

OVERSAMPLING: SMOTE

Advanced Machine Learning Imbalanced Learning: Sampling Methods Part 2

- SMOTE creates **synthetic instances** of minority class.
- Interpolate between neighboring minority instances.
- Instances are created in \mathcal{X} rather than in $\mathcal{X} \times \mathcal{Y}$.
- Algorithm: For each minority class instance:
 - Find its k nearest minority neighbors.
 - Randomly select one of these neighbors.
 - Randomly generate new instances along the lines connecting the minority example and its selected neighbor.

Learning goals

- Understand the state-of-art oversampling technique SMOTE

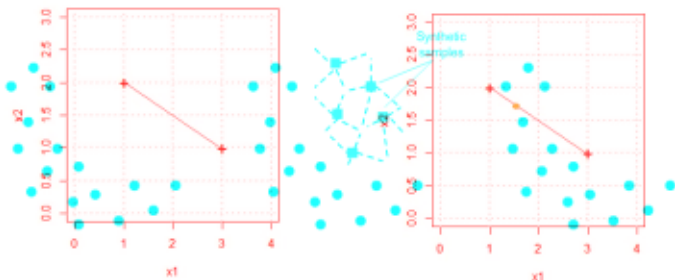


SMOTE: GENERATING NEW EXAMPLES

- Let $\mathbf{x}^{(i)}$ be the feature of the minority instance and let $\mathbf{x}^{(j)}$ be its nearest neighbor. The line connecting the two instances is
- Interpolate between neighboring minority instances.
- Instances are created in \mathcal{X} rather than in $\mathcal{X} \times \mathcal{Y}$.
- where $\lambda \in [0, 1]$
- By sampling a $\lambda \in [0, 1]$, say $\tilde{\lambda}$, we create a new instance

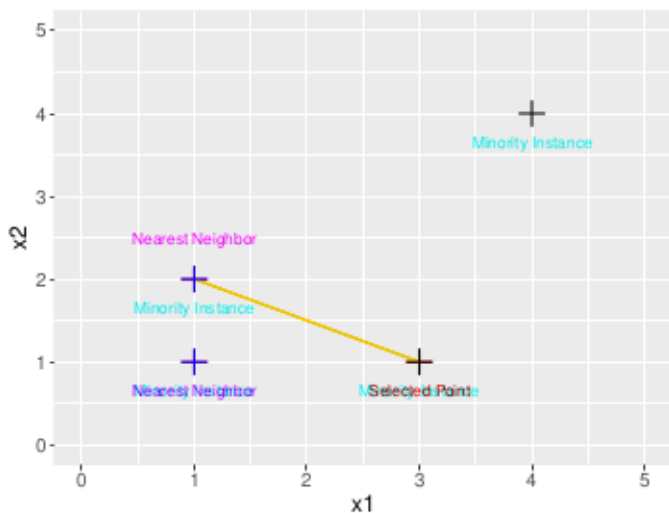
- Find its k nearest minority neighbors.
- Randomly select one $\mathbf{x}^{(j)} = \tilde{\lambda}(\mathbf{x}^{(j)} - \mathbf{x}^{(i)}) + \mathbf{x}^{(i)}$.

Example: Let $\mathbf{x}^{(i)} = (1, 2)^T$ and $\mathbf{x}^{(j)} = (3, 1)^T$. Assume $\tilde{\lambda} \approx 0.25$.



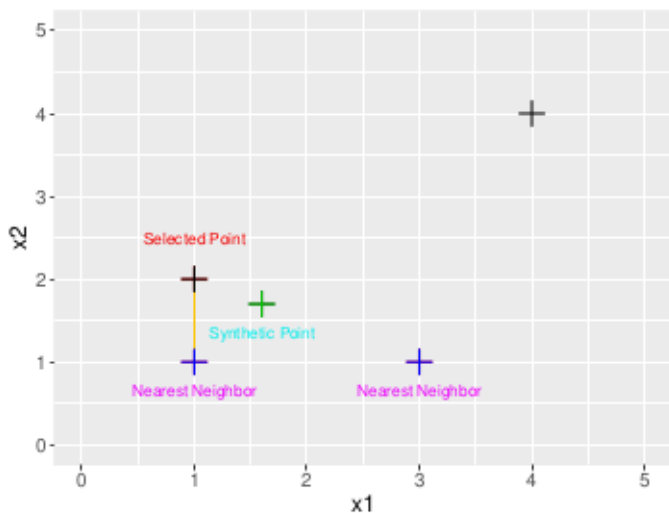
SMOTE: VISUALIZATION

For an imbalanced data situation, take four instances of the minority class. Let $K = 2$ be the number of nearest neighbors.



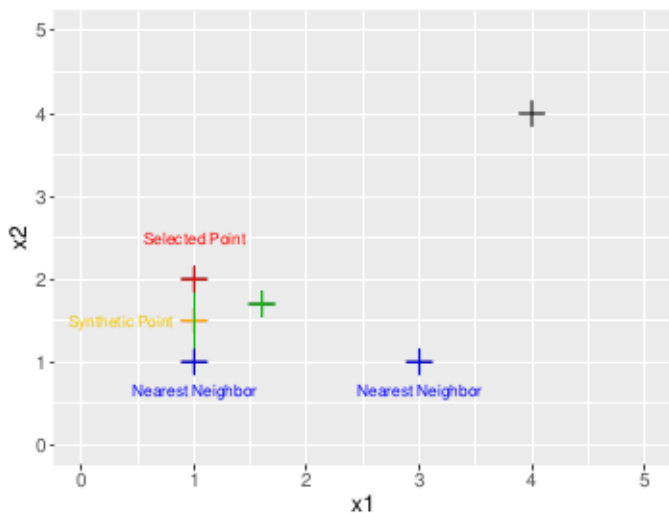
SMOTE: VISUALIZATION

For an imbalanced data situation, take four instances of the minority class. Let $K = 2$ be the number of nearest neighbors.



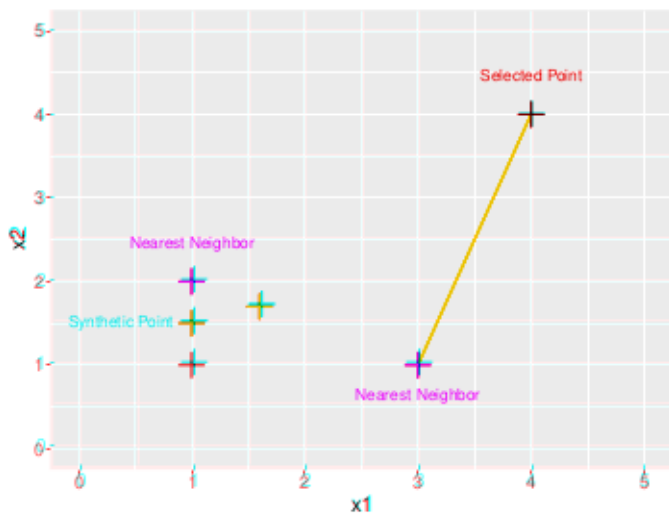
SMOTE: VISUALIZATION

For an imbalanced data situation, take four instances of the minority class. Let $K = 2$ be the number of nearest neighbors.



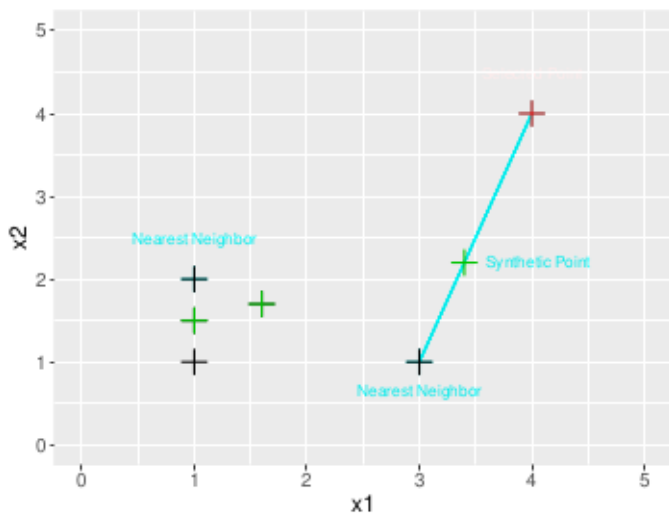
SMOTE: VISUALIZATION

For an imbalanced data situation, take four instances of the minority class. Let $K = 2$ be the number of nearest neighbors.



SMOTE: VISUALIZATION

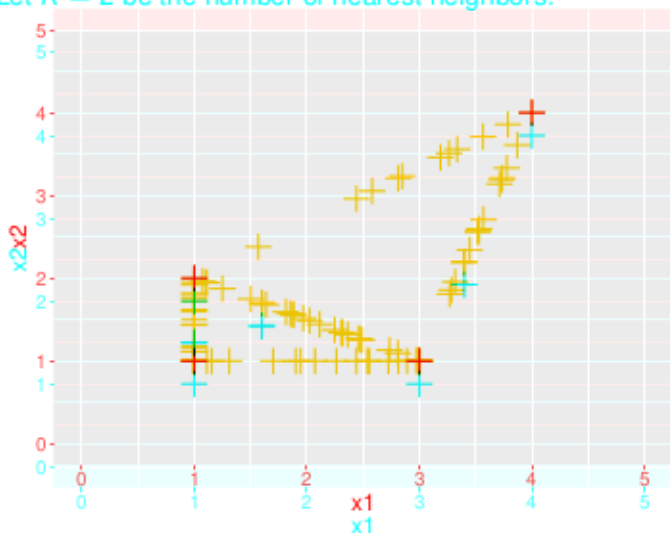
For an imbalanced data situation, take four instances of the minority class. Let $K = 2$ be the number of nearest neighbors.



SMOTE: VISUALIZATION CONTINUED

After 100 iterations of SMOTE for $K=2$ we get:

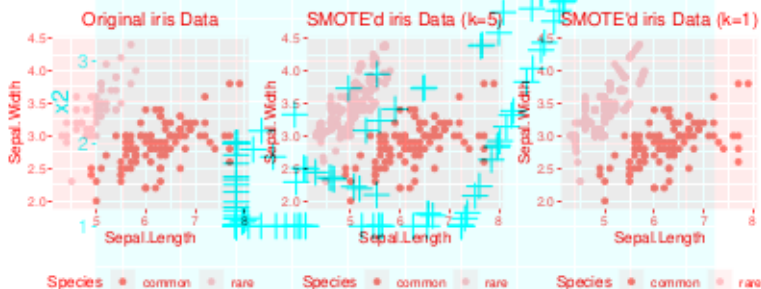
instances of the minority class. Let $K=2$ be the number of nearest neighbors.



SMOTE: EXAMPLE EXPLANATION CONTINUED

After 100 iterations of SMOTE for $K = 3$ we get:

- Iris data set with 3 classes and 50 instances per class.
- Make the data set "imbalanced":
 - relabel one class as positive
 - relabel two other classes as negative

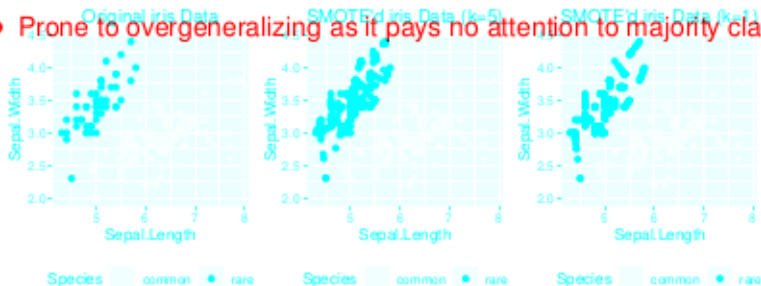


SMOTE enriches minority class feature space.



SMOTE: DISADVANTAGES

- Generalize decision region for minority class instead of making it quite specific, such as by random oversampling.
- Well-performed among the oversampling techniques and is the basis for many oversampling methods: Borderline-SMOTE, LN-SMOTE, ... (over 90 extensions!)
- Prone to overgeneralizing as it pays no attention to majority class.



SMOTE enriches minority class feature space.

COMPARISON OF SAMPLING TECHNIQUES

- Compare different sampling techniques on a binarized version of Optdigits dataset for optical recognition of handwritten digits.
- Use random forest with 100 trees, 5-fold cv, and F1-Score

basis for many oversampling methods: Borderline-SMOTE, LN-SMOTE, ... (over 90 extensions!)

- Prone to overgeneralizing as it pays no attention to majority class.

Sampling technique	Class ratio	F1-Score
None	0.11	0.9239
Undersampling	0.68	0.9538
Oversampling	0.69	0.9538
SMOTE	0.79	0.9576

- Class ratios could be tuned (here done manually).
- Sampling techniques outperform base learner.
- SMOTE leads sampling techniques, although by a small margin.



COMPARISON OF SAMPLING TECHNIQUES

- Compare different sampling techniques on a binarized version of Optdigits dataset for optical recognition of handwritten digits.
- Use random forest with 100 trees, 5-fold cv, and F_1 -Score.

Sampling technique	Class ratio	F1-Score
None	0.11	0.9239
Undersampling	0.68	0.9538
Oversampling	0.69	0.9538
SMOTE	0.79	0.9576

- Class ratios could be tuned (here done manually).
- Sampling techniques outperform base learner.
- SMOTE leads sampling techniques, although by a small margin.

