

Streaming Machine Learning Ensemble Classification

Alessio Bernardo

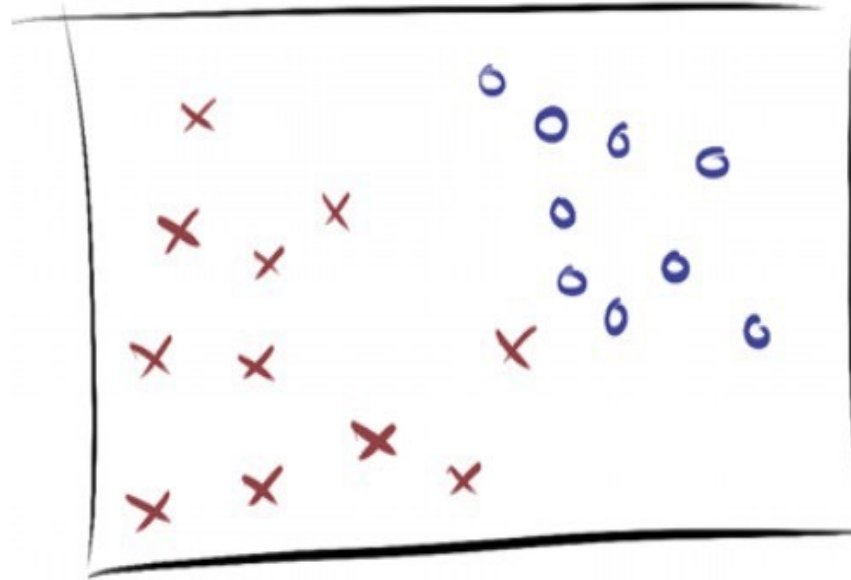
Post-doc @ Politecnico di Milano

CTO & Co-founder @ Motus ml



POLITECNICO
MILANO 1863

SML Ensemble Classification models

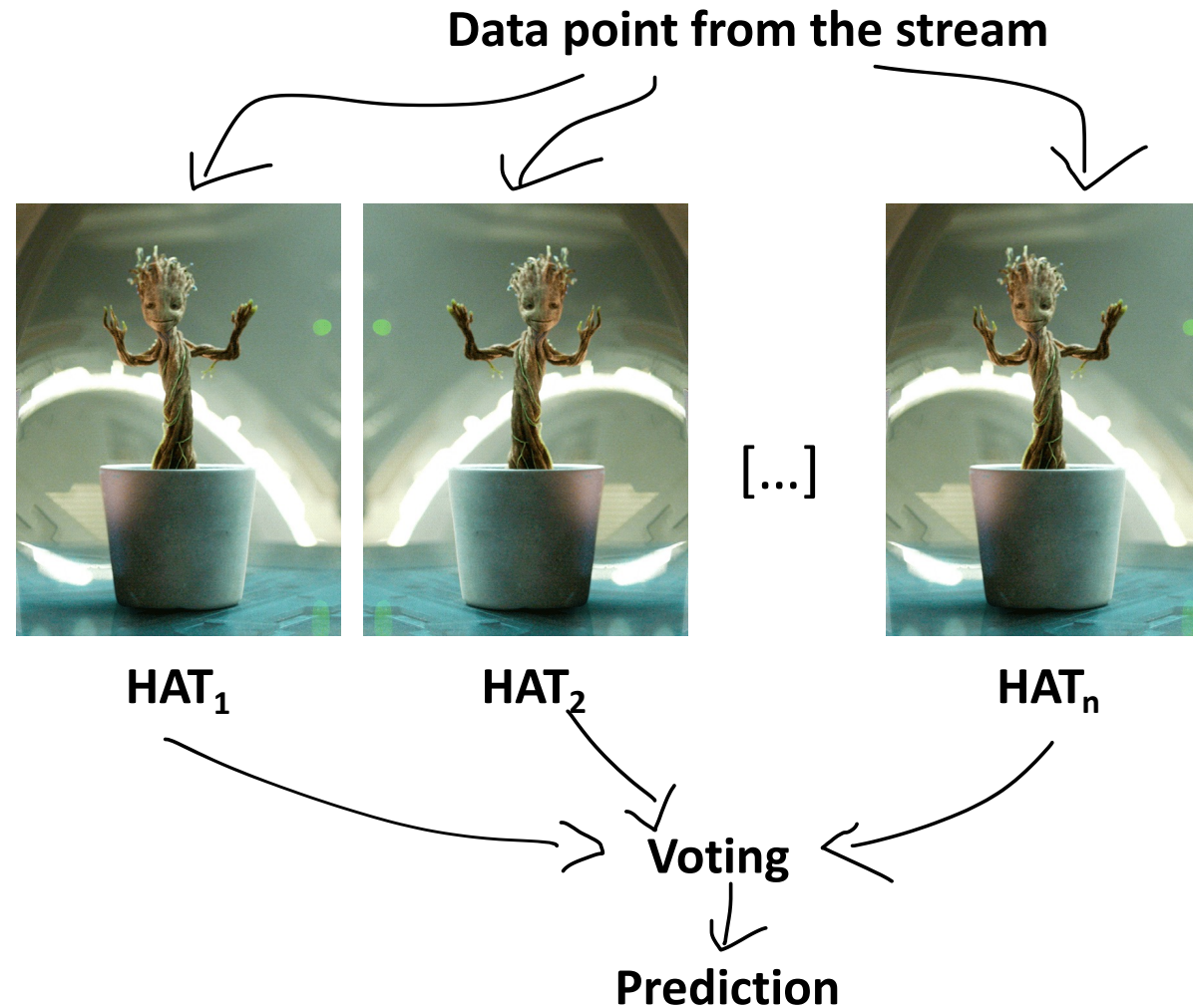


Ensemble Classifiers

*“An **ensemble** can be described as a **composition** of **multiple weak** learners to form one with (expected) **higher** predictive **performance** (strong learner), such that a weak learner is loosely defined as a learner that performs slightly better than random guessing”*

Freund and Schapire, 1997

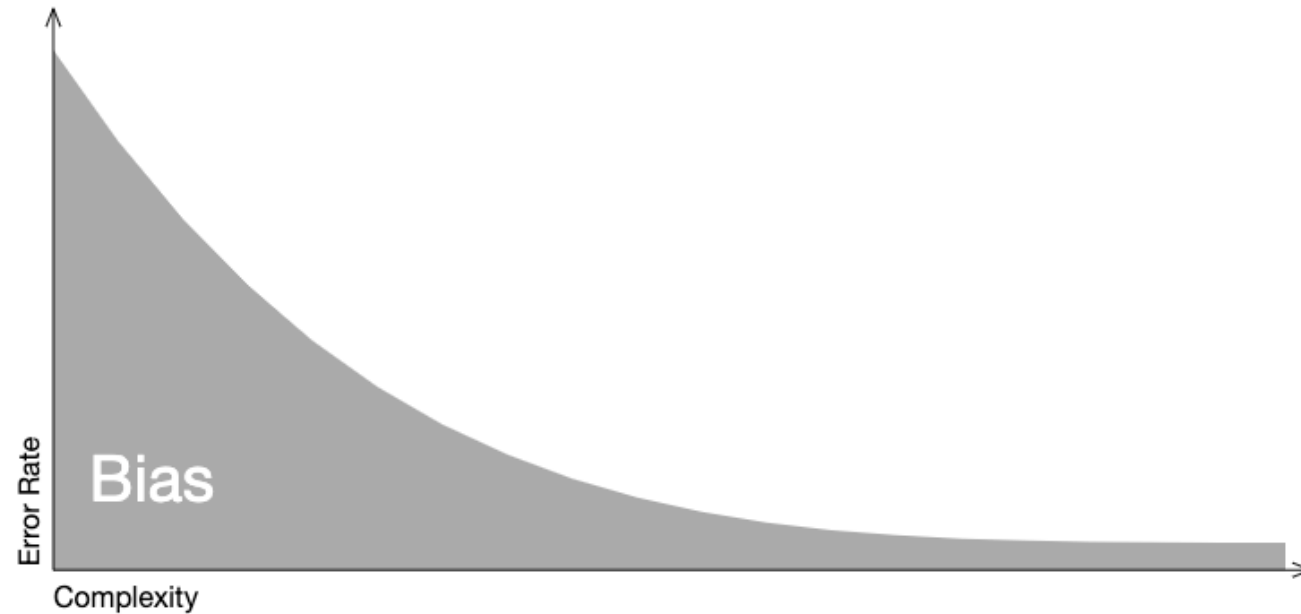
Ensemble Classifiers



Bias-Variance trade-off

Bias

When a model is less complex, it ignores relevant information, and error due to bias is high. As the model becomes more complex, error due to bias decreases.

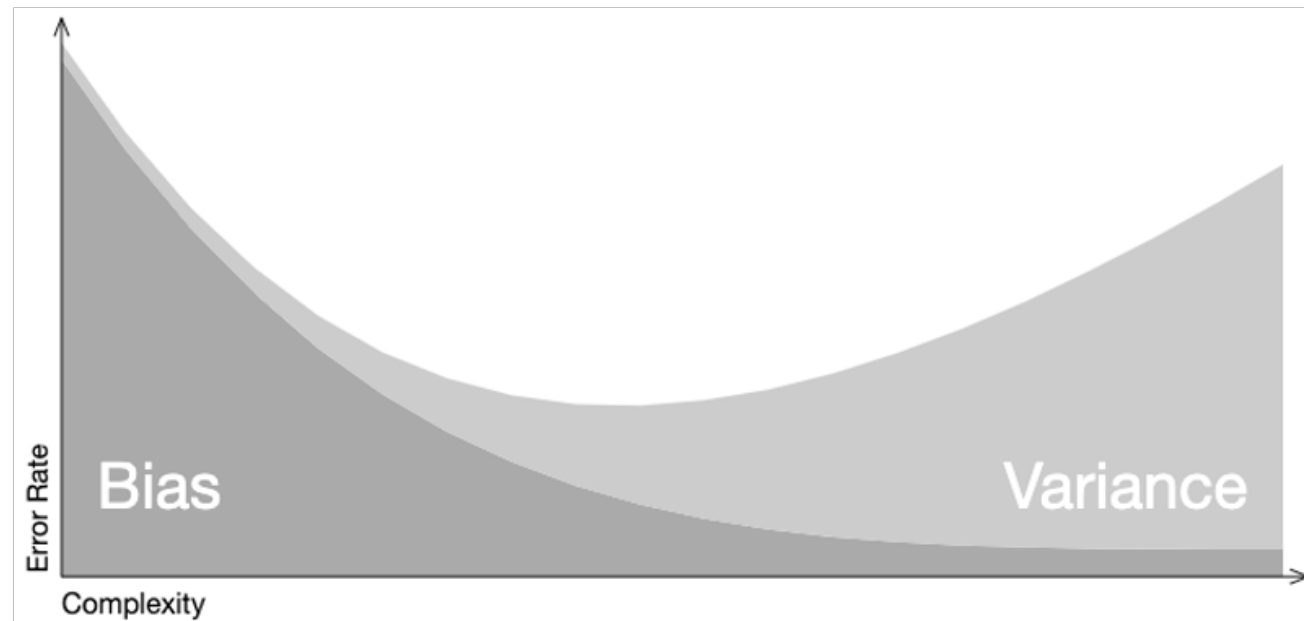


<http://www.r2d3.us/visual-intro-to-machine-learning-part-2/>

Bias-Variance trade-off

Variance

On the other hand, when a model is less complex, error due to variance is low. Error due to variance increases as complexity increases.

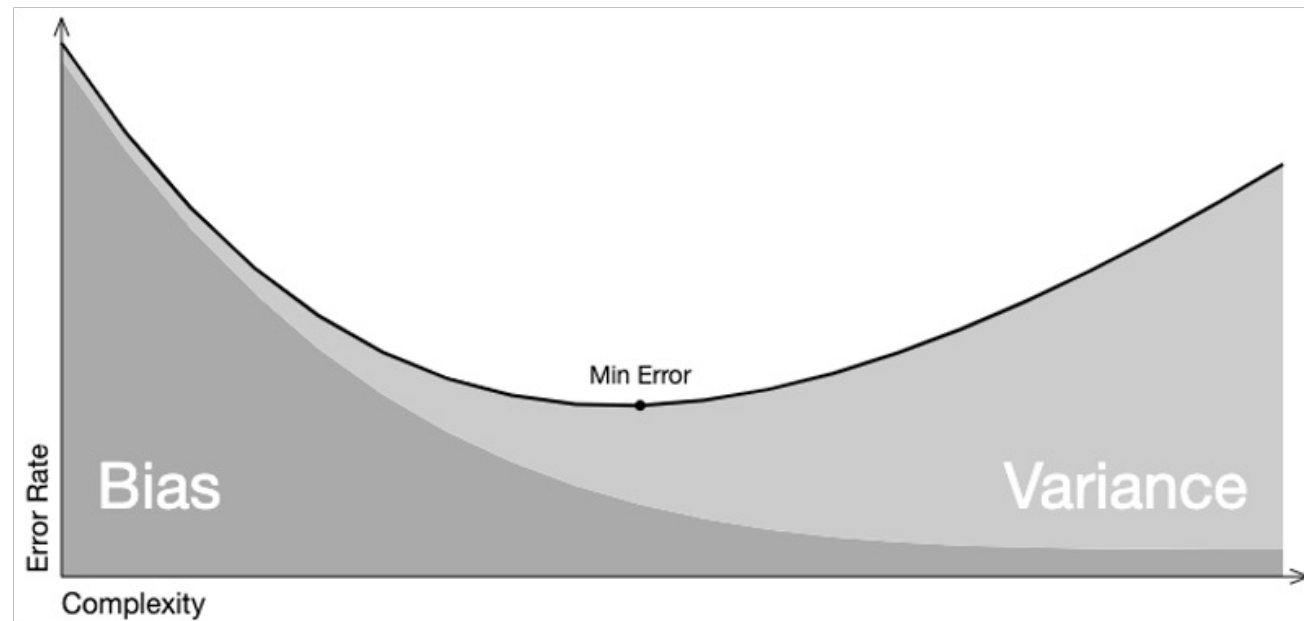


<http://www.r2d3.us/visual-intro-to-machine-learning-part-2/>

Bias-Variance trade-off

Trade-off

Overall model error is a function error due to bias and variance. The ideal model minimized error from each.



<http://www.r2d3.us/visual-intro-to-machine-learning-part-2/>

Ensemble Classifiers in ML

Bagging

- Fits **M** **independent** models and “average” their predictions in order to obtain a model with a **lower variance**...
- But we have only **one** dataset, how can we build **independent** models?

Bootstrapping

- Create **M** bootstrap samples (one for each model) from the original dataset of size **N** , created by drawing random samples with replacement. Each bootstrap contains each original sample **K** times, where **$Pr(K=k)$** follows a binomial distribution.
- **0.632** of the data points in the original sample show up in the bootstrap sample (the other **0.368** won't be present in it)

L. Breiman. **Bagging predictors**. Machine Learning, 1996

Ensemble Classifiers in ML

Bagging → Random Forests

- The **random forest** approach is a **bagging** method where **M** trees, fitted on **bootstrap samples**, are combined to produce an output with lower variance.
- To make the **M** trees a bit less **correlated** with each others: random forest also samples over features and keep only a random subset of them to build the tree.

L. Breiman. **Random Forests**. Machine Learning, 2001

Ensemble Classifiers in ML

Boosting

- **Sequential** method that combines weak models **no longer** fitted **independently** from each others.
- It fits models **iteratively** such that the training of model at a given **step depends** on the models fitted at the **previous steps**: it gives **more importance** to observations in the dataset that were **badly handled** by the **previous** models in the sequence.
- It produces an ensemble model that is in general **less biased** than the weak learners that compose it.

Y. Freund & R. Schapire. **Experiments with a new boosting algorithm**. ICML, 1996

Ensemble Classifiers in ML

Boosting → Adaptive Boosting (AdaBoost)

It puts **more weight** on **difficult** to classify instances and less on those already **handled** well:

- First, it **updates** the observations **weights** in the dataset and train a **new weak learner** with a special **focus** given to the **observations misclassified** by the current ensemble model.
- Second, it **adds** the **weak learner** to the weighted sum according to an **update coefficient** that expresses the performances of this weak model: the **better** a weak learner performs, the **more** it contributes to the strong learner.

Y. Freund & R. Schapire. **Experiments with a new boosting algorithm**. ICML, 1996

Ensemble Classifiers in ML

Boosting → Gradient Boosting

Instead of fitting a weak learner on the data at each iteration, it actually fits a new weak learner to the **residual errors** made by the previous one:

- For every instance in the training set, it calculates the **residuals** for that instance, or, in other words, the **observed value minus the predicted value**.
- Once it has done this, it **adds** a **weak learner** that tries to **predict** the **residuals** that was previously calculated.

J. H. Friedman. **Stochastic gradient boosting**. Computational statistics & data analysis, 2022

Ensemble Classifiers in ML

Stacking

- It considers heterogeneous weak learners (different learning algorithms are combined).
- It learns to combine the base models using a meta-model.
- It produces an ensemble model that is in general **less biased** than the weak learners that compose it.

K. M. Ting & I. H. Witten. **Stacking bagged and dagged models**. 1997

Ensemble Classifiers in SML

- **Diversity:** induce diversity among learners
- **Combination:** combine the predictions
- **Adaptation:** adapt to evolving data

Pro

- High Predictive performance
- Flexibility

Cons

- Computational resources

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification.** ACM, 2017

Induce Diversity

Horizontal Partitioning

- **Bagging**: build a set of M base models, with a bootstrap sample from the original dataset of size N , created by drawing random samples with replacement. Each bootstrap contains each original sample K times, where $Pr(K=k)$ follows a binomial distribution.

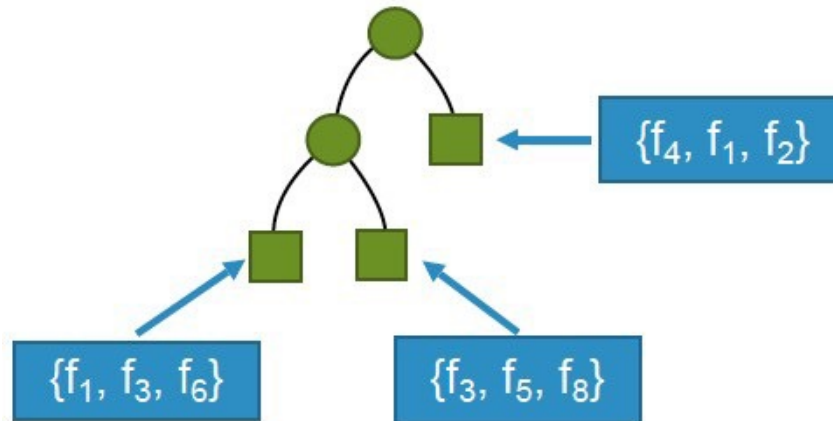
Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification**. ACM, 2017

Induce Diversity

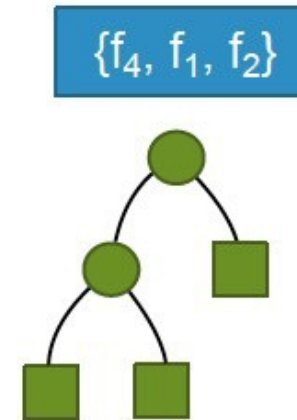
Vertical Partitioning

- **Random Subspaces:** train learners on different subsets of features

Local Randomization



Global Randomization



Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification.** ACM, 2017

Induce Diversity

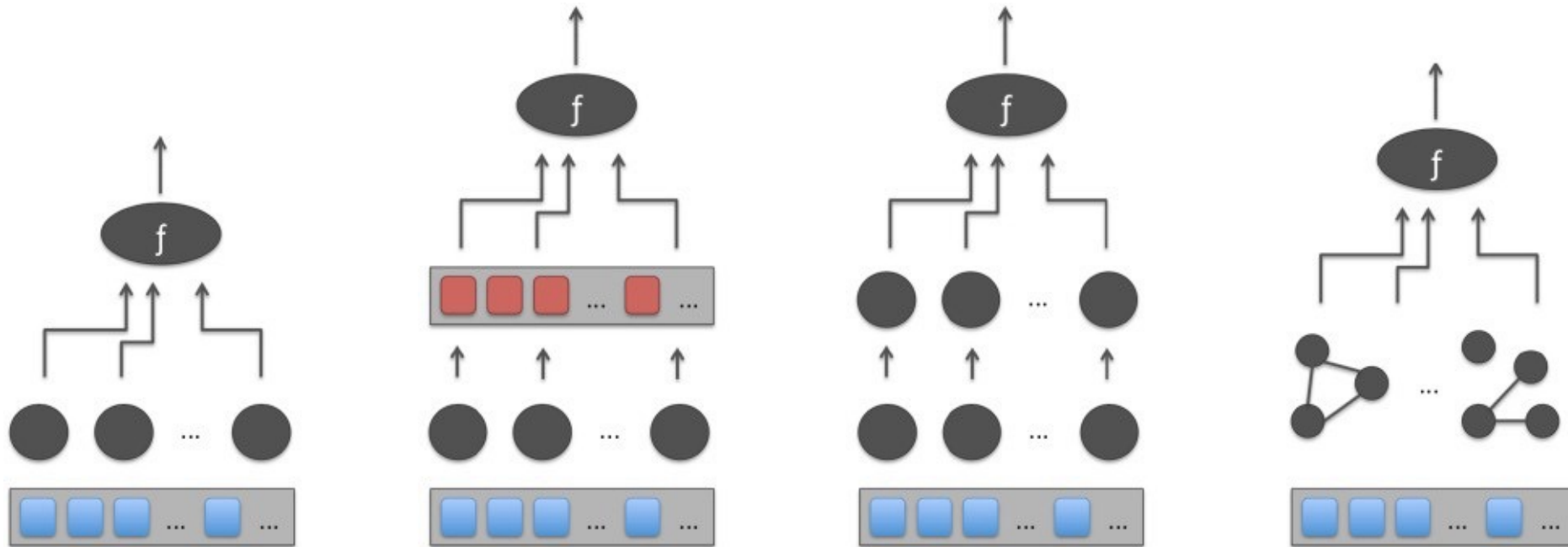
Others

- **Base Learner Manipulation:** varying parameters of the same base learner
- **Heterogeneous Base Learners (Stacking):** use heterogeneous base learners and obtain ensemble members with different biases

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification.** ACM, 2017

Combination Architecture

● Base learners □ Instances



Flat

Meta-Learner

Hierarchical

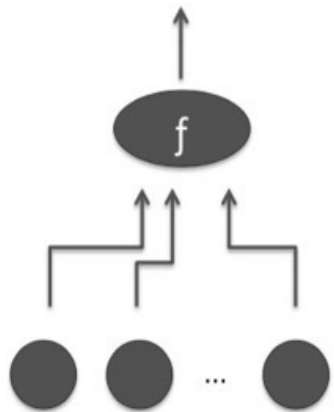
Network

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification.** ACM, 2017

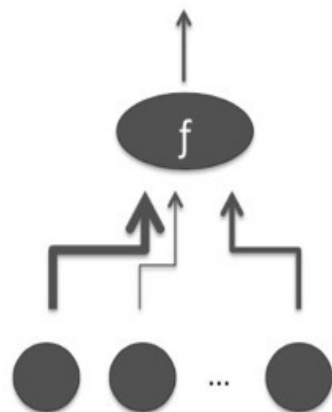
Combination

Voting

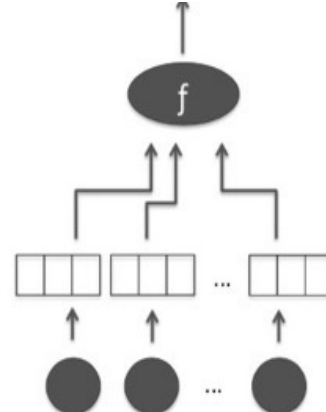
● Base learners □ Instances



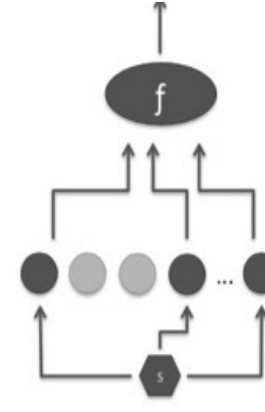
Majority



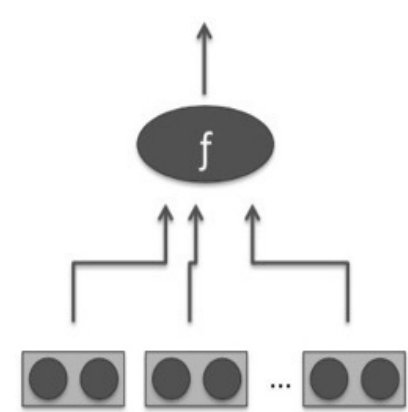
Weighted Majority



Rank



Abstaining



Relational

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification.** ACM, 2017

Adaptation

Cardinality

- **Fixed:** fixed numbers of base learners
- **Dynamic:** add classifiers on the fly

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. **A survey on ensemble learning for data stream classification**. ACM, 2017

Online Bagging

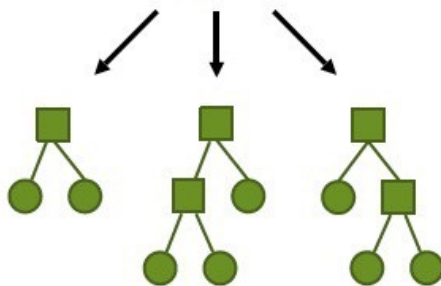
- Since data streams are supposed to be unbounded (large N), the binomial distribution tends to a **Poisson(1)** distribution.



For each learner...

$$k = \text{Poisson}(\lambda = 1)$$

Train model using (x^t, y^t)
with weight k

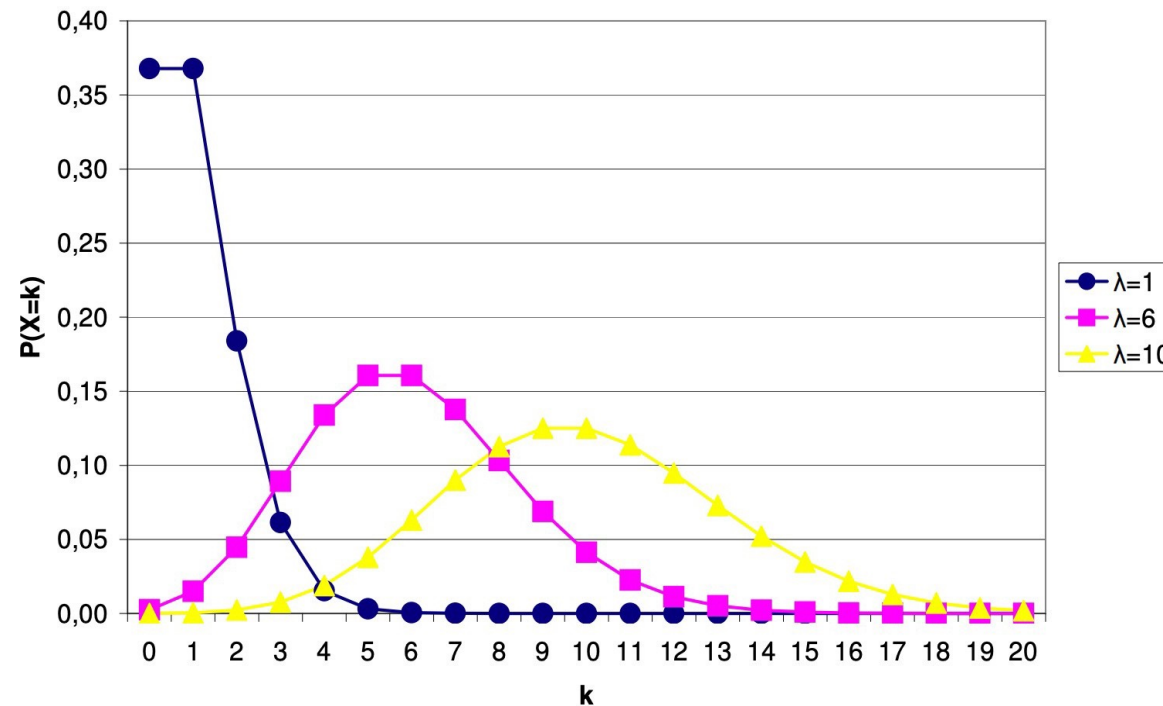


**Train learners on different
subsets of instances**

Oza and Russel **Online bagging and boosting**. Artificial Intelligence and Statistics, 2001

Leveraging Bagging

- Add an **ADWIN** drift detector per base learner
- Use more weight during training - **Poisson(6)**



Bifet, G. Holmes, and B. Pfahringer **Leveraging bagging for evolving data streams**. PKDD, 2010

Adaptive Random Forest (ARF)

- **Base Learners:** Hoeffding Trees
- **Diversity:** Leveraging Bagging + Local Random Subspaces
- **Combination:**
 - Flat architecture
 - Weighted majority voting
- **Adaptation:** Adaptive window + warning period (train background learners)

H. M. Gomes et al. **Adaptive random forests for evolving data stream classification**. Machine Learning, 2017

Streaming Random Patches (SRP)

- **Base Learners:** User choice
- **Diversity:** Leveraging Bagging + **Global** Random Subspaces
- **Combination:**
 - Flat architecture
 - Weighted majority voting
- **Adaptation:** Adaptive window + warning period

Gomes, Read and Bifet. **Streaming Random Patches for Evolving Data Stream Classification**. ICDM, 2019



Exercise 4: Stream Ensemble Classification





Credits

- Albert Bifet DATA STREAM MINING 2020-2021 course at Telecom Paris
- Alessio Bernardo & Emanuele Della Valle

Streaming Machine Learning Ensemble Classification

Alessio Bernardo

Post-doc @ Politecnico di Milano

CTO & Co-founder @ Motus ml



POLITECNICO
MILANO 1863