Machine Learning —

Application in Climate Science

Is there a unique definition of ML?

- Wikipedia —

“Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction”

- Example:

Amazon uses ML to suggest items based on the purchase histories of other customers.

Another Example: Netflix

Sparse matrix completion

- Make predictions about what movies you'll like
- based on the ratings of other users

- <u>Data is short in supply</u> as one user has rated only a few movies

- So, you put together this matrix of
  - All users that have ever rated things on Netflix and
  - All the movies

- Hope to fill in some missing data

Simply put ML is about:

- Constructing computer programs that automatically improve with experience.
- ML employs techniques from the fields of
  - computer science,
  - statistics, and
  - artificial intelligence, among others.

My Background

- Economics and Management Science
  - Statistics for Business Application
  - Regression testing, Business Forecasting, Econometrics

- Data Science
  - Statistical Modeling — MLE, Bayesian Inference, GLM…
  - Machine Learning —
    - Decision Trees,
    - Random Forests,
    - Clustering: K-means, Spectral,
    - Boosting: AdaBoost

Climate Informatics → (Climate Sc. + Data Sc)

- Dr. Claire Monteleoni, Assistant Professor of Computer Science at George Washington University

- She co-founded the Climate Informatics Workshop with NASA climatologist, Gavin Schmidt.

- Her most recent work — applying ML to track several climate models that make predictions about climate change where the data collected is used to adjust ***how each of the models' output is weighed***

One Dilemma in Climate Science

- “to be able to say that we are having warming, we have to be able to say what the temperature was in the past.”

- Records of past are few and far between

- ML can also be used to better understand what the climate looked like in past.

Main types of climate data

- Past: *Historical data*
  - Limited
  - Very heterogeneous

- Present: *Observation data*
  - Increasingly measured
  - Large quantities

- Past, Present, Future: *Climate model simulations*
  - Vast, high-dimensional
  - Encodes scientific domain knowledge
  - Information lost in discretizations
  - Future predictions cannot be validated

Improving predictions of ensembles

- No one model predicts best all the time, for all variables
- Average predictions over all models is better predictor

- ML Approaches:
    - Tracking Climate Models (TCM)
    - Neighborhood-Augmented TCM (geospatial influence)
    - Multi-model Regression with spatial smoothing
    - Climate prediction via Matrix Completion

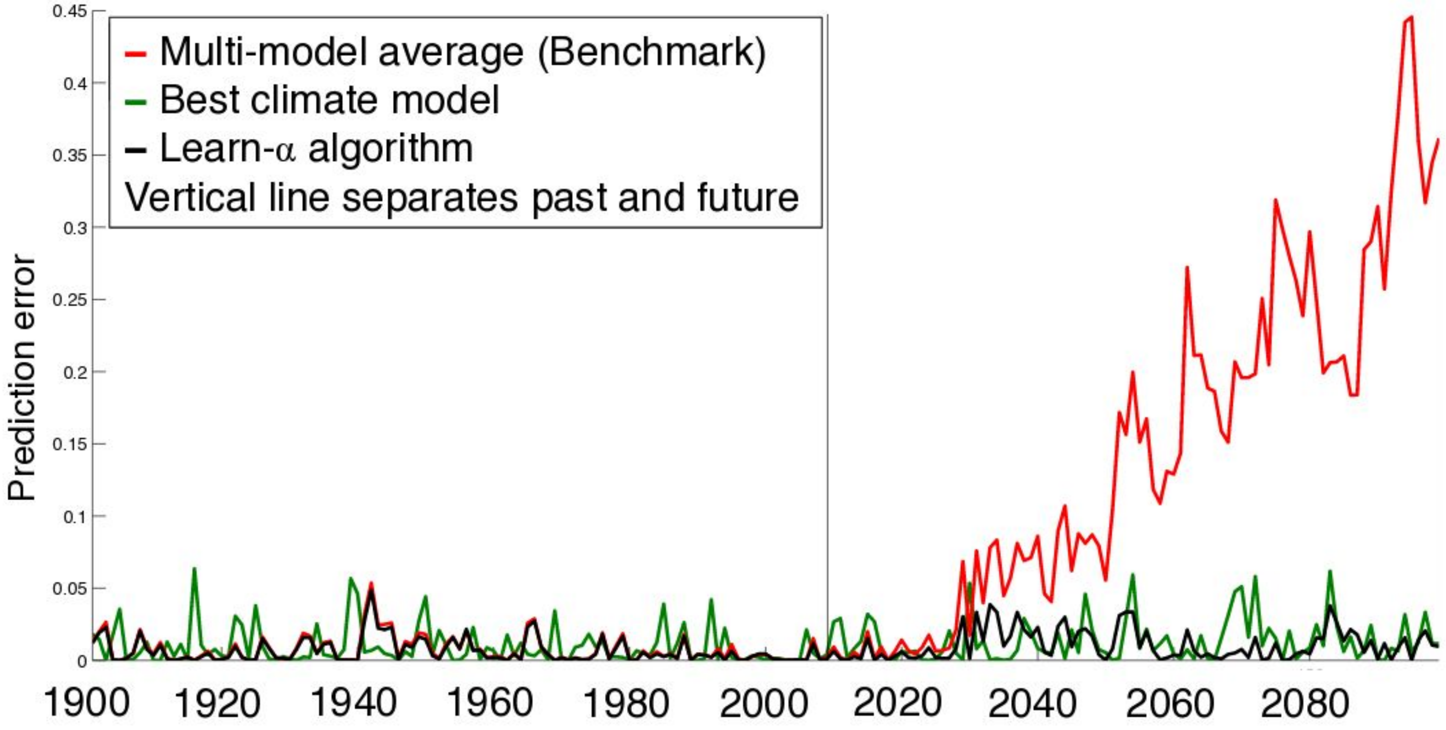
Adaptive weighted prediction

- Average prediction weights all models equally

- Weighted average prediction gives varying weights to each models based on past performances

- Adaptive weighted average prediction identifies current best predicting model vs one that quickly switching to other models
  - Tradeoff: how often the identity of the best model switches

- Online Learning: Non stationary data
  - Learns the switching rate: level of non-stationarity



M, Schmidt, Saroha & Asplund, SAM 2011
(CIDU 2010)

Learning curves

Track a set of expert predictors under changing observations

ML and Data mining collaborations with CS

- **Atmospheric Chemistry**, e.g. Musicant et al. ‘07 (‘05)
- **Meteorology,** e.g. Fox-Rabinovitz et al. ‘06
- **Seismology**, Kohler et al. ‘08
- **Oceanography**, e.g Lima et al. ‘09
- **Mining/modeling Climate Data**, e.g. Steinback et al. ‘03, Steinhaeuser et al. ‘10, Kumar ‘10

ML and Climate Modeling

- **Data-drive climate models**, Lozano et al. ‘09
- **ML techniques inside a climate model, or for calibration**, e.g. braverman et al. ‘06
- **ML techniques with ensembles of climate models**:
    - Regional models: Sain et al. ‘10
    - Global Climate Models (GCM): TCM

ML and Air Quality

- Inferring Air Quality for Station Location Recommendation Based on Urban Big Data
  - Infer real-time air quality of any arbitrary location given environmental data and data from very sparse monitoring locations.
  - Determine the best locations to establish new monitor stations to improve the inference quality

  - Design a semi-supervised inference model
    - utilizing existing monitoring data
    - together with heterogeneous city dynamics, including meteorology, human mobility, structure of road networks, and point of interests
  - Propose an entropy-minimization model to suggest the best locations to establish new monitoring stations.

  - Evaluate the proposed approach using Beijing air quality data, resulting in clear advantages over a series of state-of-the-art and commonly used methods.

ML and Air Quality

- U-Air: When Urban Air Quality Inference Meets Big Data
  - Infer the real-time and fine-grained air quality information throughout a city
  - Air quality data reported by existing monitor stations and other data sources such as meteorology, traffic flow, human mobility, structure of road networks, and point of interests
  - Propose a semi-supervised learning approach that consists of two separated classifiers
    - A spatial classifier based on an artificial neural network (ANN) — takes spatially-related features (e.g., the density of POIs and length of highways) as input to model the spatial correlation between air qualities of different locations.
    - A temporal classifier based on a linear-chain conditional random field (CRF), involving temporally-related features (e.g., traffic and meteorology) to model the temporal dependency of air quality in a location.

ML and Air Quality

- Deriving high-resolution urban air pollution maps using mobile sensor nodes
  - Real-time pollution assessment
  - Analyze one of the largest spatially resolved ultrafine particles (UFP) data set containing over 50 million measurements
  - More than two years using mobile sensor nodes installed on top of public transport vehicles in the city of Zurich, Switzerland.
  - Develop land-use regression models to create pollution maps with a high spatial resolution of 100 m × 100 m.
  - Compare the accuracy of the derived models across various time scales and observe a rapid drop in accuracy for maps with sub-weekly temporal resolution.
  - Propose a novel modeling approach that incorporates past measurements annotated with metadata into the modeling process.
  - Achieve a 26% reduction in the RMSE – a standard metric to evaluate the accuracy of air quality models– of pollution maps with semi-daily temporal resolution.