIMITIES



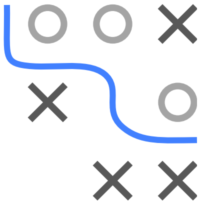
- One of the most useful tools in random forests
- A measure of similarity ("closeness" or "nearness") of observations derived from random forests
- Can be calculated for each pair of observations
- Definition:
  - The proximity between two observations $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ is calculated by measuring the number of times that these two observations are placed in the same terminal node of the same tree of the random forest, divided by the number of trees in the forest
  - The proximity of observations $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ can be written as $\text{prox}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$
  - The proximities form an intrinsic similarity measure between pairs of observations
- The proximities of all observations form a symmetric $n \times n$ matrix.

RANDOM FOREST PROXIMITIES / 2



- Algorithm:
  - Once a random forest has been trained, all of the training data is put through each tree (both in- and out-of-bag).
  - Every time two observations $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ end up in the same terminal node of a tree, their proximity is increased by one.
  - Once all data has been put through all trees and the proximities have been counted, the proximities are normalized by dividing them by the number of trees.

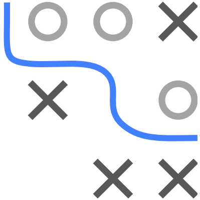
USING RANDOM FOREST PROXIMITIES



- Imputing missing data:
  1. Replace missing values for a given variable using the median of the non-missing values
  2. Get proximities
  3. Replace missing values in observation $\mathbf{x}^{(i)}$ by a weighted average of non-missing values, with weights proportional to the proximity between observation $\mathbf{x}^{(i)}$ and the observations with the non-missing values

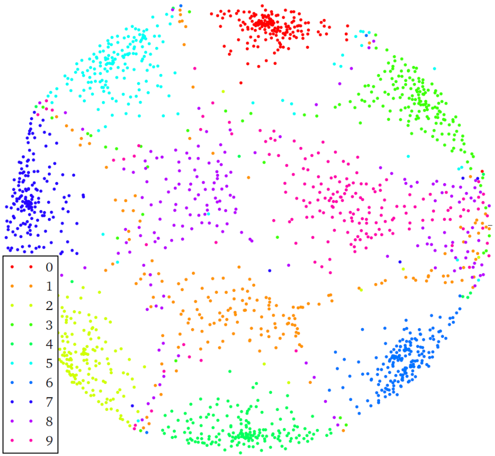
  Steps 2 and 3 are then iterated a few times.
- Locating outliers:
  - An outlier is an observation whose proximities to all other observations are small
  - Measure of outlyingness can be computed for each observation in the training sample

USING RANDOM FOREST PROXIMITIES / 2



  - If the measure is unusually large, the observation should be carefully inspected
- Identifying mislabeled data:
  - Instances in the training data set are sometimes labeled ambiguously or incorrectly, especially in “manually” created data sets.
  - Proximities can help in finding them: they often show up as outliers in terms of their proximity values.
- Visualizing the forest:
  - The values $1 - \text{prox}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$ can be thought of as distances in a high-dimensional space
  - They can be projected onto a low-dimensional space using metric multidimensional scaling (MDS)
  - Metric multidimensional scaling uses eigenvectors of a modified version of the proximity matrix to get scaling coordinates

USING RANDOM FOREST PROXIMITIES / 3



Proximity plot for a 10-class handwritten digit classification task.

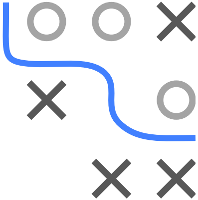


image from G. Louppe (2014) *Understanding Random Forests* arXiv:1407.7502.

USING RANDOM FOREST PROXIMITIES / 4

- The figure depicts the proximity matrix learnt for a 10-class handwritten digit classification task
  - proximity matrix distances projected onto the plane using multidimensional scaling
  - samples from the same class form identifiable clusters, which suggests that they share a common structure
  - also shows the fact for which classes errors occur, e.g., digits 1 and 8 have high within-class variance and have overlaps with other classes