# Concept drift adaptation

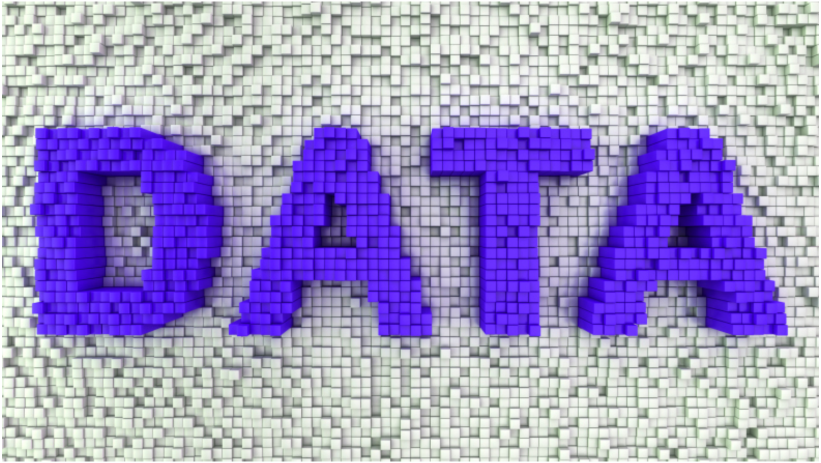
Author: Ward Schodts
Dozent: Pieter-Jan Kindermans

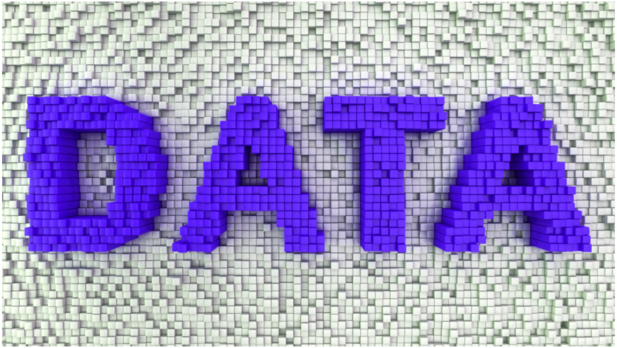
Hot Topics in Machine Learning
Technische Universität Berlin

July 18, 2016

# Agenda

# Introduction






Total volume of data generated: $> 3$ zetabytes

# Introduction

- Traditionally all this data is processed in an *offline* mode

# Introduction

- Traditionally all this data is processed in an *offline* mode
- Due to the big amount of data
- And the expansion in forms of streams

# Introduction

- Traditionally all this data is processed in an *offline* mode
- Due to the big amount of data
- And the expansion in forms of streams
- We can’t do this anymore if we want to keep up.

- Traditionally all this data is processed in an *offline* mode
- Due to the big amount of data
- And the expansion in forms of streams
- We can't do this anymore if we want to keep up.

$\rightarrow$ We need an online solution

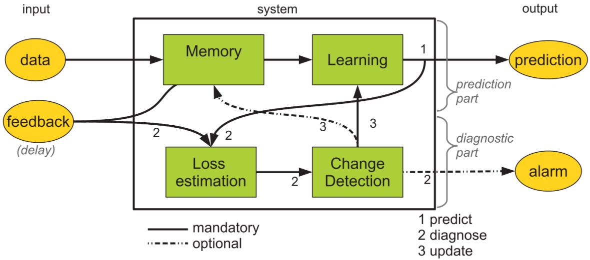
# Agenda

***Adaptive Learning Algorithm:***
*can be seen as advanced incremental learning algorithms that are able to adapt to evolution of the data-generating process over time.*

It needs to be able to:

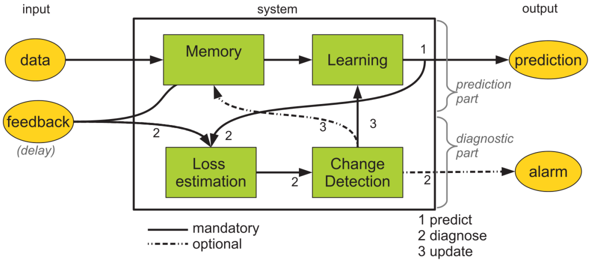
1. detect concept drift (and adapt if needed) as soon as possible;
2. distinguish drifts from noise and be adaptive to changes, but robust to noise; and
3. operate in less than example arrival time and use not more than a fixed amount of memory for any storage.

# Online Adaptive Learning Procedure



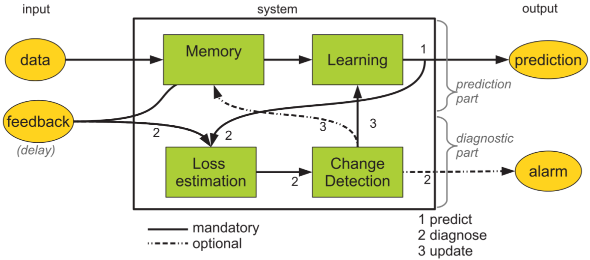

1. *Predict.* When new example $X_t$ arrives, a prediction $\hat{y}_t$ is made using the curren model $L_t$.

# Online Adaptive Learning Procedure




1. *Predict.* When new example $X_t$ arrives, a prediction $\hat{y}_t$ is made using the curren model $L_t$.
2. *Diagnose.* After some time, the true label $y_t$ is received and we can estimate the loss as $f(\hat{y}_t, y_t)$.

# Online Adaptive Learning Procedure




1. *Predict.* When new example $X_t$ arrives, a prediction $\hat{y}_t$ is made using the curren model $L_t$.
2. *Diagnose.* After some time, the true label $y_t$ is received and we can estimate the loss as $f(\hat{y}_t, y_t)$.
3. *Update.* We can use the example $(X_t, y_t)$ for the model update to obtain $L_{t+1}$.

# Concept drift I

***Concept drift:***
*Formally, concept drift between time point $t_0$ and time point $t_1$ can be defined as:*

$$\exists X : p_{t_0}(X, y) \neq p_{t_1}(X, y)$$

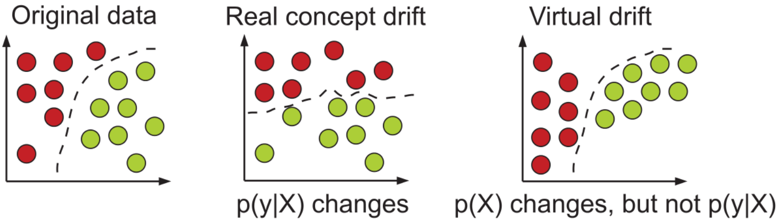
*where $p_{t_0}$ denotes the joint distribution at time $t_0$ between the set of input variables $X$ and the target variable $y$.*

$$\exists X : p_{t_0}(X, y) \neq p_{t_1}(X, y)$$

***Real concept drift:***
*refers to changes in* $p(y|X)$*. Such changes can happen either with or without change in* $p(X)$*. It's also called concept shift or conditional change.*

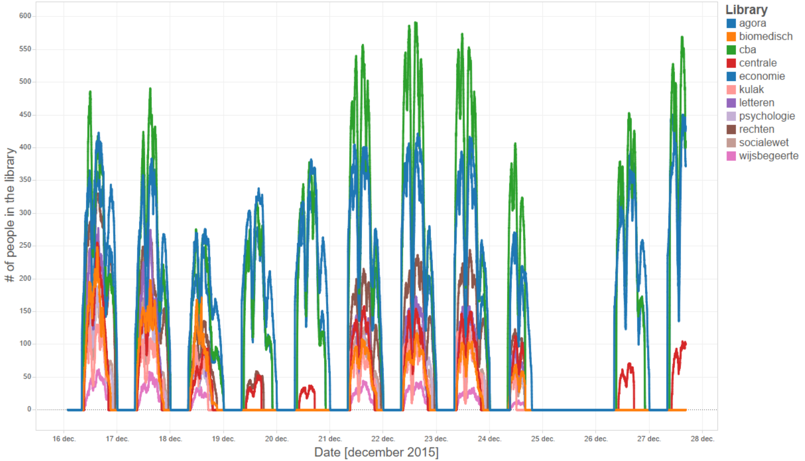
***Virtual concept drift:***
*happens if the distribution of the incoming data chagnes (i.e., p(X) changes) without affecting* $p(y|X)$*.*

# Example: real vs. virtual drift



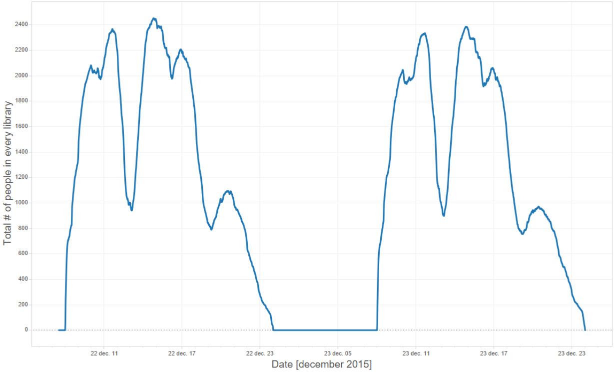

*Circles* represent instances.
Different *colors* represent different classes.

# A practical example: # of people in a library




# of people in the KU Leuven libraries

# A practical example: # of people in a library

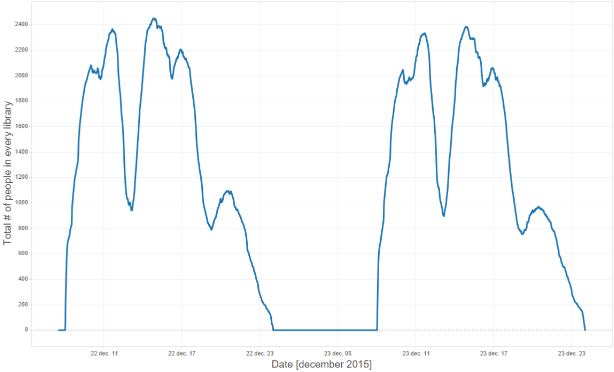



Sum of people in all KU Leuven libraries

**Task:** predict the total amount of people in all libraries

If same amount of people go to different libraries:

# A practical example: # of people in a library

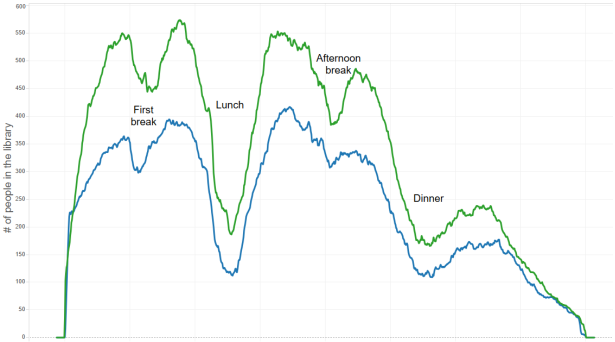



Sum of people in all KU Leuven libraries

**Task:** predict the total amount of people in all libraries

If same amount of people go to different libraries: $\rightarrow$ virtual drift

# A practical example: # of people in a library

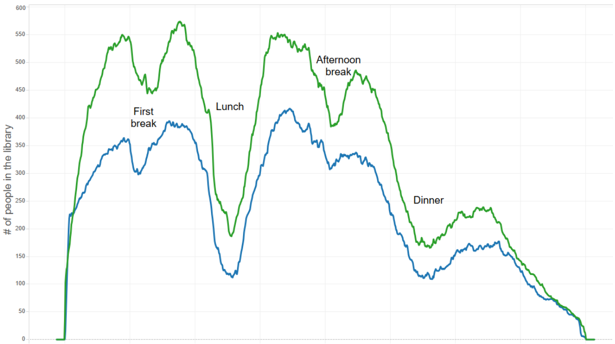



Sum of people in all KU Leuven libraries

**Task:** predict percentage of people in a specific library on a monday

If some people go to different libraries:

# A practical example: # of people in a library

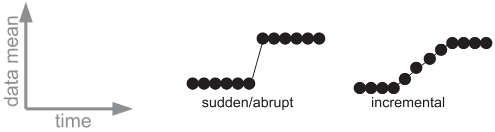



Sum of people in all KU Leuven libraries

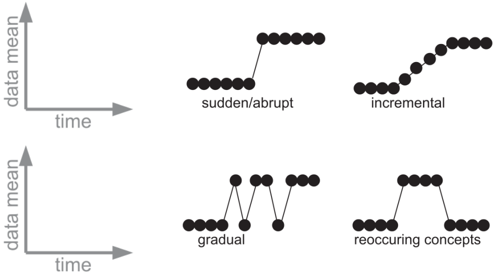
**Task:** predict percentage of people in a specific library on a monday

If some people go to different libraries: $\rightarrow$ real drift

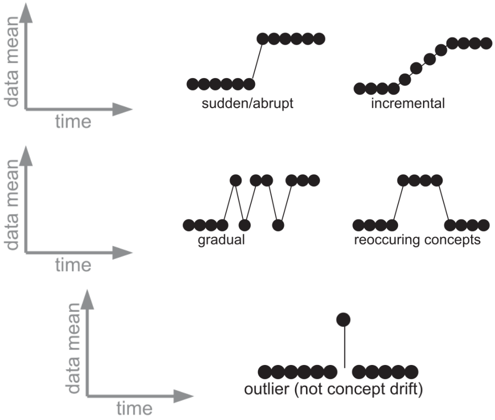
# Changes in drift over time



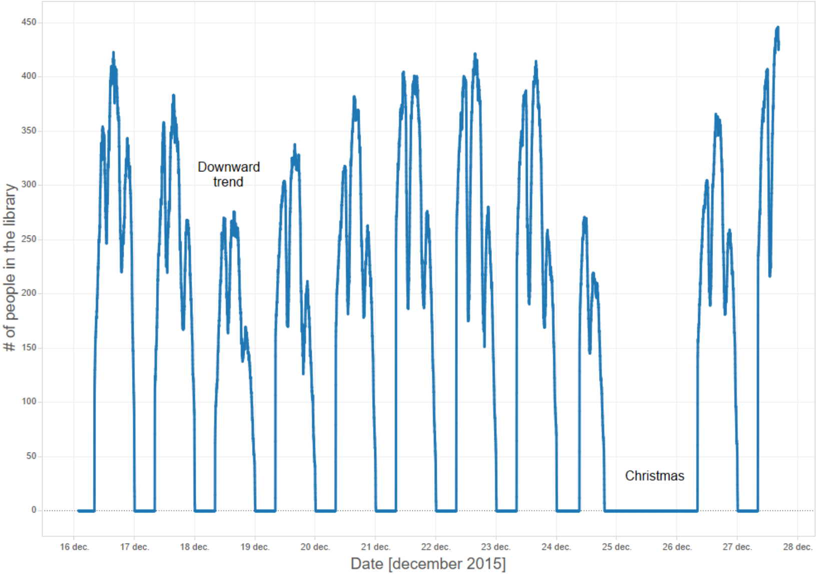
# Changes in drift over time



# Changes in drift over time



# Changes in drift over time: library example



# Illustrative Applications

- Management and strategic planning

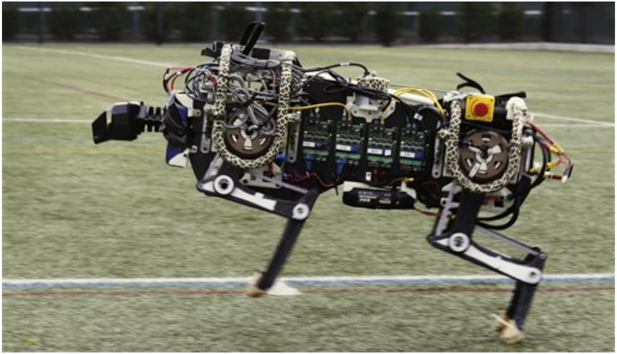
- Personal assistance and information
- Ubiquitous environment applications

# Illustrative Applications

- Management and strategic planning
- Personal assistance and information
- Ubiquitous environment applications

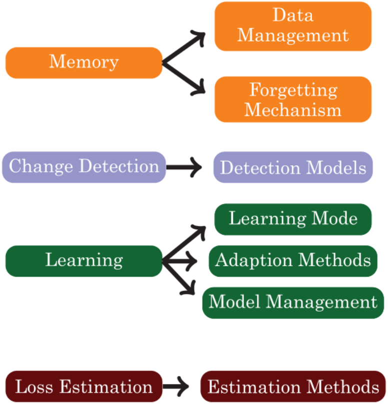
# Illustrative Applications

- Management and strategic planning
- Personal assistance and information
- Ubiquitous environment applications

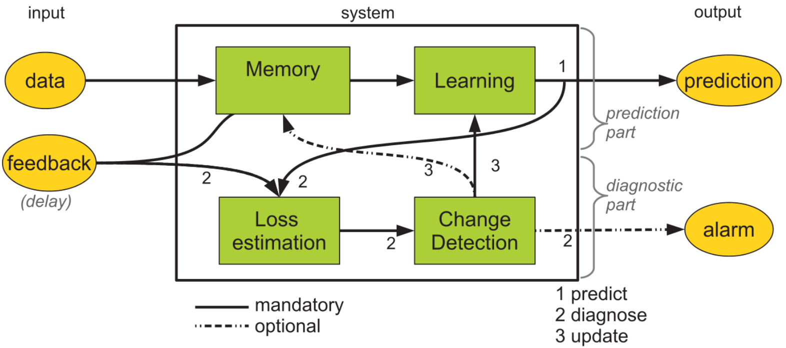


# Agenda

# Methods for concept drift adaptation

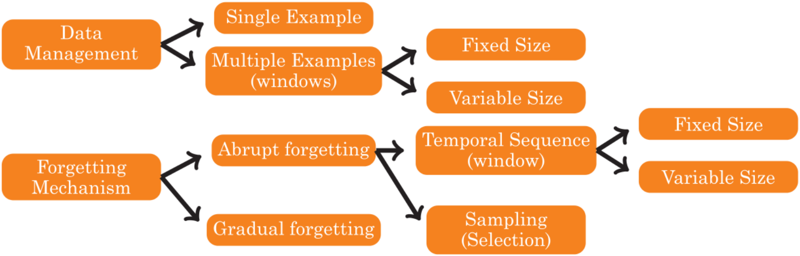


# Methods for concept drift adaptation

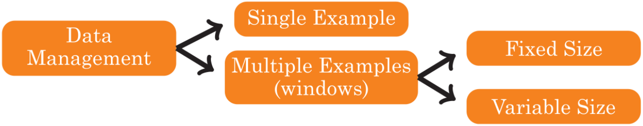


# Agenda

# Memory in concept drift Adaptation



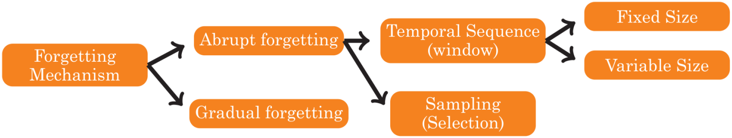
# Data Management




Often is assumed that the most recent data is the most informative.

- **Single example:** almost no memory needed, more error prone. Sometimes necessary because of speed and memory constraints (e.g. in online learning and streaming)
- **Multiple examples:** a set of recent examples:
  - *Fixed size:* often used as baseline
  - *Variable size:* depending on idication of a change detector

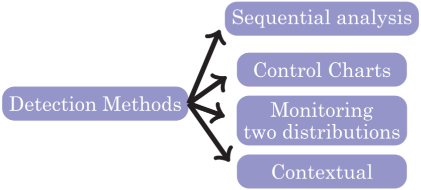
# Forgetting mechanism




- **Abrupt forgetting:** a window considered for learning.
- **Gradual forgetting:** examples are not really discarded, they are associated with weights.

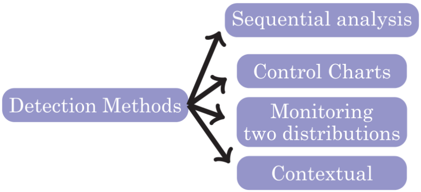
# Agenda

# Change detection



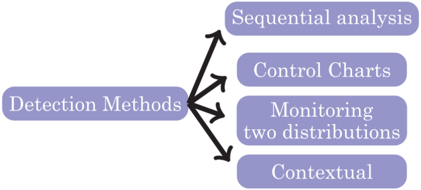

- **Sequential analysis:** detectors use sequential analysis e.g. *Sequential Probability Ratio Test*
- **Statistical Process Control (SPC):** standard statistical techniques to predict change.
- **Monitoring distributions on two different time windows:**

# Change detection




- **Sequential analysis:** detectors use sequential analysis e.g. *Sequential Probability Ratio Test*
- **Statistical Process Control (SPC):** standard statistical techniques to predict change.
- **Monitoring distributions on two different time windows:**
  - one fixed reference window (summarizes the past information)
  - one sliding window over the the most recent examples
  - distributions are compared using statistical tests

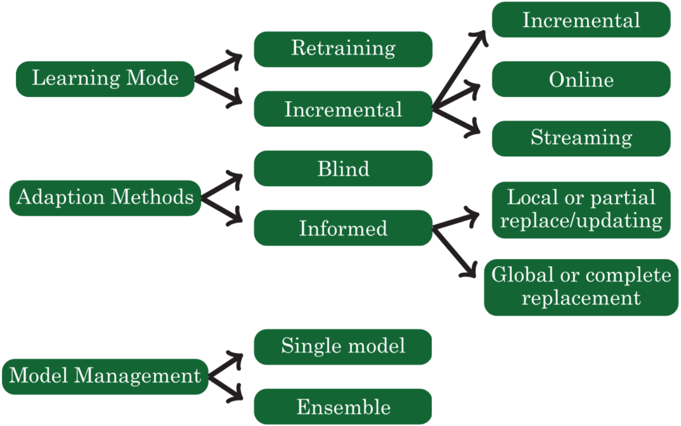
## Change detection




- **Sequential analysis:** detectors use sequential analysis e.g. *Sequential Probability Ratio Test*
- **Statistical Process Control (SPC):** standard statistical techniques to predict change.
- **Monitoring distributions on two different time windows:**
- **Contextual approaches:** a learning techinique that identifies intervals with stable hidden context

# Agenda

# Learning


Learning Mode
Retraining
Incremental
Incremental
Online
Streaming
Adaption Methods
Blind
Informed
Local or partial replace/updating
Global or complete replacement
Model Management
Single model
Ensemble

# Learning mode


If new labeled examples are available the model might be updated:

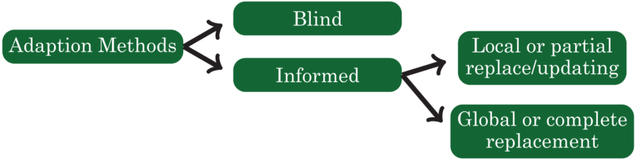
- **Retraining:** start from scratch, using bufferd data.

# Learning mode


If new labeled examples are available the model might be updated:

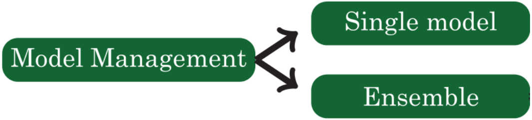
- **Retraining:** start from scratch, using bufferd data.
- **Incremental:** update the model using most recent data:
  - **Incremental:** modify and update statistics in the model upon arrival from new examples (can still have acces to old values)
  - **Online:** update the current model with most recent example $\rightarrow$ errorprone if last example is misclassified
  - **Streaming:** are online algorithms for a high-speed continious flow of data. Limited memory and normally only one pass over the data.

# Adaption Methods




- Manage the adaptation of the predictive model
  - **Blind:** without explicit detection of changes. (fixed-size sliding windows)
  - **Informed:** action depend triggers:
    - *Global Replacement:* the model is rebuild (linear regression, naive Bayes...)
    - *Local Replacement:* if the model is modular (e.g. a decision tree) only a part (e.g. a subtree) is replaced

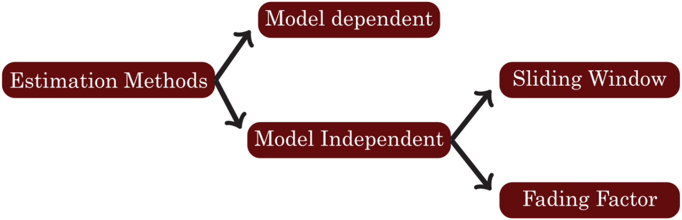
# Adaption Methods



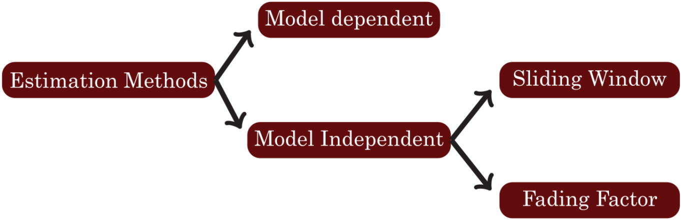

- **Single model:** only a single model is used
- **Ensemble:** multiple models are used together and do voting

# Agenda

# Loss estimation in concept drift Adaptation

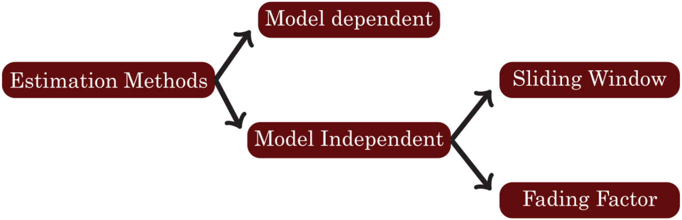


# Loss estimation in concept drift Adaptation




- **Model Dependent**
  - Estimator needs to learn from the current model
  - e.g. Klinkenbeg and Joachims [4] use it in SVM

# Loss estimation in concept drift Adaptation




- **Model Independent**
  - Can be applied immediately
  - **Sliding Window:** one small (quick) and one large (slower) together
  - **Fading Factor:** a small FF detects change earlier

# Agenda

# Evaluation

For any machine-learning technique we want to evaluate we need:

1. Performance evaluation metrics chosen according to the goal of a learning task

# Evaluation

For any machine-learning technique we want to evaluate we need:

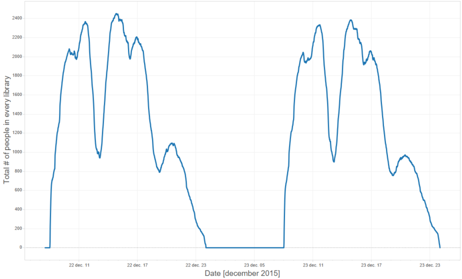
1. Performance evaluation metrics chosen according to the goal of a learning task
2. A methodology allowing to compute the coresponding estimates (here in the streaming setting)

# Performance evaluation metrics

- Traditional accuracy measures can be used:
  - Precision and recall
  - Weighted average
  - Mean absolute sclaed errors
- Use baseline approaches for some settings:
- Evaluating change detection methods:

# Performance evaluation metrics

- Traditional accuracy measures can be used:
- Use baseline approaches for some settings:




- Evaluating change detection methods:

# Performance evaluation metrics

- Traditional accuracy measures can be used:
- Use baseline approaches for some settings:
- Evaluating change detection methods:
  - Probability of true change detection
  - Probability of false alarms
  - Delay of detection

# Experimental Design

- Cross-validation is not directly applicable
- Taking snapshots at different times
- Other techniques needed

# Evaluation of Time-Ordered Data

- **Holdout:** keeping a subset
- **Interleaved Test-Then-Train:** check every instance first
- **Controlled Permutations:** run multiple test with permutated copies of the data
  - (averaging accuracy might mask adaptation properties)

# Agenda

# Conclusion

- Concept drift is present in different application domains!

# Conclusion

- Concept drift is present in different application domains!
- Concept drift is not only studied in machine learning and data mining or pattern recognition (e.g. processmining).

# Conclusion

- Concept drift is present in different application domains!
- Concept drift is not only studied in machine learning and data mining or pattern recognition (e.g. processmining).
- Next challenges:
  - Scalability
  - Robustness & reliability
  - Moving from *black-box* adaptation to a more interpretable interpretation.
  - Reducing dependence on timely and acurate feedback

# Thank you for your attention

Questions?

# References I


João Gama et al. “A survey on concept drift adaptation”. In: *ACM Computing Surveys (CSUR)* 46.4 (2014), p. 44.

Blue Ridge Global. *Demand Forecasting*. URL: http://blueridgeglobal.com/wp-content/uploads/2012/09/demand-forecasting.png.

Harvard. *Data and Cloud computing*. URL: http://gking.harvard.edu/files/gking/files/dataandcloudcomputing.jpg?m=1430186249.

Ralf Klinkenberg and Thorsten Joachims. “Detecting Concept Drift with Support Vector Machines”. In: *In Proceedings of the Seventeenth International Conference on Machine Learning (ICML*. Morgan Kaufmann, 2000, pp. 487–494.

# References II

MIT. *Cheetah*. URL: http://cdn.phys.org/newman/gfx/news/hires/2014/1-mitengineers.jpg.

Netflix. *Netflix logo*. URL: http://lifehacking.nl/wp-content/uploads/netflix-logo.png.