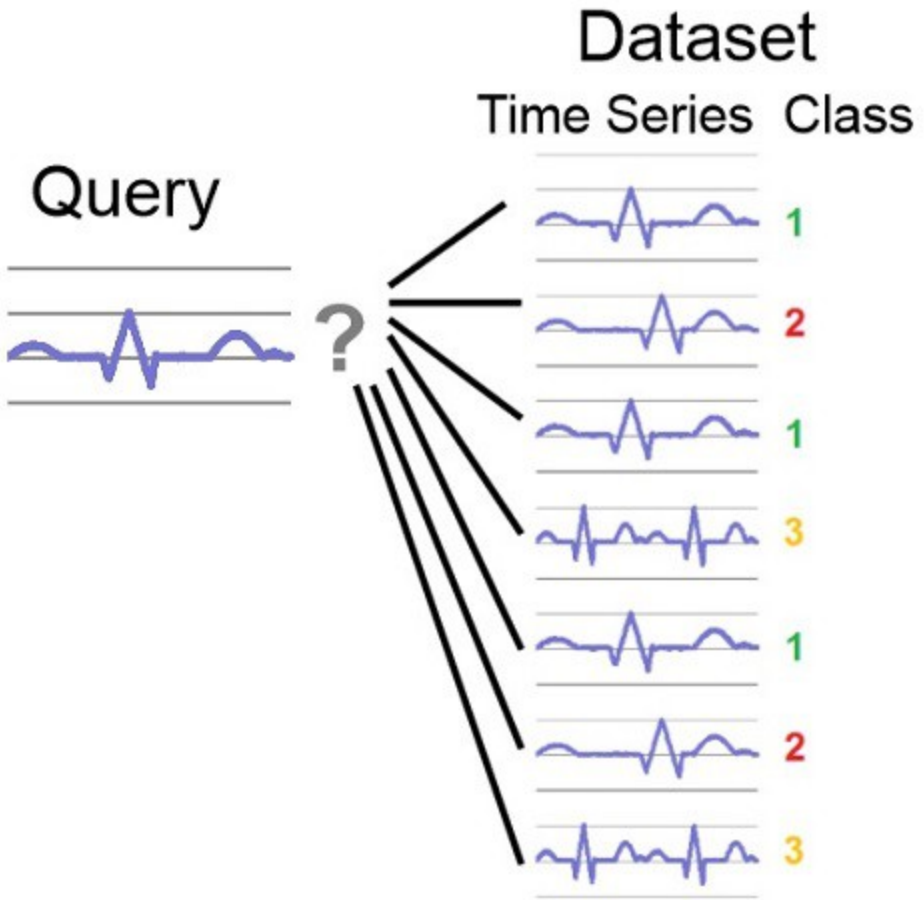
Mining and Processing Biomedical Data

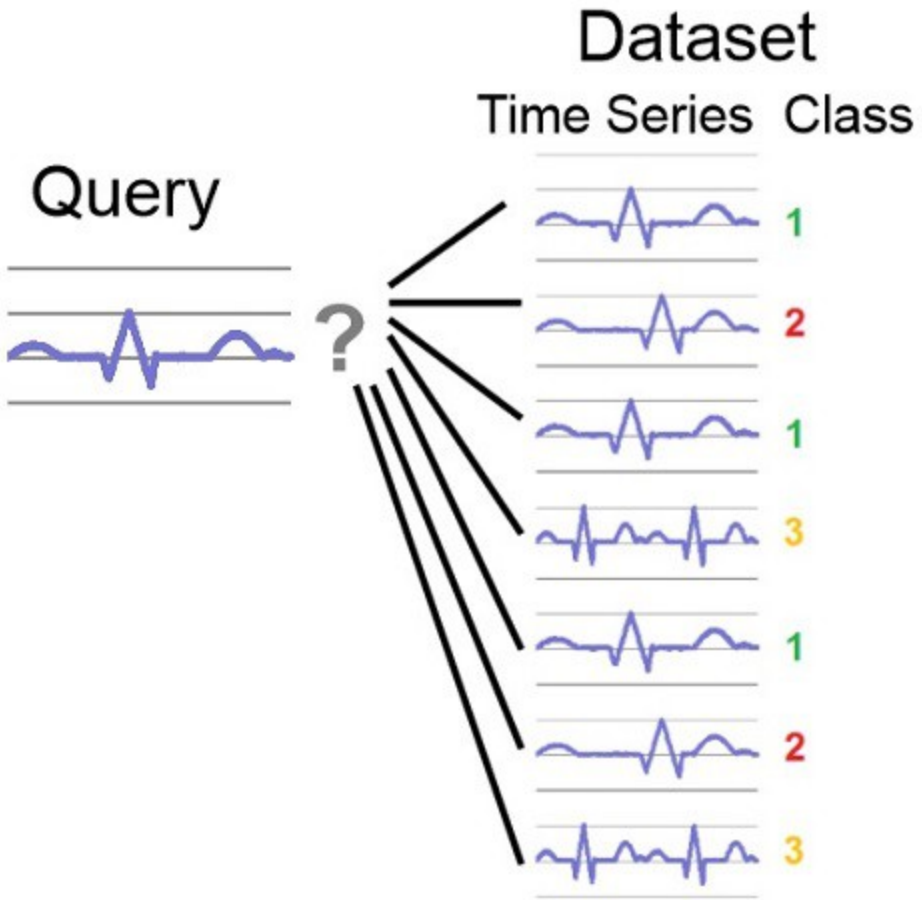
Dr. rer. nat. Krisztian Buza
adiunkt naukowy
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw, Poland
chrisbuza@yahoo.com

Speeding-up the Classification of Time-Series via Instance Selection

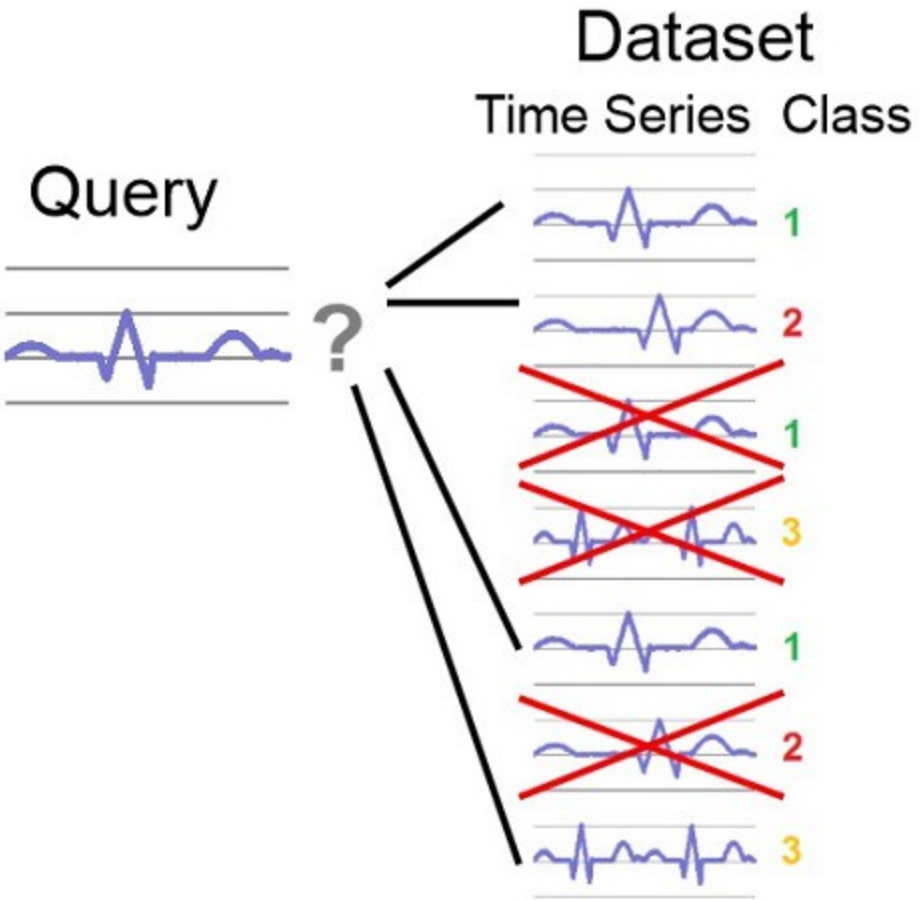
Standard nearest neighbor: Comparison to ***all*** train time series



Standard nearest neighbor: Comparison to ***all*** train time series

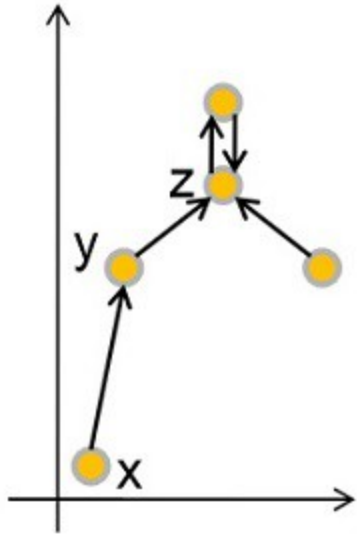


With instance selection: Comparison to the ***selected*** train time series

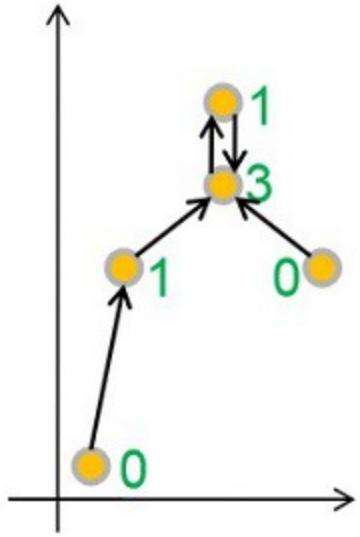


- Instance *y* is a good (bad) *k*-nearest neighbor of the instance *x* if
  (i) *y* is one of the *k*-nearest neighbors of *x*, and
  (ii) both have the same (different) class labels.

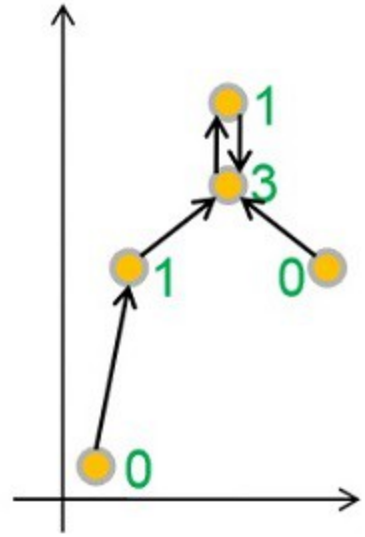
- Instance *y* is a good (bad) *k*-nearest neighbor of the instance *x* if
  (i) *y* is one of the *k*-nearest neighbors of *x*, and
  (ii) both have the same (different) class labels.

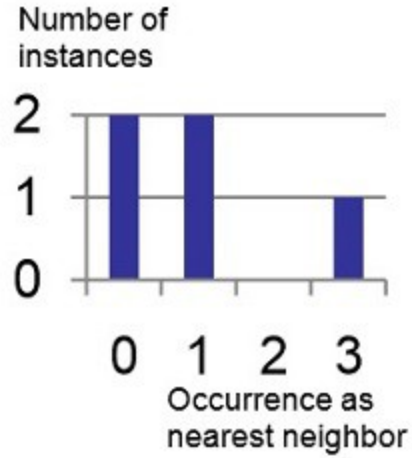


- Instance *y* is a good (bad) *k*-nearest neighbor of the instance *x* if
  (i) *y* is one of the *k*-nearest neighbors of *x*, and
  (ii) both have the same (different) class labels.

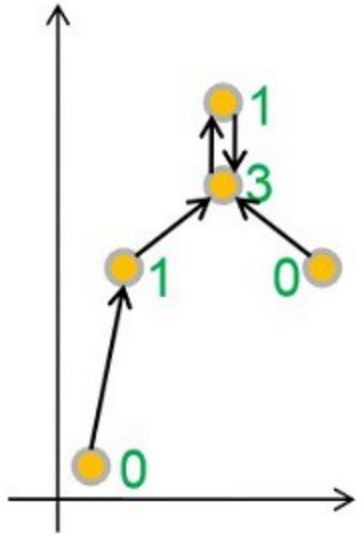


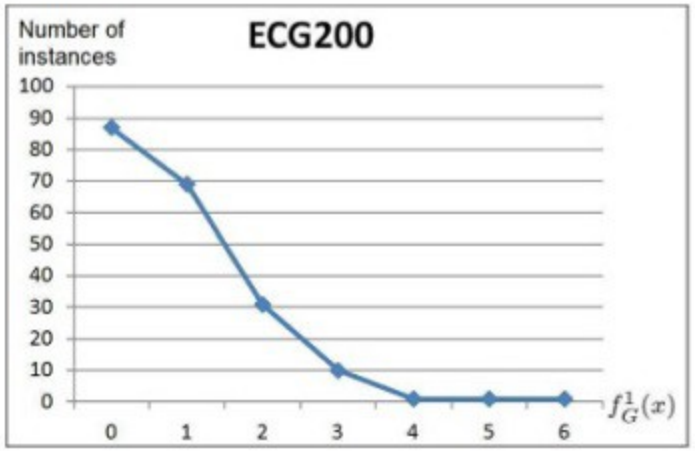
- Instance *y* is a good (bad) *k*-nearest neighbor of the instance *x* if
  (i) *y* is one of the *k*-nearest neighbors of *x*, and
  (ii) both have the same (different) class labels.

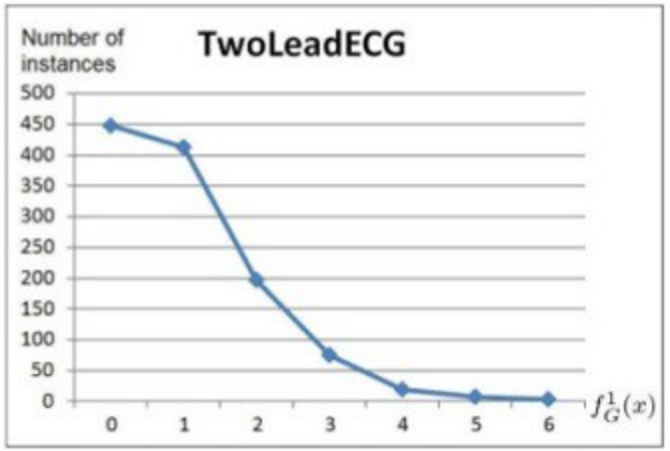


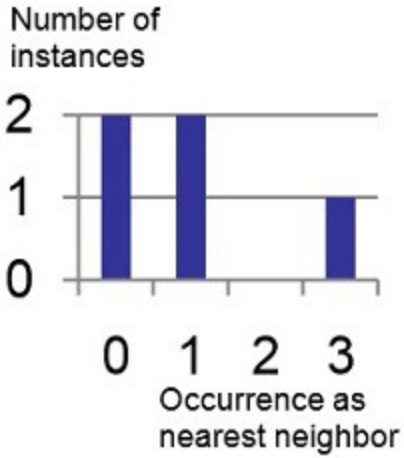


- Instance *y* is a good (bad) *k*-nearest neighbor of the instance *x* if
  (i) *y* is one of the *k*-nearest neighbors of *x*, and
  (ii) both have the same (different) class labels.
- The distribution of good (bad) nearest neighbors is substantially **skewed** → **good (bad) hubs**









Distribution of good 1-nearest neighbors for some ECG datasets

$f_G^k(x)$, $f_B^k(x)$ Number of instances that have *x* as one of their **good / bad** *k*-nearest neighbors

$$f_N^k(x) = f_G^k(x) + f_B^k(x)$$

- Good 1-occurrence score: $f_G(x) = f_G^1(x)$
- Relative score: $f_R(x) = \frac{f_G^1(x)}{f_N^1(x)+1}$
- Xi's score: $f_{Xi}(x) = f_G^1(x) - 2f_B^1(x)$
- **Hub-based Selection:**
  rank instances based on one of these scores, and select the top-ranked instances

See also:
Buza, K., Nanopoulos, A., & Schmidt-Thieme, L. (2011). INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification. LNCS Vol. 6635, Springer

Semi-supervised Classification of Time Series

- Basic assumption of (conventional) classification
  - Unlabeled (test) data originates from the same (or at least very similar) distribution as the labeled training data

- Basic assumption of (conventional) classification
  - Unlabeled (test) data originates from the same (or at least very similar) distribution as the labeled training data
- Semi-supervised classification
  - Unlabeled data may originate from slightly different distribution
  - Many unlabeled instances (time series) are available while training the model → the „structure“ of the data can be learned using both labeled and unlabeled instances

Example

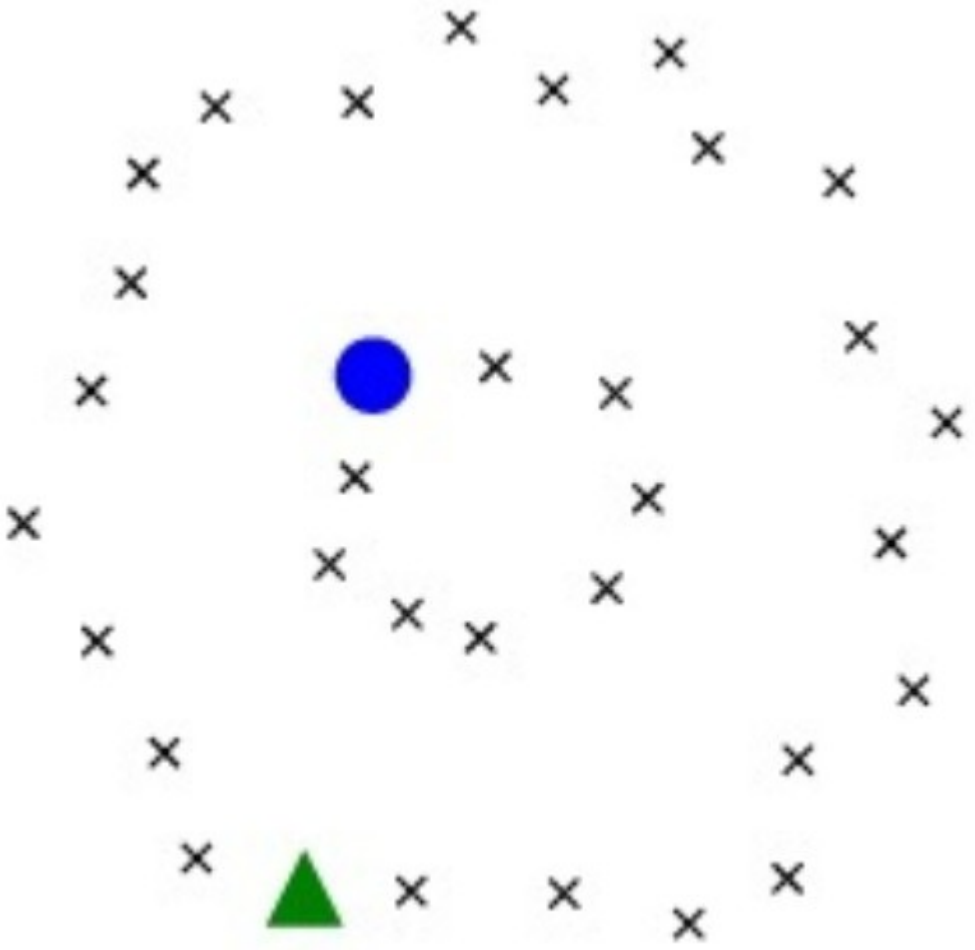


Figure was made by Kristóf Marussy

Example

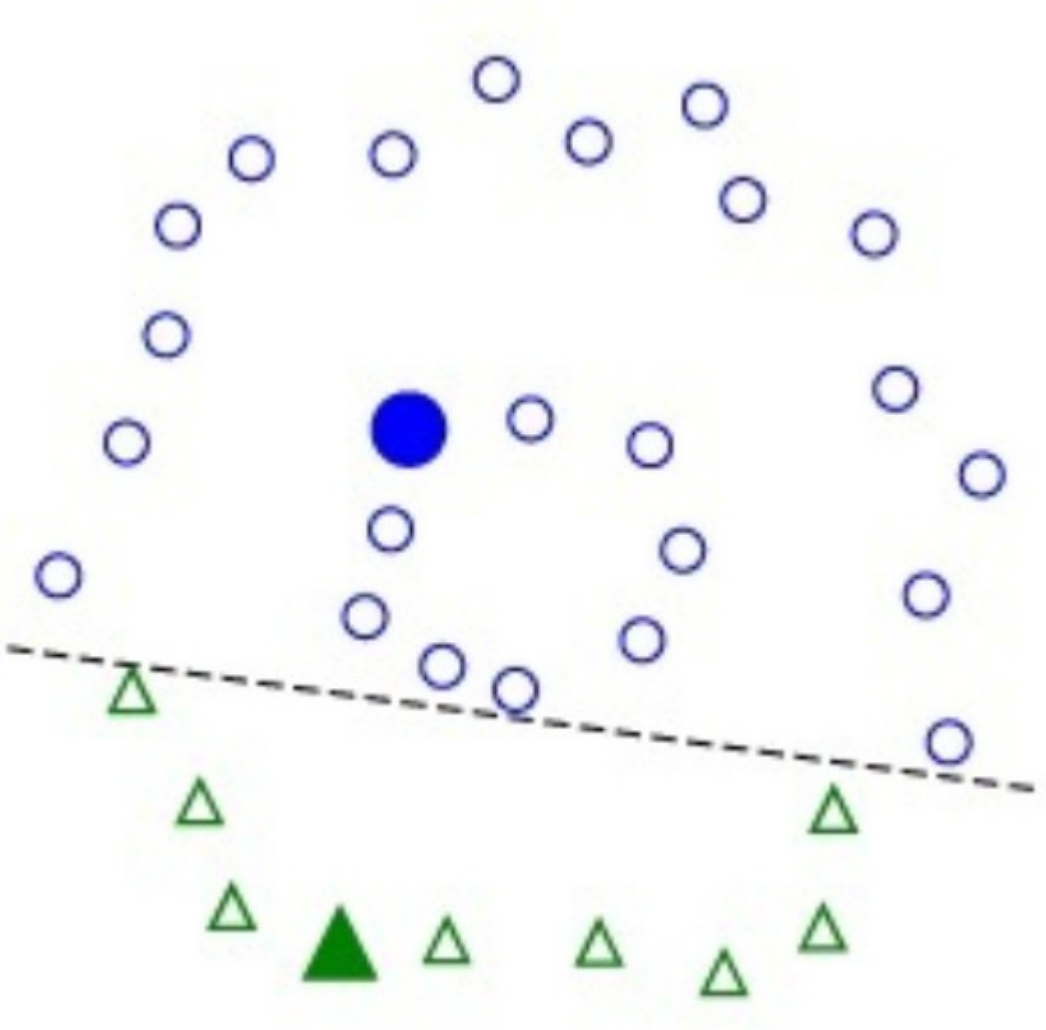


Figure was made by Kristóf Marussy

Example

An instance
that is close
to its nearest
labeled neighbor

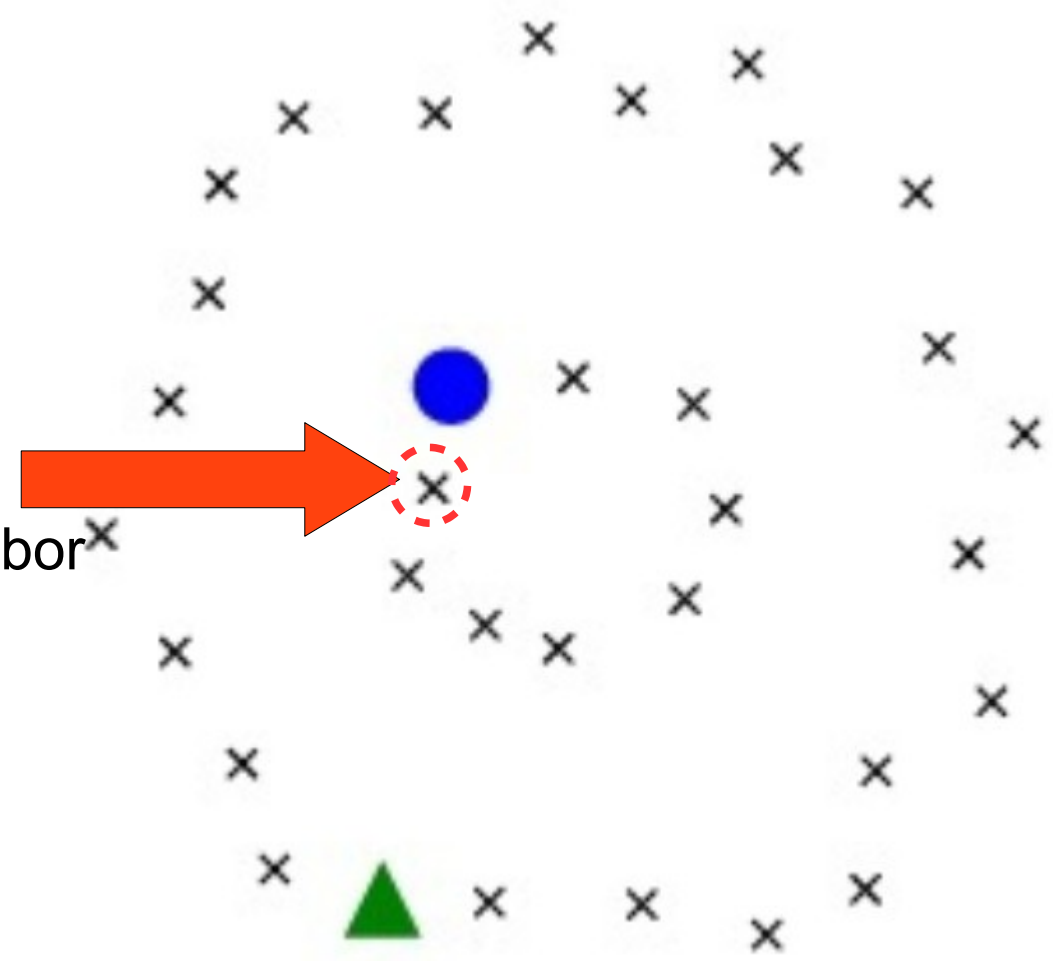
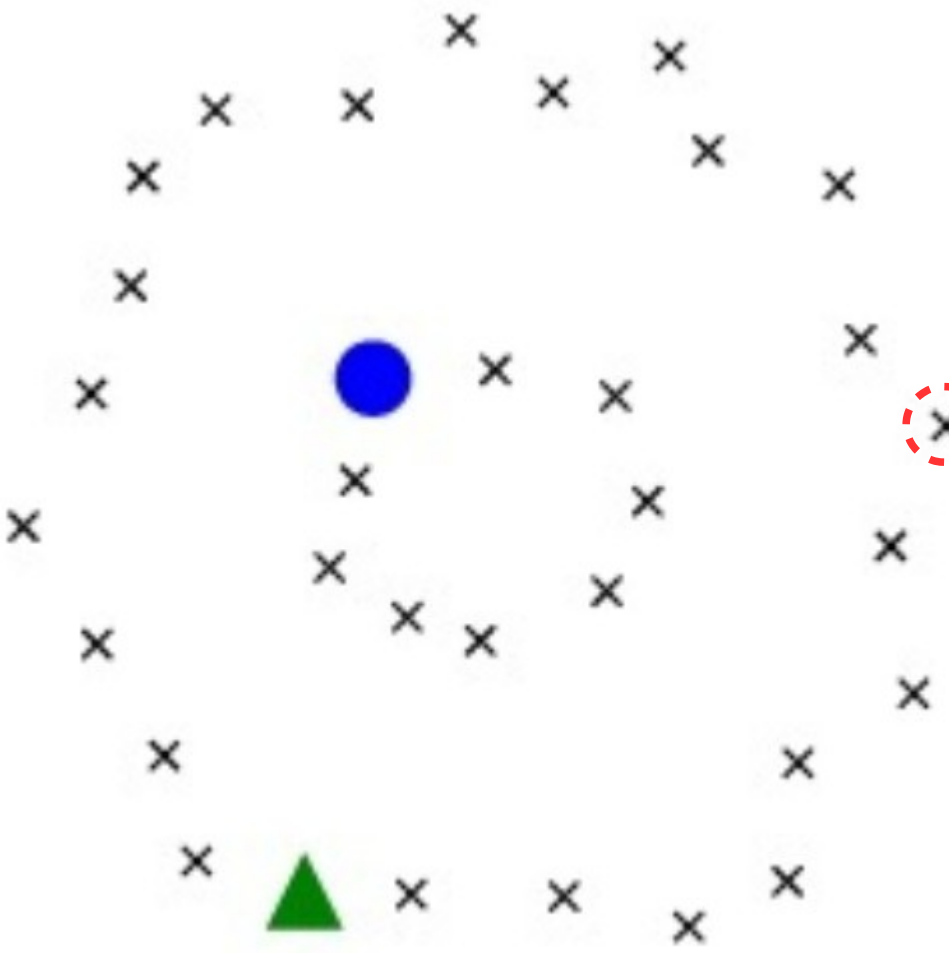


Figure was made by Kristóf Marussy

Example



An instance that is distant from its nearest labeled neighbor

Figure was made by Kristóf Marussy

Example

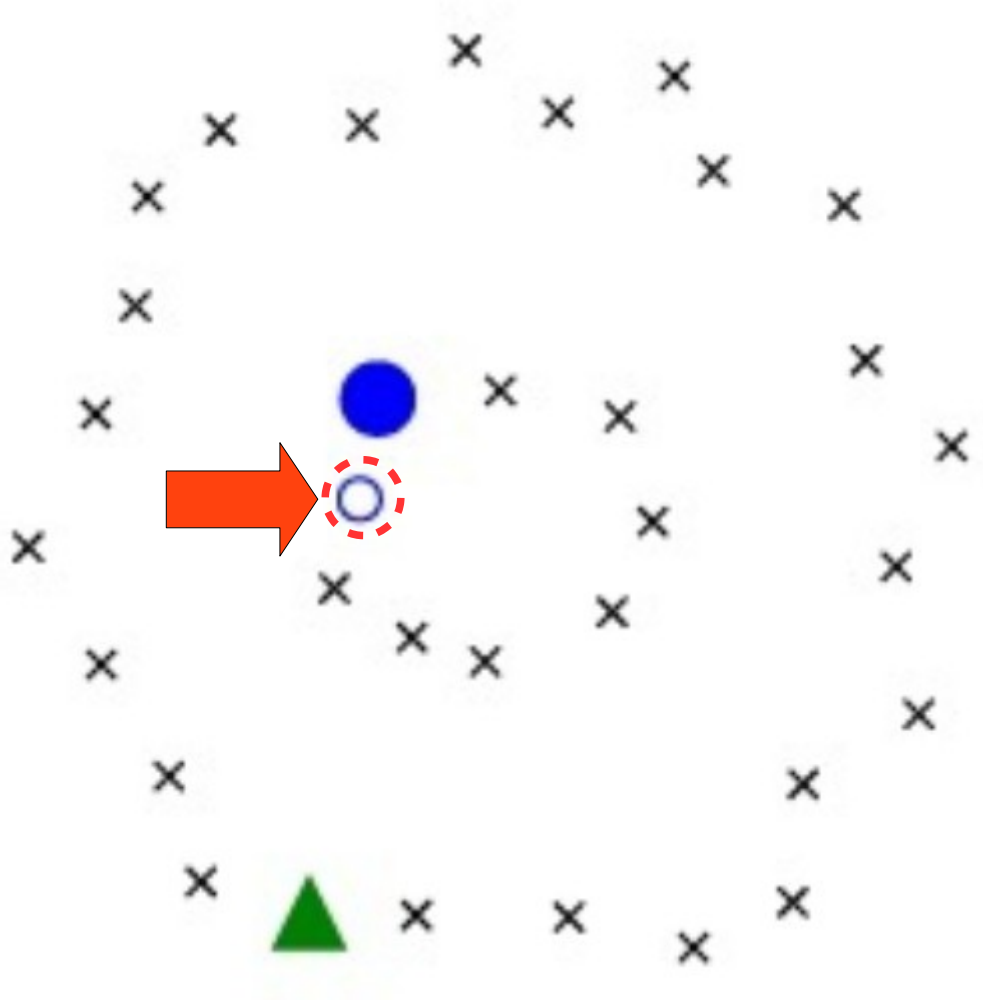


Figure was made by Kristóf Marussy

Example

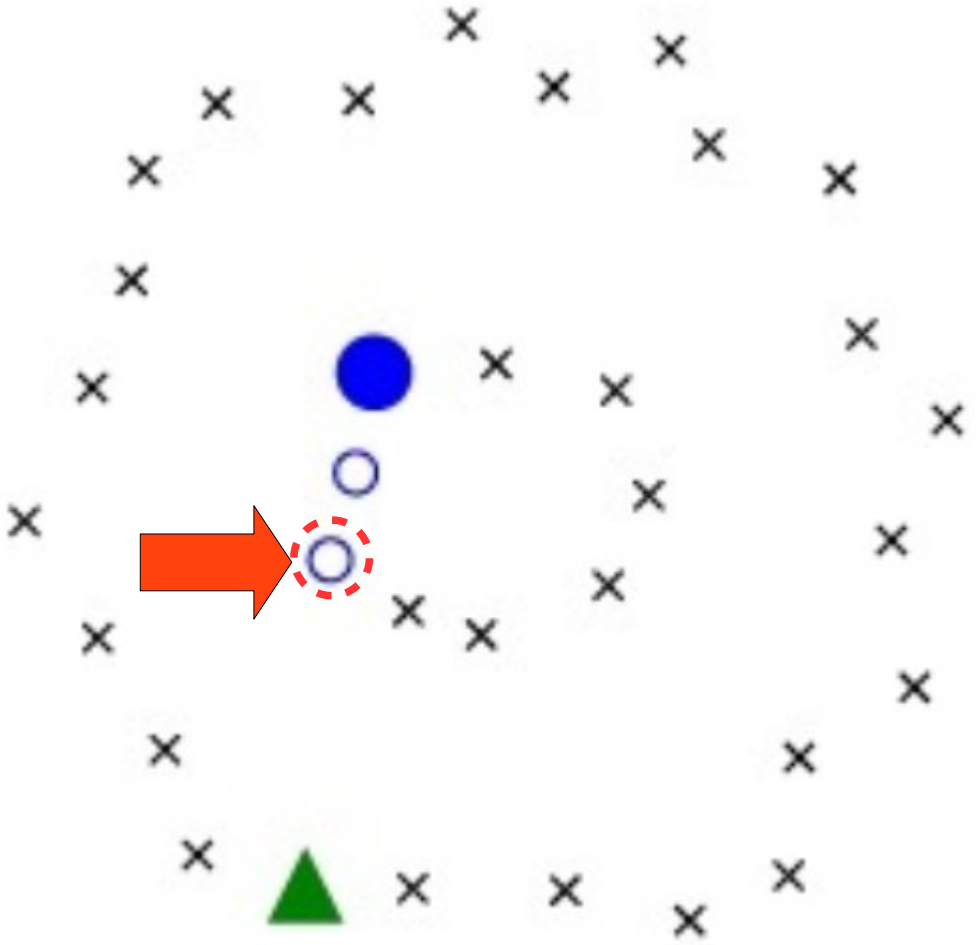


Figure was made by Kristóf Marussy

Example

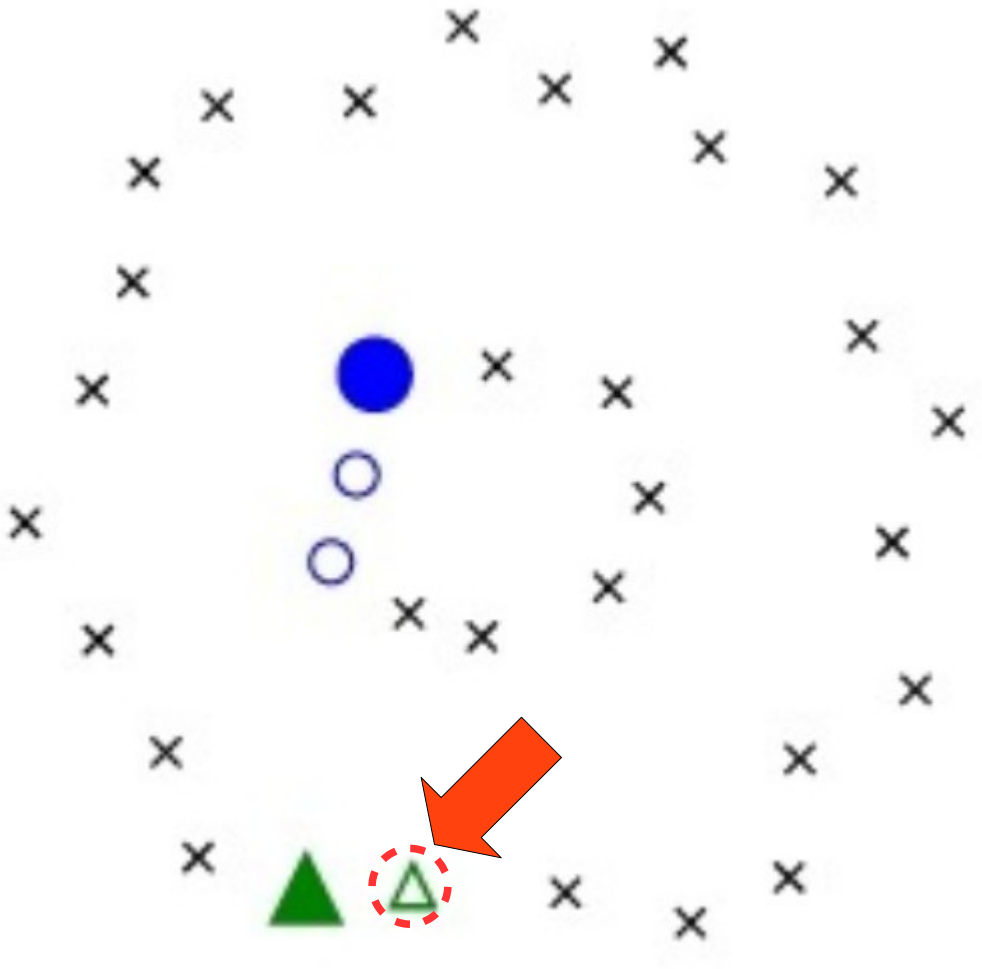


Figure was made by Kristóf Marussy

Example

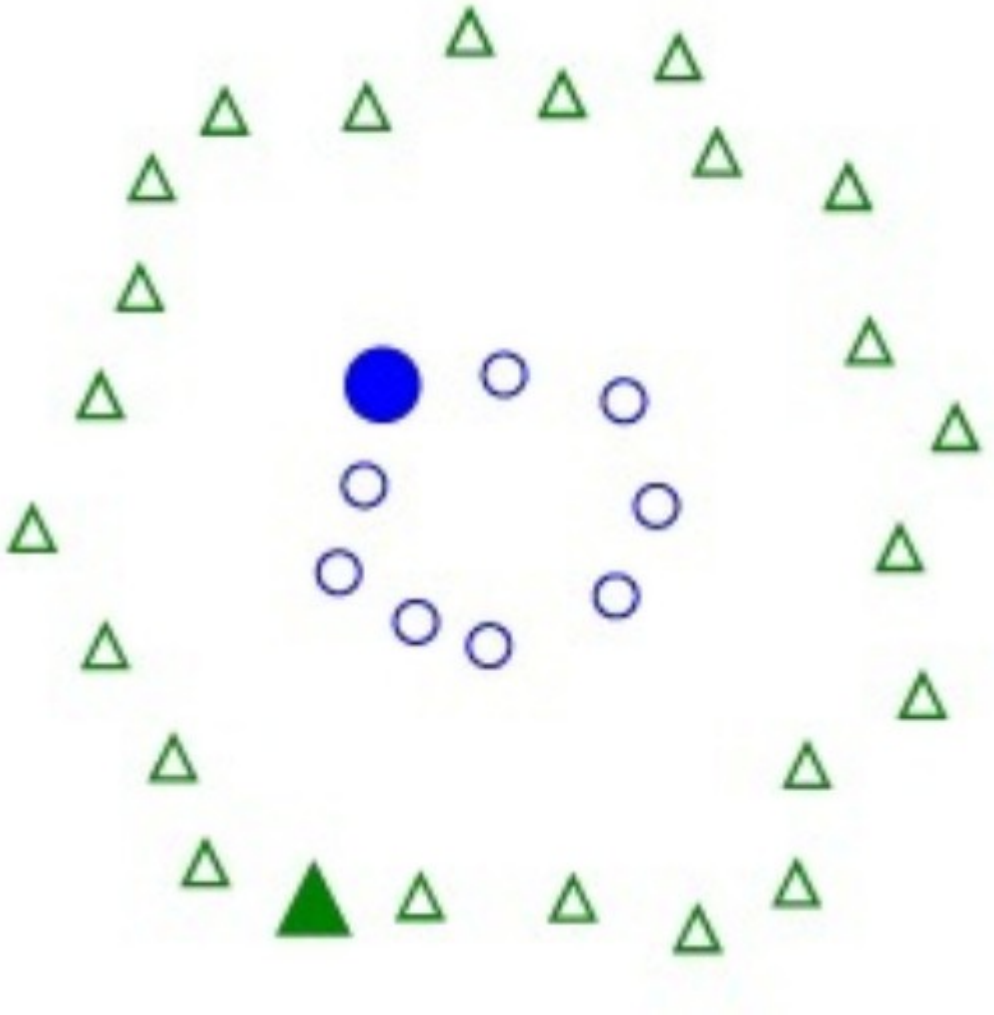
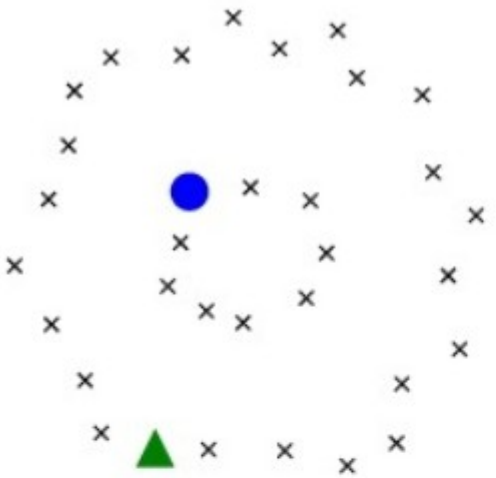
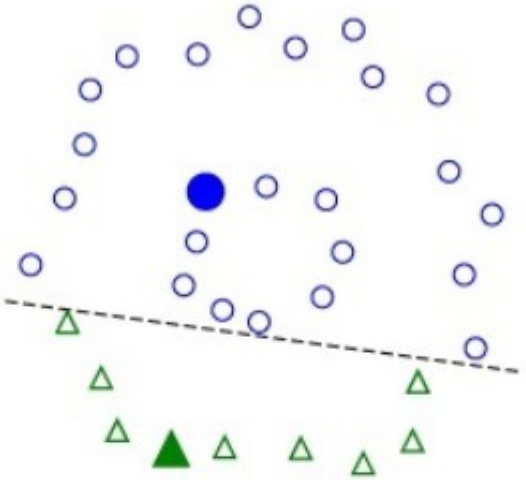


Figure was made by Kristóf Marussy

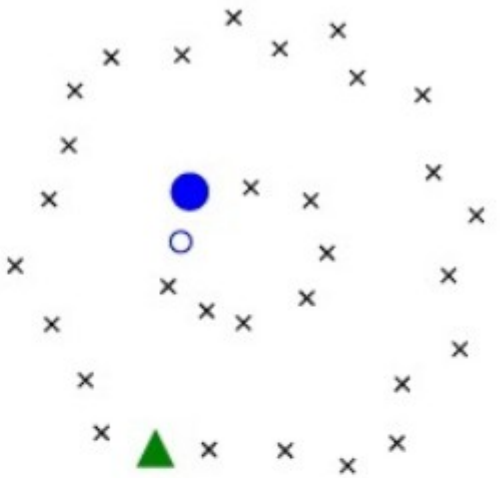
Example



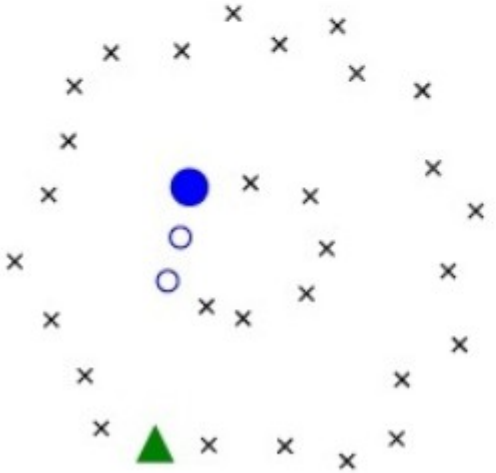
The training set.



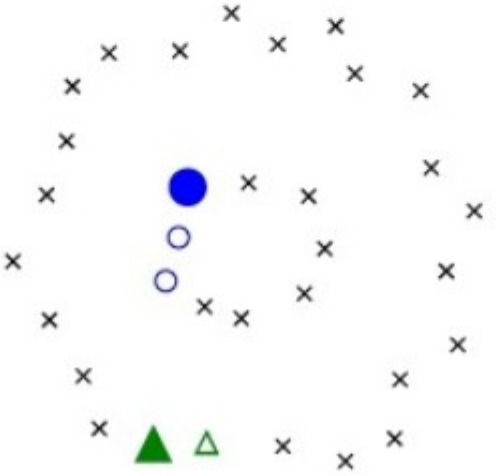
Decision boundary with supervised training.



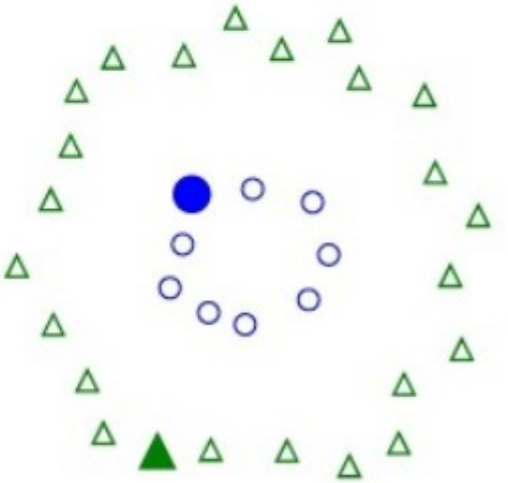
1st iteration of self-training.



2nd iteration of self-training.



3rd iteration of self-training.



Classification with self-training.

Figure was made by Kristóf Marussy

Some works on semi-supervised classification of time series

- Marussy, K., Buza K.: SUCCESS: A New Approach for Semi-Supervised Classification of Time-Series, In: ICAISC, LNCS Vol. 7894. pp. 437-447, Springer (2013)
- Nguyen, M.N., Li, X., Ng, S.K.: Positive unlabeled leaning for time series classification. In: Walsh, T. (ed.) IJCAI. pp. 1421-1426. IJCAI/AAAI (2011)
- Ratanamahatana, C.A., Wanichsan, D.: Stopping criterion selection for efficient semi-supervised time series classication. In: Lee, R.Y. (ed.) Software Engineering, Articial Intelligence, Networking and Parallel/Distributed Computing, Studies in Computational Intelligence, vol. 149, pp. 1-14. Springer (2008)
- Zhong, S.: Semi-supervised sequence classication with HMMs. IJPRAI 19(2), 165-182 (2005)