Hierarchical models for the rainfall forecast

DATA MINING APPROACH

Thanh-Nghi Do

dtnghi@cit.ctu.edu.vn

June - 2014

Introduction

Problem

large scale → GCM

small scale → models

Aim

- Statistical downscaling local daily precipitation from large scale weather variables obtained by GCM

Introduction

Contributions

- Hierarchical models for the rainfall forecast
- Learning the classification model to categorise the day into the class: no rain, light rain, moderate rain, heavy rain and violent rain;
- Training the regression models to predict the rainfall from data of each class
- Experimental results on datasets collected from SEA - START show that the our hierarchical model is the most accurate rainfall prediction compared to linear regression, k neighbor neighbors, decision trees, support vector machines

Related works

Dynamic downscaling

- high-resolution regional climate models nested in a GCM to obtain local weather variables

Statistical downscaling

- Statistical relationship between large-scale GCM outputs and local weather variables (low cost)
- Weather typing method
- Stochastic weather generators
- Resampling
- **Regression**

Related works

Pasini, 2009

- Neural network: non-linear regression
- Non-linear relationship between large-scale variables (predictors of GCM) and regional/local variables (predictands)

Tripathi et al., 2006

- Support vector machine: non-linear regression
- Non-linear relationship between predictors & predictands

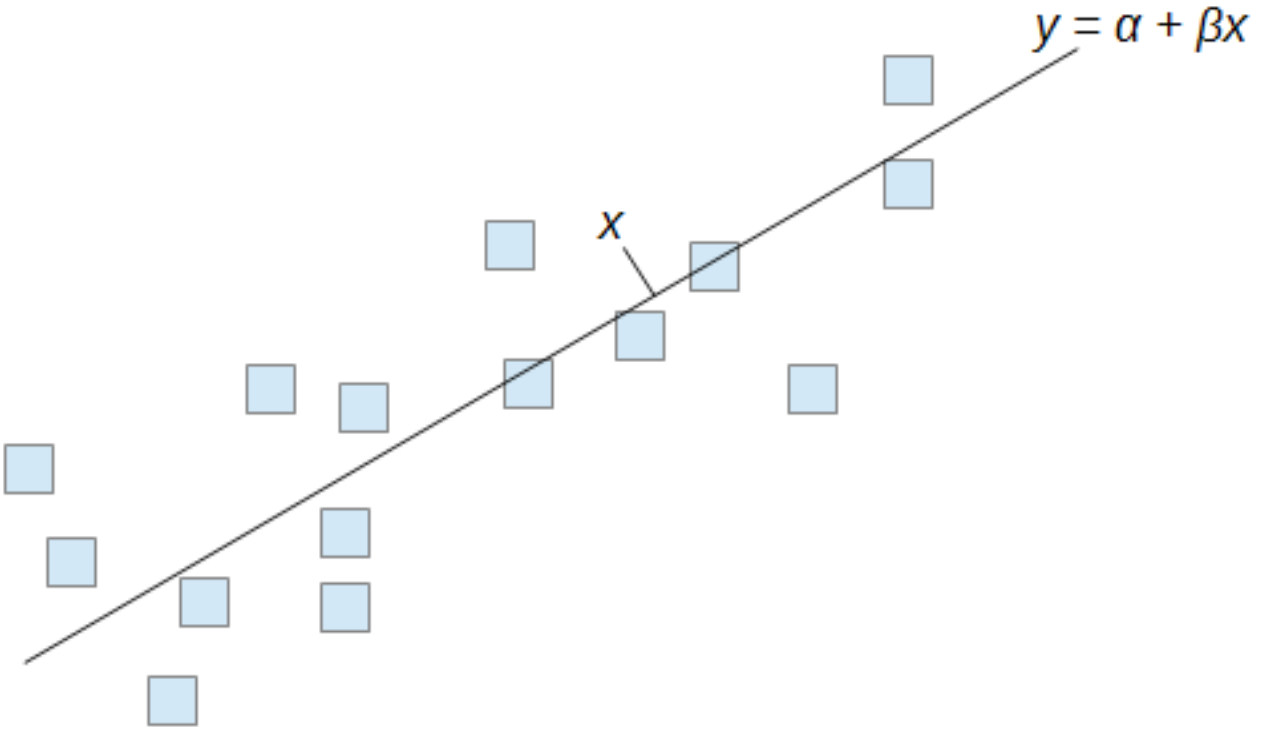
Chen et al., 2010

- Support vector machines: non-linear classification/regression
- Classification: dry/wet
- Regression: precipitation of wet days

Linear models

Linear regression

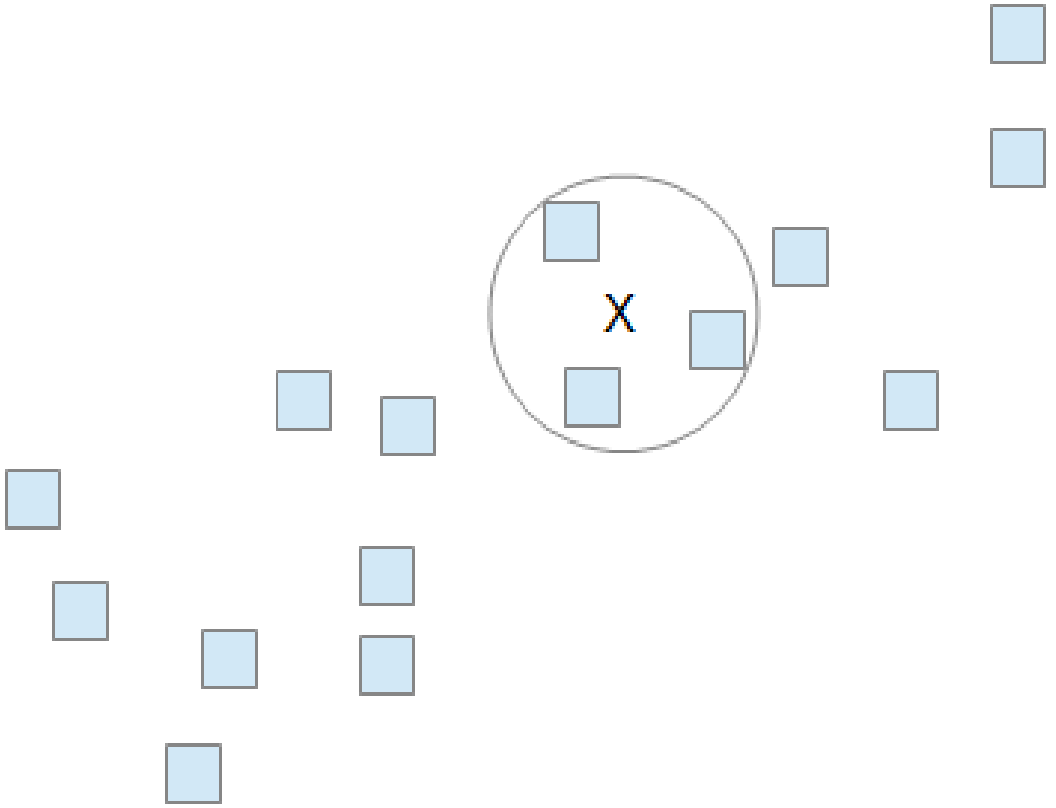
- Simplest model: regression
- Non-parametric
- Numerical data



Data mining tools

k nearest neighbors

- Simplest models: classification, regression
- Parameter k=1,2,...
- Error bound: 2 x Bayes risk
- Long time for prediction
- Numerical data

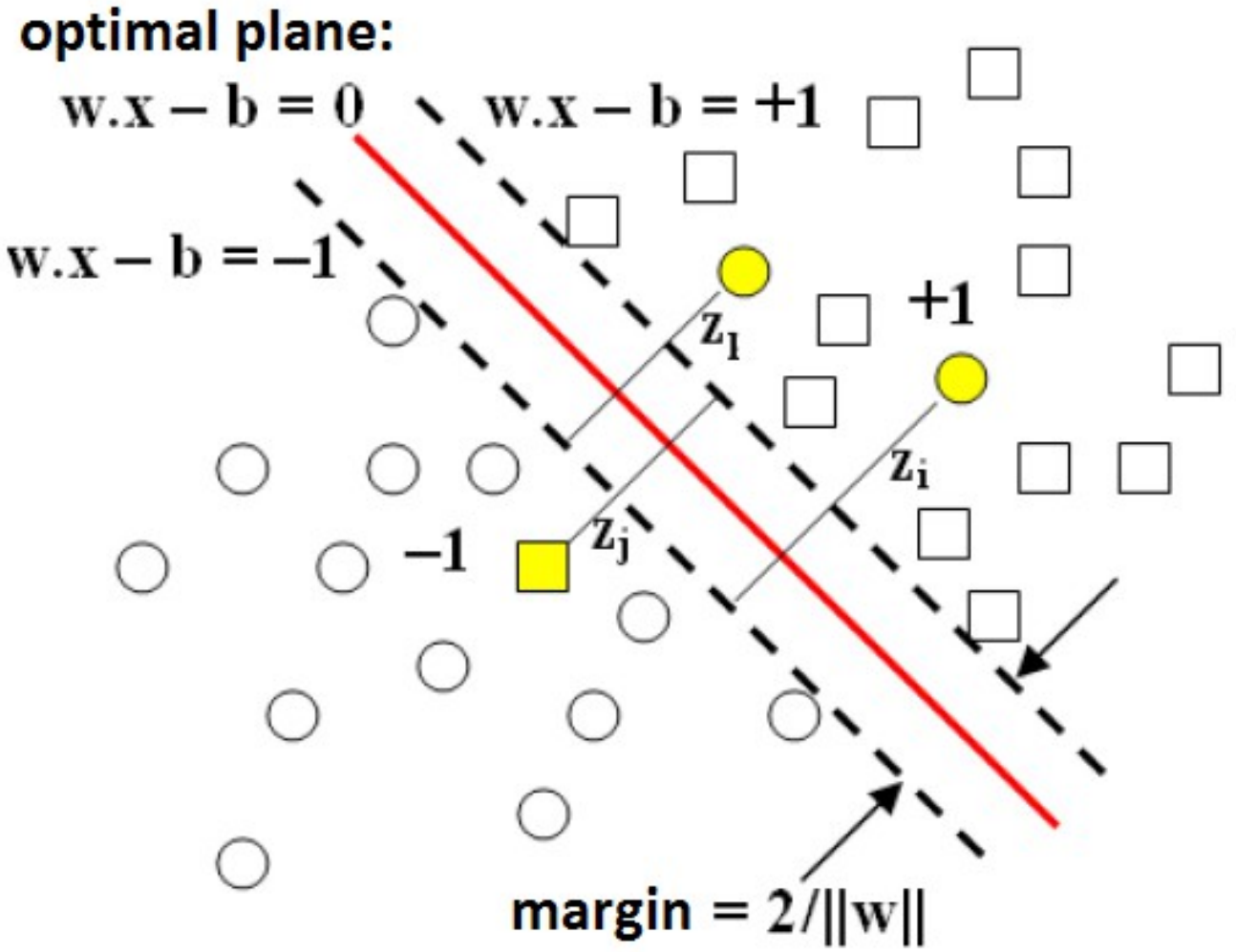


Data mining tools

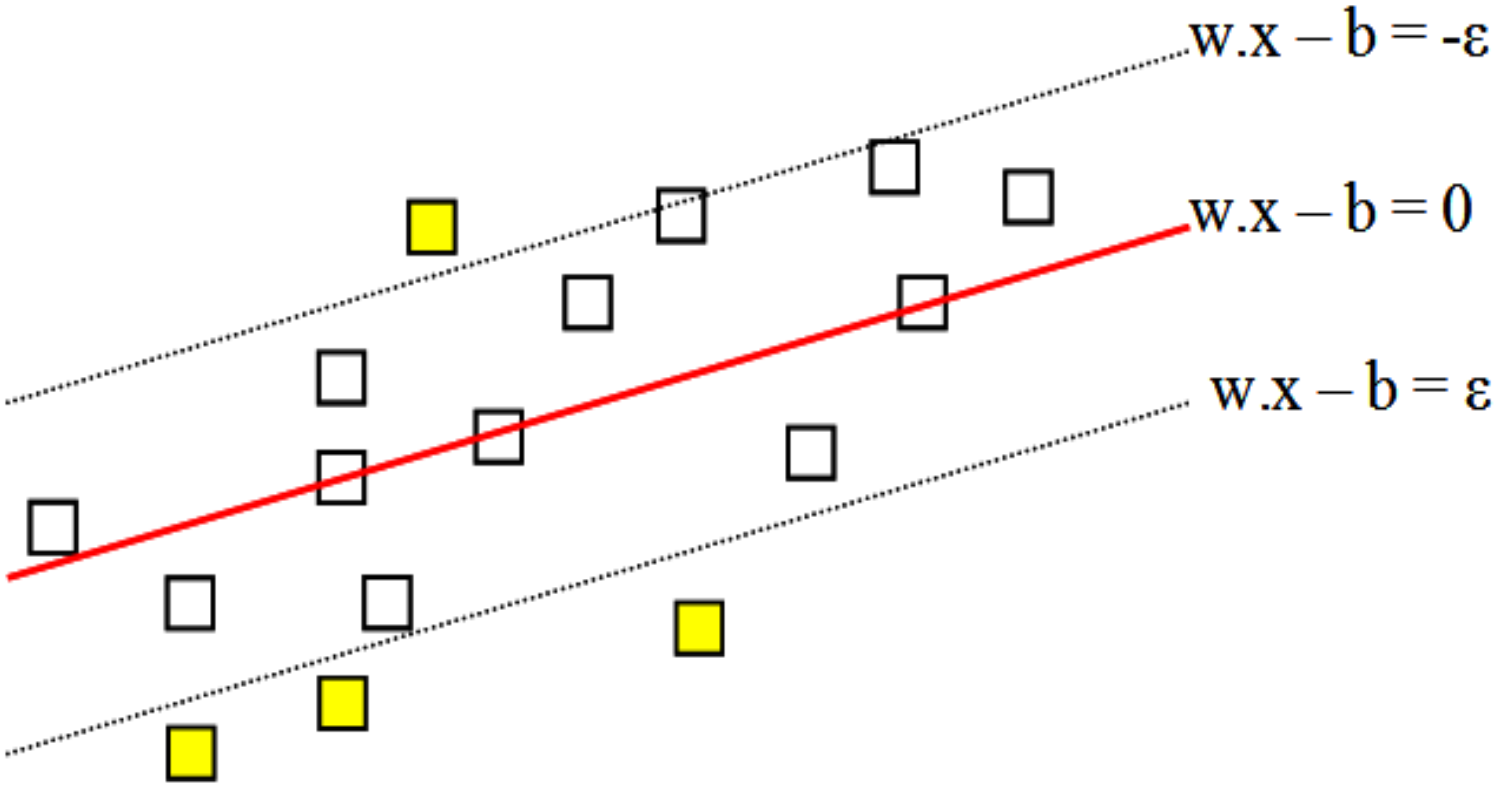
Support vector machine

- Accurate models: classification, regression
- SVM + kernel function = model
- Linear/non-linear tasks
- Global optimum
- Hyper-parameters
- Long training time
- Numerical data

Support vector classification



Support vector regression



Data mining tools

Decision tree

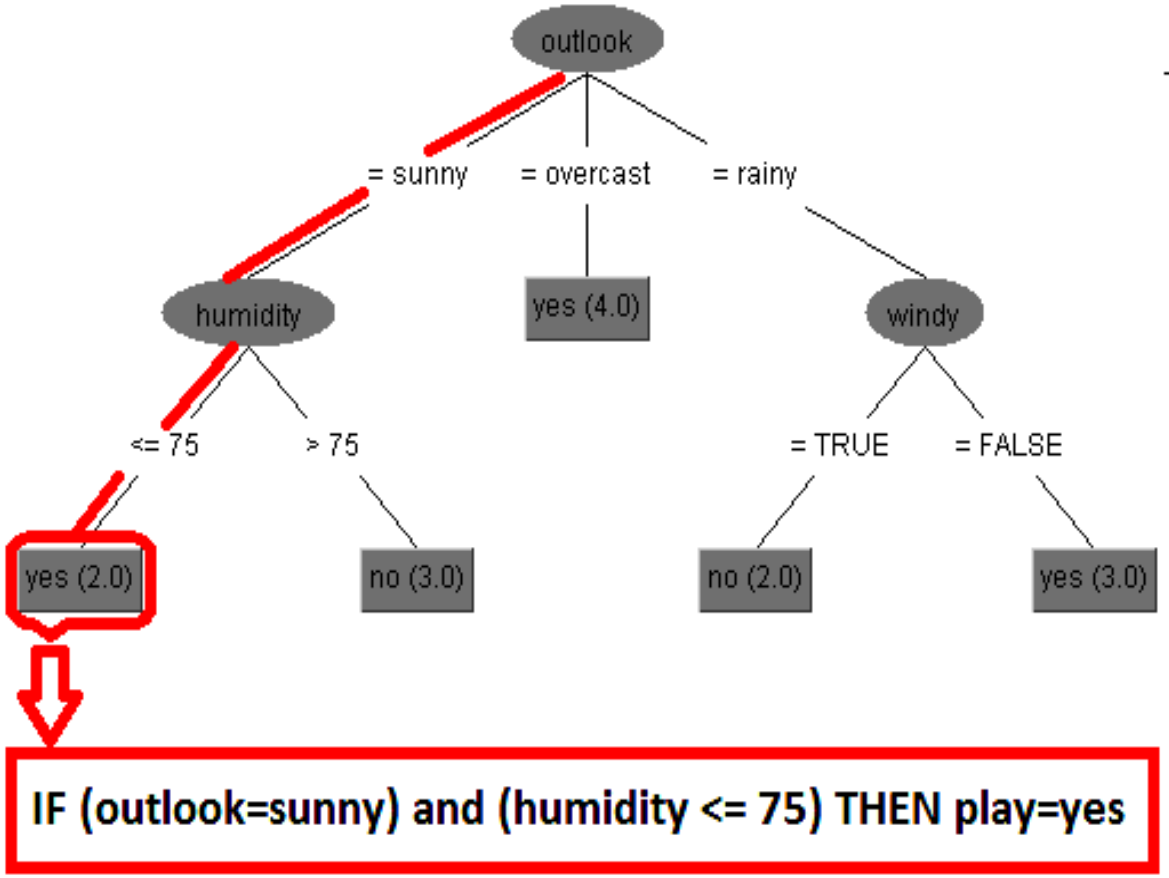
- Classification, regression
- Simple, popular, good performance, intuitive model
- Non-parametric
- Any type of data
- Less accurate than SVM

Bagging and random forest

- Ensemble of decision trees: more accurate than a single one
- Classification, regression
- Simple, popular, good performance
- Non-parametric
- Training time: faster than SVM

Decision tree

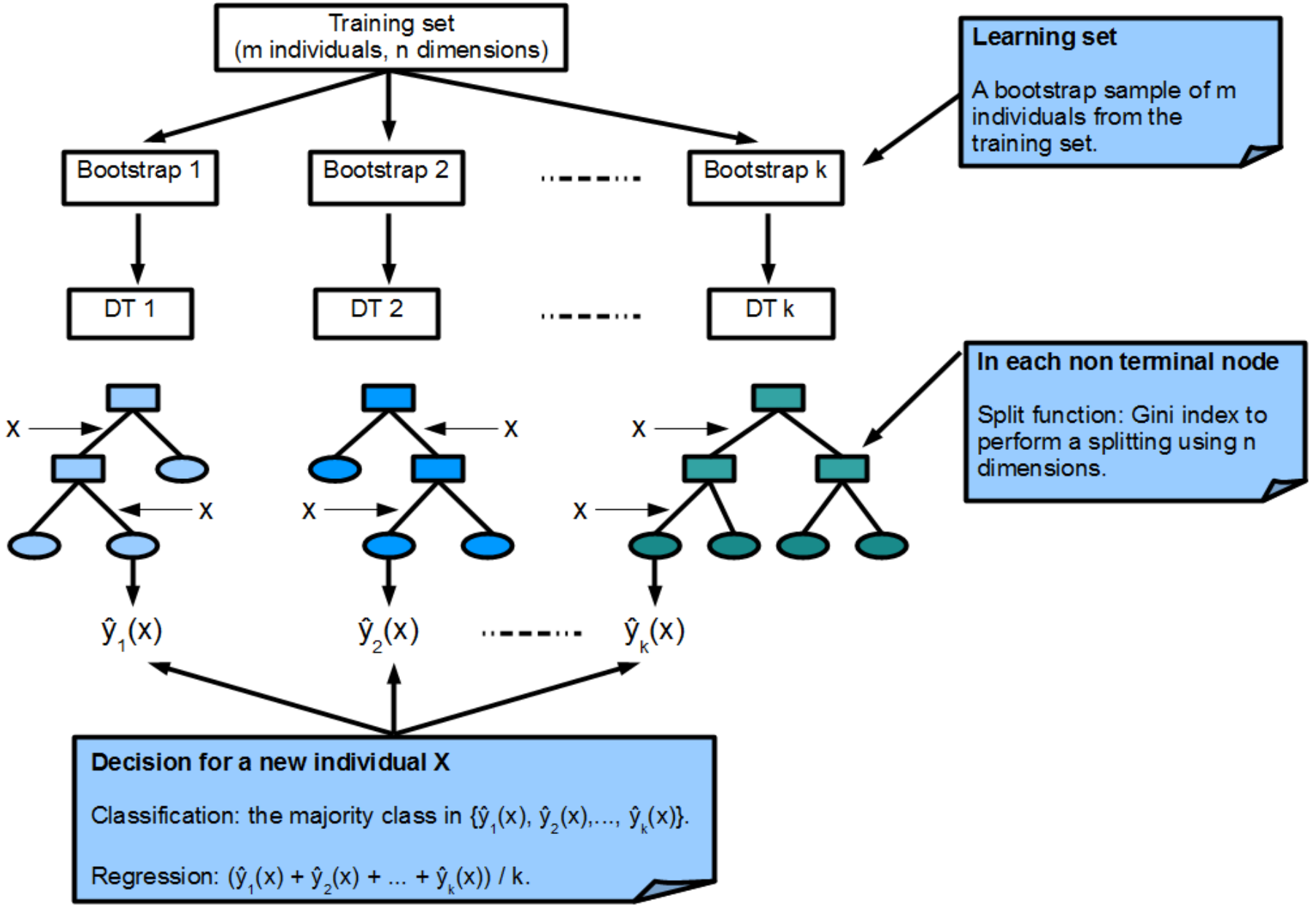
- Classification: minimize impurity (Shannon entropy, Gini index)
- Regression: minimize variance



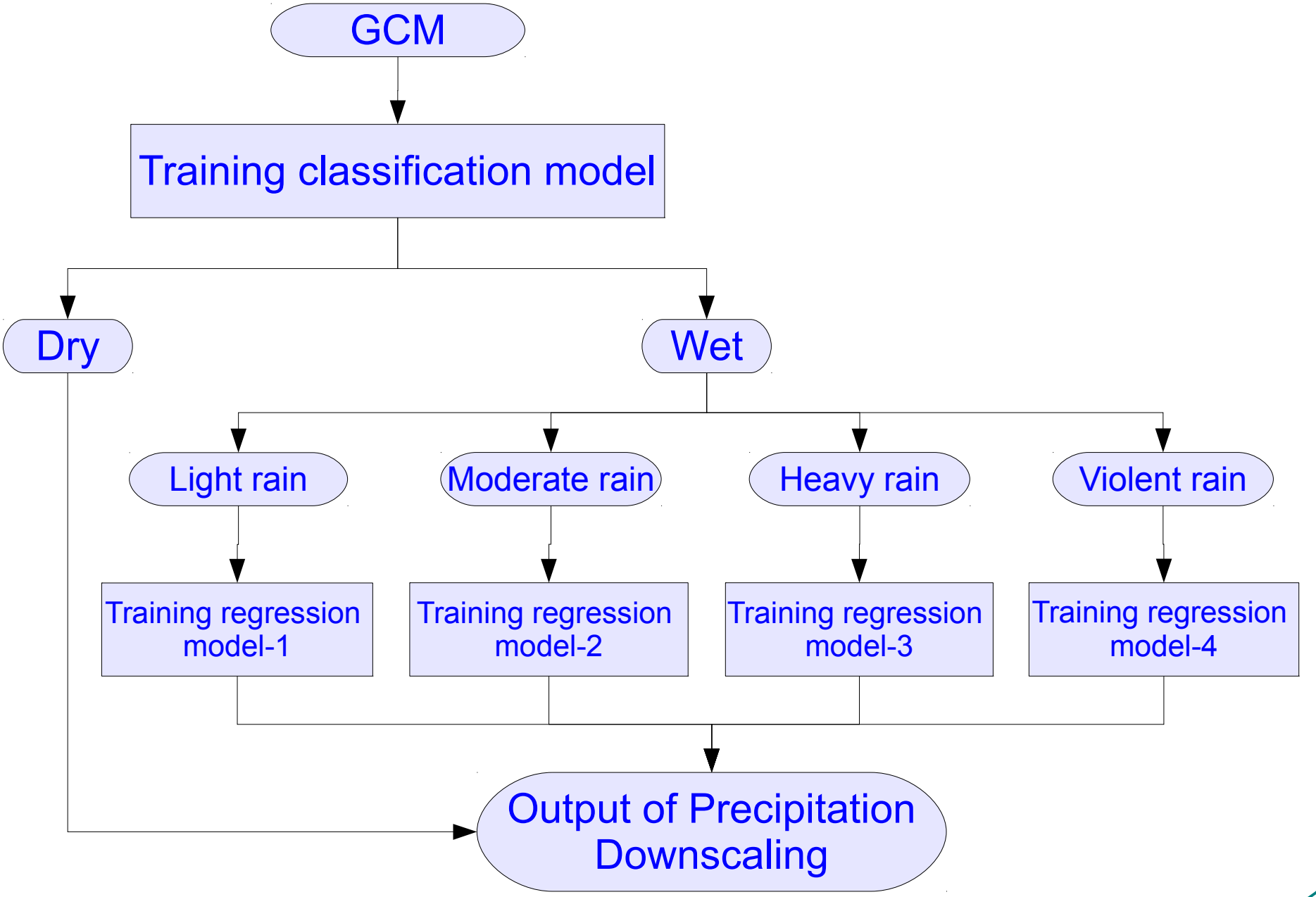
outlook	temp.	hum.	windy	Play
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
...	...	...	...	...
...	...	...	...	...

IF (outlook=sunny) and (humidity <= 75) THEN play=yes

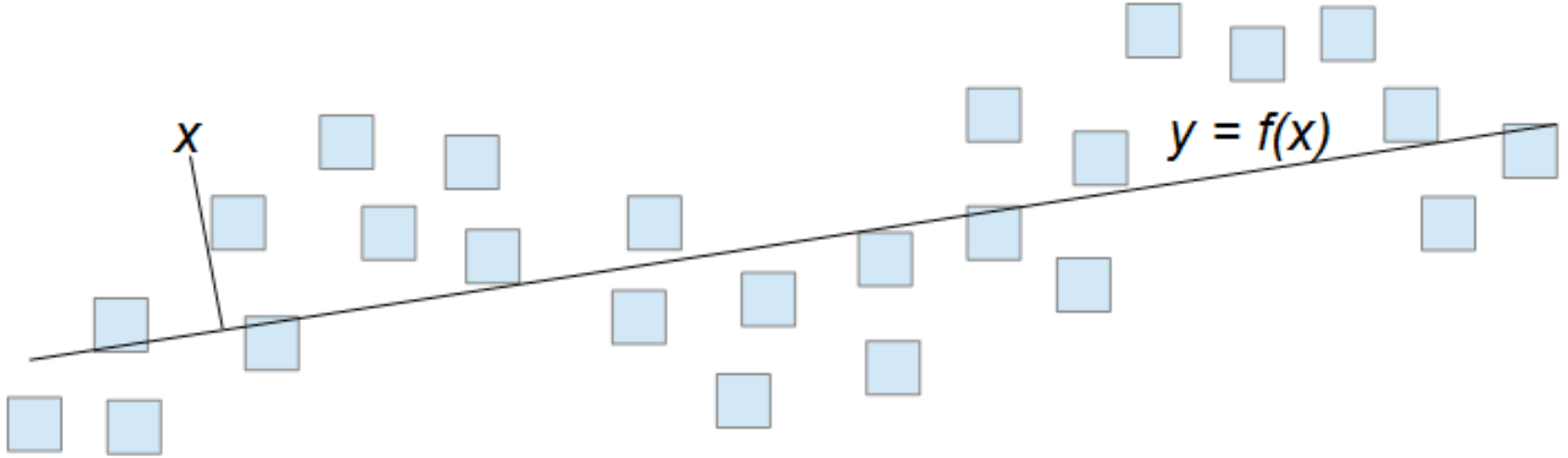
Bagging

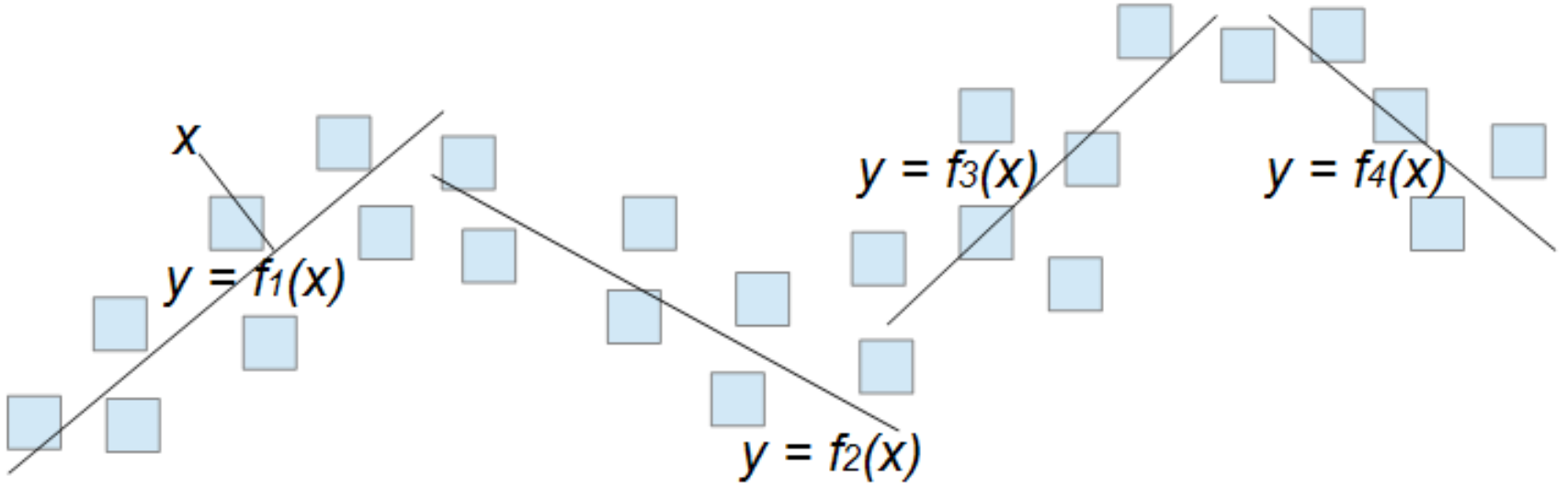

Training set
(m individuals, n dimensions)
Bootstrap 1
Bootstrap 2
Bootstrap k
DT 1
DT 2
DT k
x
x
x
x
x
x
$\hat{y}_1(x)$
$\hat{y}_2(x)$
$\hat{y}_k(x)$
Learning set
A bootstrap sample of m individuals from the training set.
In each non terminal node
Split function: Gini index to perform a splitting using n dimensions.
Decision for a new individual X
Classification: the majority class in $\{\hat{y}_1(x), \hat{y}_2(x), \ldots, \hat{y}_k(x)\}$.
Regression: $(\hat{y}_1(x) + \hat{y}_2(x) + \ldots + \hat{y}_k(x)) / k$.

Hierarchical models



1 model vs. 4 models





Data collection

Dataset (predictors)

- Website: http://cc.start.or.th/
- 1980-2006: 26 years
- Can Tho: LON = 105.8 and LAT = 10.2
- Tmax (°C)
- Tmin (°C)
- Precipitation (mm)
- Wind speed (m/s)
- Wind direction (degree from north)
- Solar Radiation (W/m^2)

Software programs

Algorithms

- Implemented in **R** language
- Linear regression (LM)
- k nearest neighbors (kNN)
- Decision trees (DT)
- Bagging of decision trees (BagR)
- Support vector machines (SVM) using a RBF kernel
- Hierarchical approach: BagC-BagR, SVC-SVR

Evaluation

Measures

- Mean squared error (MSE)
- Mean absolute error (MAE)

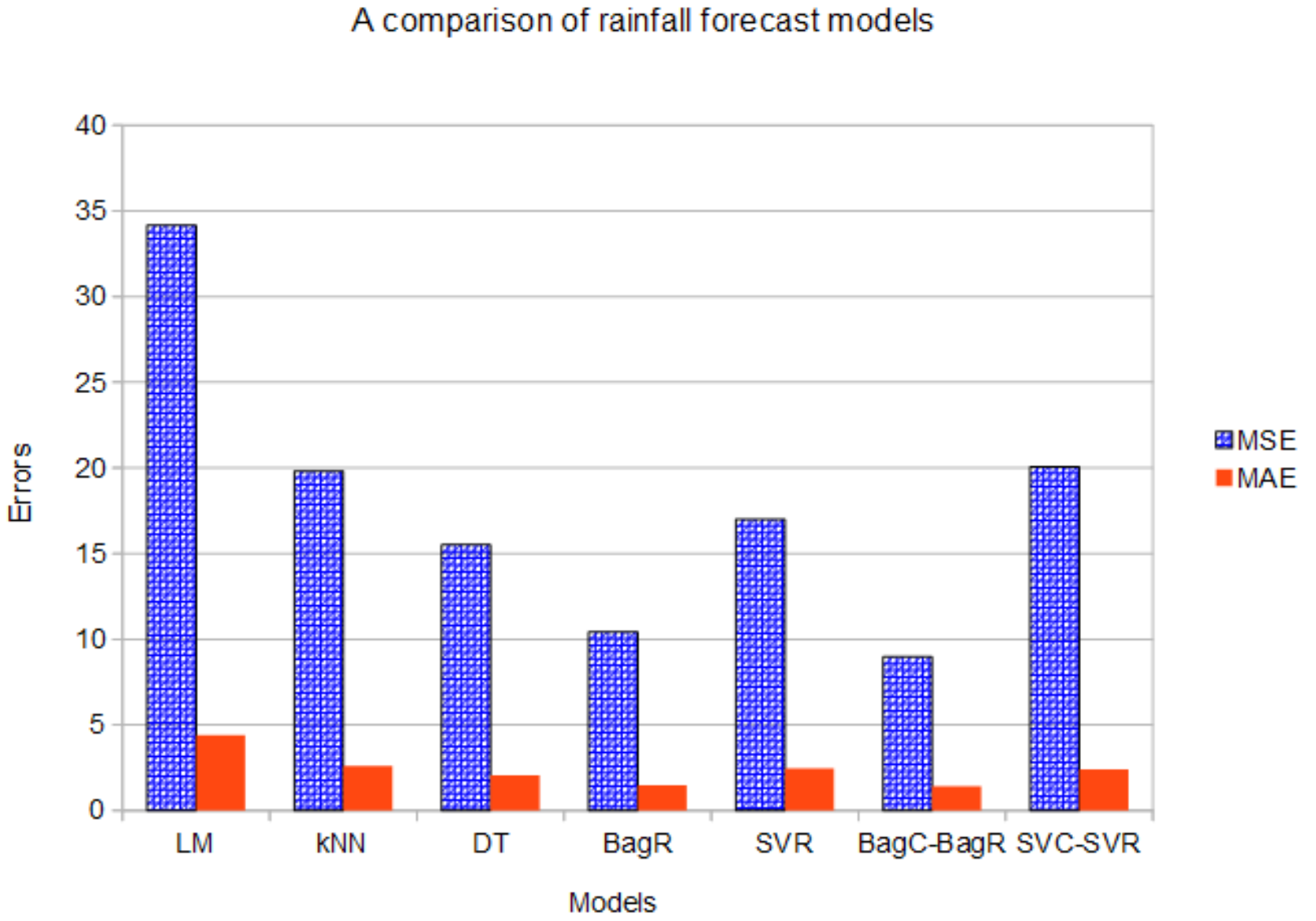
Hold-out protocol

- Repeat 3 times
- Randomly splitting the dataset into the training set (2/3 fullset) and testing set (1/3 fullset)
- Learning prediction models from the training set
- Evaluating the prediction models using the testing set
- Averaging the prediction results

Prediction results

Methods	MSE	MAE
Linear regression (LM)	34.1613	4.4062
k nearest neighbors (kNN with *k=5*)	19.8154	2.5938
Decision tree regression (DT with *minobj=5*)	***15.5085***	***2.0525***
Bagging (Bag with *#trees=100*)	10.4347	1.4814
Support vector regression (SVR-RBF, $\gamma=0.01$, $\varepsilon=0.1$, $C=10^4$)	17.0064	2.4554
Hierarchical model (BagC-BagR with *#trees=100*)	**8.9708**	**1.4337**
Hierarchical model (SVC-SVR-RBF)	20.0745	2.4058

Prediction results

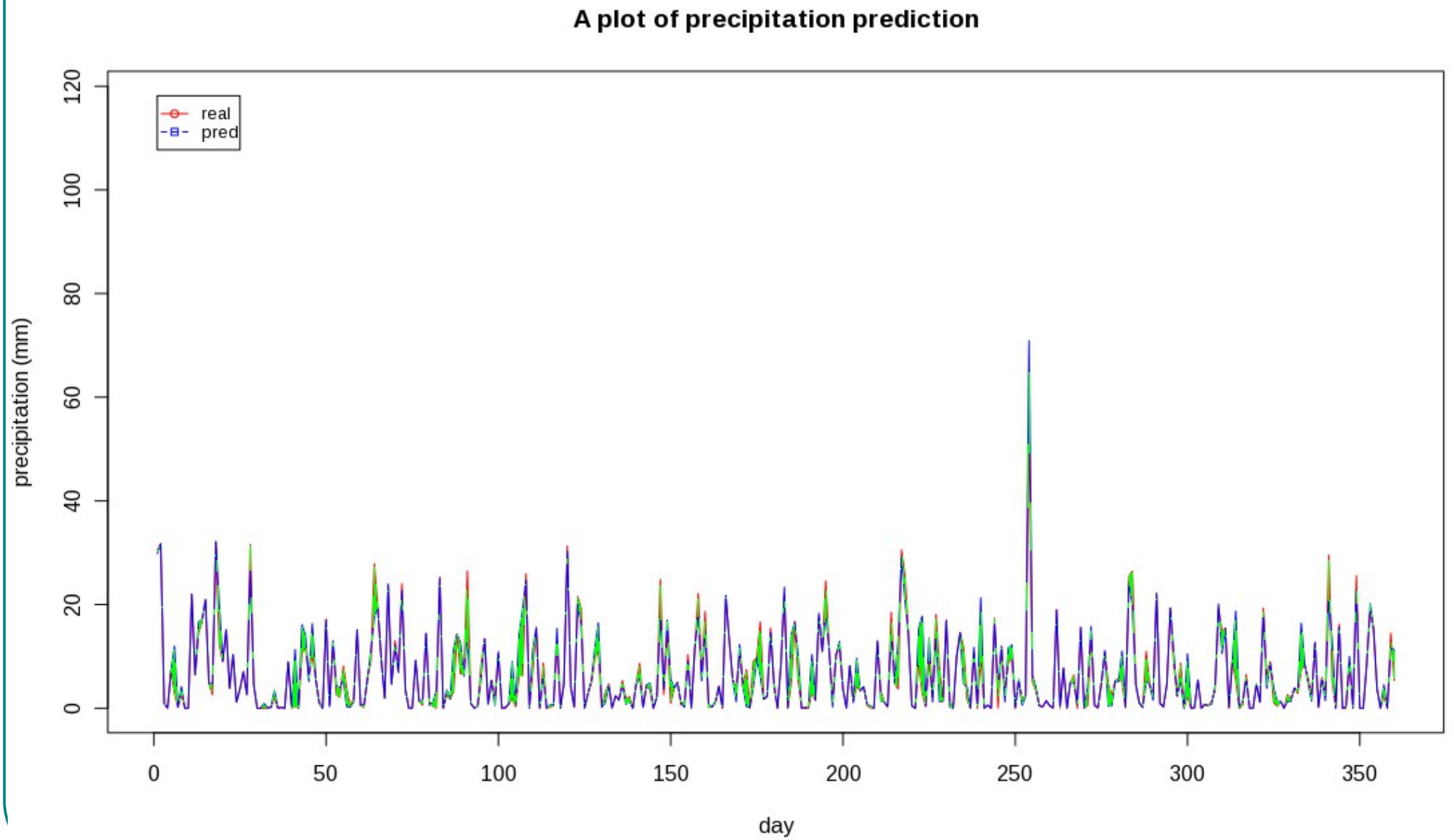


Prediction results

Comments

- Linear regression: not suited for rainfall forecast
- Simple models (kNN, Decision trees): very competitive
- Non-linear models (SVM, Bagging) and hierarchical approach (BagC-BagR, SVC-SVR): most accurate
- Bagging, BagC-BagR: non-parametric, more accurate than SVM

Prediction of BagC-BagR



Future works

To do

- Preprocessing: smoothing and cleaning data
- Improving prediction results of hierarchical models
- Hierarchical models: non-linear regression models for other problems, including water level, number of telephone calls, etc.

Thanks for your attention!

Question?