ML Toolbox

Over- and Under-Sampling

Over- and under-sampling

Dealing with class imbalance in model training

- Post-hoc adjustments of class composition in training data
- Over-sampling
  - Sample cases of the minority class with replacement
- Under-sampling
  - Draw a random sample of the majority class
- Create synthetic minority instances
- **Hybrid techniques**

→ Evaluation data should still reflect class imbalance

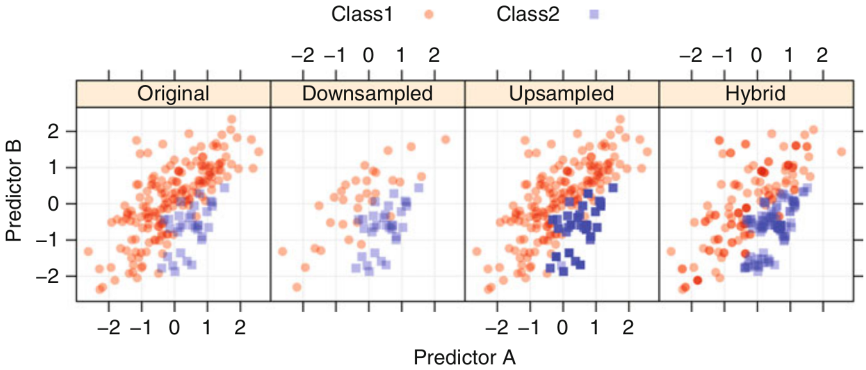
SMOTE

A synthetic minority over-sampling technique (SMOTE)

1. Initialize SMOTE
   - Set the number of neighbors K, number of iterations N
2. For each minority instance T and N
   1. Find the K nearest neighbors of the minority instance
   2. Sample one of the K nearest neighbors
   3. Multiply the distance between the sampled neighbor and T by a random number $\{0, 1\}$
   4. Create a synthetic minority instance at the coordinate of step 3
3. Optional: Combine w. down-sampling of the majority class

Comparison

Figure: Down-sampling, up-sampling, SMOTE[1]



[1] Kuhn and Johnson (2013)

Software Resources

Resources for R

- `caretEnsemble`
  1. Create a list of caret models via `caretList`
  2. Combine with `caretEnsemble` or `caretStack`
- `SuperLearner`, `subsemble`
- `smotefamily`, `DMwR`

References

Breiman, L. (1996). Stacked Regressions. *Machine Learning 24*(1), 49–64.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over–Sampling Technique. *Journal of Artificial Intelligence Research 16*(1), 321–357.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Leblanc, M., Tibshirani, R. (1996). Combining Estimates in Regression and Classification. *Journal of the American Statistical Association 91*(436), 1641–1650.

Sill, J., Takacs, G., Mackey, L., Lin, D. (2009). *Feature-Weighted Linear Stacking*. `https://arxiv.org/abs/0911.0460`

Sapp, S., van der Laan, M. J., Canny, J. (2014). Subsemble: an ensemble method for combining subset-specific algorithm fits. *Journal of Applied Statistics 41*(6), 1247–1259.

van der Laan, M. J., Polley, E. C., Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology 6*(1).

Wolpert, D. (1992). Stacked Generalization. *Neural Networks 5*, 241–259.