ASSIGNMENT 2

GROUP 21

SOPHIE RAIN (01425316),
LUCAS UNTERBERGER (01325438),
PETER STROPPA (01326468).

datasets

- WINE
- MUSHROOMS
- CONGRESS
- AMAZON

datasets overview

Dataset	# samples	# attributes	Missing values?	# classes	classes eq. important?	Types of attributes
Amazon	750	10 000	no	50	yes	ordinal
Congress	217	18	yes (unknown)	2	yes	categorical (2 values)
Wine	1599	11	no	11	no	numeric
Mushrooms	8124	22	yes	2	no	categorical (2-10)

methods

- K-Nearest-Neighbors
- Support Vector Machines
- Random Forest

pre-processing

scaling and selection

- Min-Max-Scaling
- Z-Score-Scaling
- Feature Selection
- Handmade weighted scaling

encoding

- Ordinal Encoding
- One Hot Encoding
- TF-IDF

pre-processing

Weighted scaling

$If\ correlation < 0:$

- $$\frac{max(feature) - feature}{max(feature) * |correlation|}$$

$If\ correlation > 0:$

- $$\frac{feature}{max(feature) * |correlation|}$$

TF-IDF

results

Best result	kNN	Random Forest	SVM
Wine		perfectly	perfectly
Congress			97% accuracy
Mushroom	perfectly	perfectly	perfectly
Amazon		67% accuracy	

conclusion

diffrent datasets

- kNN really good for clustered data
- SVM highly depend on kernel
- Don‘t forget human intuition

Diffrent methods

- SVM & RF outperformed
- kNN troubles with many classes
- RF profits of heterogeneous data