Machine Learning

A Scientific Method or Just a Bag of Tools?

Don Hush

Machine Learning Team

Group CCS-3, Los Alamos National Laboratory

Machine Learning Toolbox

- Fisher’s Linear Discriminant
- Nearest Neighbor
- Neural Networks (backprop)
- Decision Trees (CART, C4.5)
- Boosting
- Support Vector Machines
- K–Means Clustering
- Principle Component Analysis (PCA)
- Expectation-Maximization (EM)
- ... and many more

A Day at Work with the ML Toolbox

- Job Assignment: Design a system that uses the Tufts Artificial Nose to detect trichloroethylene (TCE).
- Tufts Data Collection:
  - 760 samples with TCE
  - 352 samples without TCE
- Tool: Support Vector Machine (SVM)

A Day at Work ...

- The SVM Tool:
  - produces a classifier and
  - reports a classification error rate of 18%

A Day at Work ...

- The SVM Tool:
  - produces a classifier and
  - reports a classification error rate of 18%
  - *However, when the classifier is deployed it produces an error rate of 38%*

A Day at Work ...

- The SVM Tool:
  - produces a classifier and
  - reports a classification error rate of 18%
  - *However, when the classifier is deployed it produces an error rate of 38%*
- One of the (many) reasons why this is unacceptable: The naive classifier, *that predicts NOT-TCE for every sample*, produces an error rate of 10%.

A Day at Work ...

- The SVM Tool:
  - produces a classifier and
  - reports a classification error rate of 18%
  - *However, when the classifier is deployed it produces an error rate of 38%*
- One of the (many) reasons why this is unacceptable: The naive classifier, *that predicts NOT-TCE for every sample*, produces an error rate of 10%.
- What Went Wrong?

A Day at Work ...

- The SVM Tool:
  - produces a classifier and
  - reports a classification error rate of 18%
  - *However, when the classifier is deployed it produces an error rate of 38%*
- One of the (many) reasons why this is unacceptable: The naive classifier, *that predicts NOT-TCE for every sample*, produces an error rate of 10%.
- What Went Wrong?

READ THE MANUAL

The SVM Manual Entry

- SVMs ... *assume that the operating environment is characterized by a stationary random process and that the training data is sampled from that process ...*

The SVM Manual Entry

- SVMs ... *assume that the operating environment is characterized by a stationary random process and that the training data is sampled from that process* ...

- Since the fraction of TCE samples in the training data is 0.7 and the fraction on the operating environment is 0.1, *the TCE problem violates the assumptions!*

What Should We Do?

- tweak the SVM tool

What Should We Do?

- tweak the SVM tool
- use a different tool from the toolbox

What Should We Do?

- tweak the SVM tool
- use a different tool from the toolbox
- *design a tool specifically for the TCE problem*

How do we design a new tool?

Scientific Problem Formulation

Key Ingredients:

- Specify a performance criterion – a measure of the quality of the model

Scientific Problem Formulation

Key Ingredients:

- Specify a performance criterion – a measure of the quality of the model
- Identify and characterize the information available to design the model

Scientific Problem Formulation

Key Ingredients:

- Specify a performance criterion – a measure of the quality of the model
- Identify and characterize the information available to design the model ... two types
  - Empirical (EMP), i.e. data
  - First Principles Knowledge (FP)

Scientific Problem Formulation

Key Ingredients:

- Specify a performance criterion – a measure of the quality of the model
- Identify and characterize the information available to design the model
- Establish a validation procedure – a way to evaluate (or estimate) the performance of a proposed model

Scientific Problem Formulation

Key Ingredients:

- Specify a performance criterion – a measure of the quality of the model
- Identify and characterize the information available to design the model
- Establish a validation procedure – a way to evaluate (or estimate) the performance of a proposed model
  ... two methods
  - Empirical Tests
  - Theoretical Analysis

Scientific Problem Formulation

Key Ingredients:

- Specify a performance criterion – a measure of the quality of the model
- Identify and characterize the information available to design the model
- Establish a validation procedure – a way to evaluate (or estimate) the performance of a proposed model

All of this is done *before* we develop a solution method.

$ML^* = $ **ML + Scientific Method**

The ML^* Approach:

1. Construct a scientific problem formulation.

$ML^* =$ ML + Scientific Method

The ML^* Approach:

1. Construct a scientific problem formulation.
2. Determine its feasibility. If not feasible, go to step 1.

$ML^* =$ ML + Scientific Method

The ML^* Approach:

1. Construct a scientific problem formulation.
2. Determine its feasibility. If not feasible, go to step 1.
3. If feasible then use any means necessary to determine a solution method that is guaranteed to be
   - practical (e.g. computationally feasible), and
   - provide good performance (e.g. near optimal)

$ML^* =$ ML + Scientific Method

The ML^* Approach:

1. Construct a scientific problem formulation.
2. Determine its feasibility. If not feasible, go to step 1.
3. If feasible then use any means necessary to determine a solution method that is guaranteed to be
   - practical (e.g. computationally feasible), and
   - provide good performance (e.g. near optimal)

   (although obvious, *very few tools are designed to provide such guarantees!*)

An Example: Applying ML^* to the Supervised Classification Problem

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f(x)\right]$ to data point x.

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f(x)\right]$ to data point x.
- Performance Criterion: classification error rate, $e(f)$

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f(x)\right]$ to data point x.
- Performance Criterion: classification error rate, $e(f)$
- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (labeled) training data is sampled from that process

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f(x)\right]$ to data point x.
- Performance Criterion: classification error rate, $e(f)$
- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (labeled) training data is sampled from that process
- Validation:
  - Empirical: hold-out, cross-validation, bootstrap

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f(x)\right]$ to data point x.
- Performance Criterion: classification error rate, $e(f)$
- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (labeled) training data is sampled from that process
- Validation:
  - Empirical: hold-out, cross-validation, bootstrap *allows us to compare methods, but does not tell us how close we are to optimal*

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f(x)\right]$ to data point x.
- Performance Criterion: classification error rate, $e(f)$
- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (labeled) training data is sampled from that process
- Validation:
  - Empirical: hold-out, cross-validation, bootstrap
  - Theoretical: Statistics + Computer Science

ML^* + Supervised Classification

- A model f assigns the label $sign\left[f\left(x\right)\right]$ to data point x.
- Performance Criterion: classification error rate, $e(f)$
- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (labeled) training data is sampled from that process
- Validation:
  - Empirical: hold-out, cross-validation, bootstrap
  - Theoretical: Statistics + Computer Science *probably approximately correct (PAC) analysis*

Support Vector Machines (SVMs)

PAC Result: With mild assumptions on the distribution the SVM with n training samples requires

$$\boxed{O(n^2 \log n)}$$

computation to produce a classifier f_n with performance

$$\boxed{e(f_n) - e^* \leq cn^{-r} \quad (whp)}$$

where e^* is the theoretical minimum error and the rate $0 < r < 1$ depends on the distribution.

Support Vector Machines (SVMs)

PAC Result: With mild assumptions on the distribution the SVM with n training samples requires

$$\boxed{O(n^2 \log n)}$$

computation to produce a classifier f_n with performance

$$\boxed{e(f_n) - e^* \leq cn^{-r} \quad (whp)}$$

where e^* is the theoretical minimum error and the rate $0 < r < 1$ depends on the distribution.

Observation: This result addresses the major practical concerns:

- performance (of the actual classifier produced)
- computation (of the actual algorithm used)
- generality (applies to very large class of distributions)

Applying ML^* to the TCE Problem

ML^* + TCE Problem

Scientific Problem Formulation: Same as supervised classification except that

ML^* + TCE Problem

Scientific Problem Formulation: Same as supervised classification except that

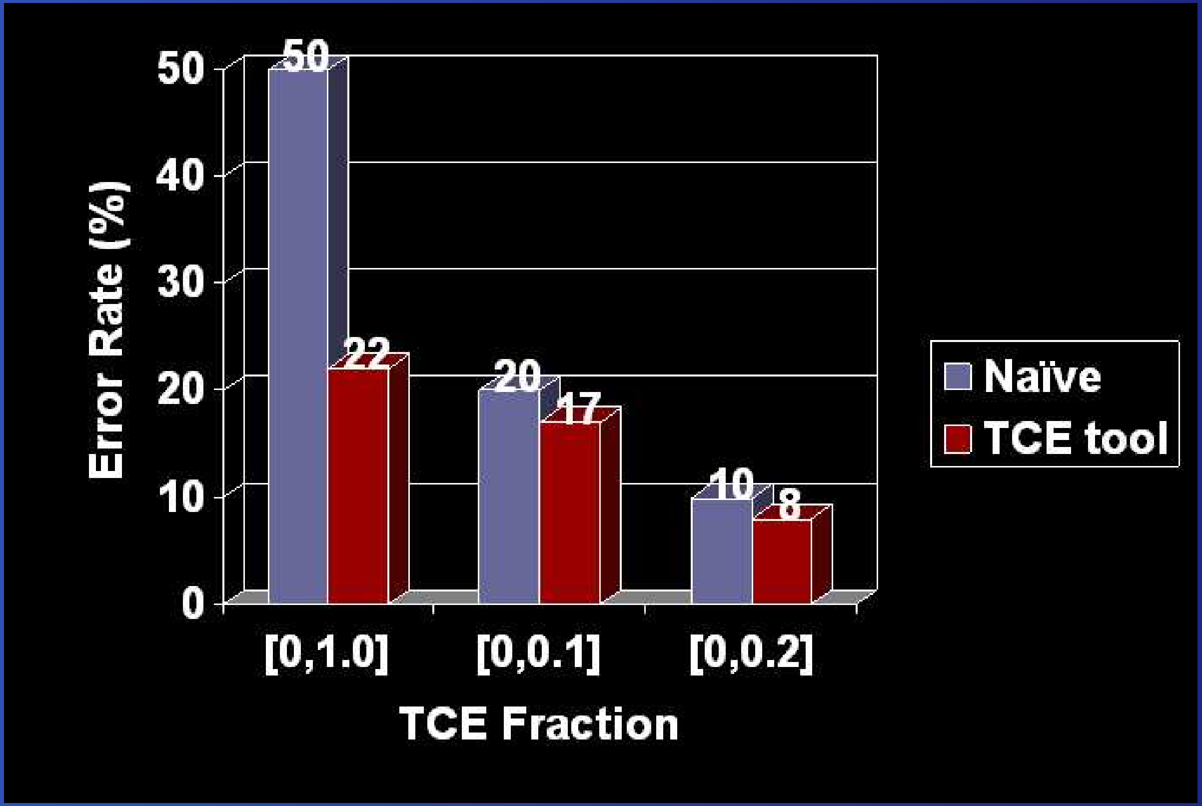
- we assume a nonstationary random process because the fraction of time that TCE is present varies over the range $[0, a]$.

ML^* + TCE Problem

Scientific Problem Formulation: Same as supervised classification except that

- we assume a nonstationary random process because the fraction of time that TCE is present varies over the range $[0, a]$.
- the performance criterion is the error rate for the worst possible value in the range $[0, a]$.
  (this is a min–max problem)

TCE Tool



Error Rate (%)
50
40
30
20
10
0
50
22
20
17
10
8
[0,1.0]
[0,0.1]
[0,0.2]
TCE Fraction
Naïve
TCE tool

Impacts of ML^* on Data Driven Modeling

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
  Direct Method: choose a model that minimizes an empirical risk function that is *calibrated* with respect to the performance criterion.

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
  Paradigm Shift?: replace *maximum likelihood, maximum entropy, least squares, plug–in, etc.* with *calibrated empirical risk minimization*

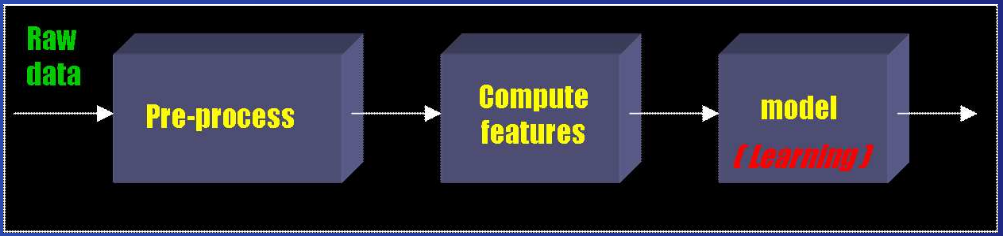
Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

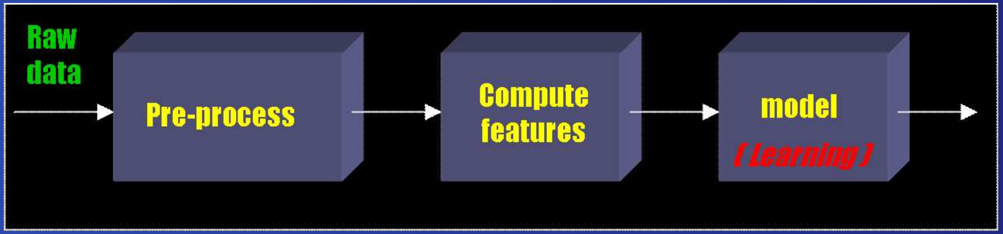
Traditional Approach:



Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

Traditional Approach:

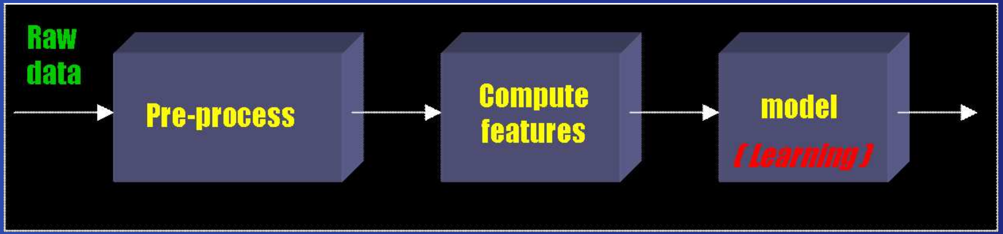


New Approach:

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

  Traditional Approach:

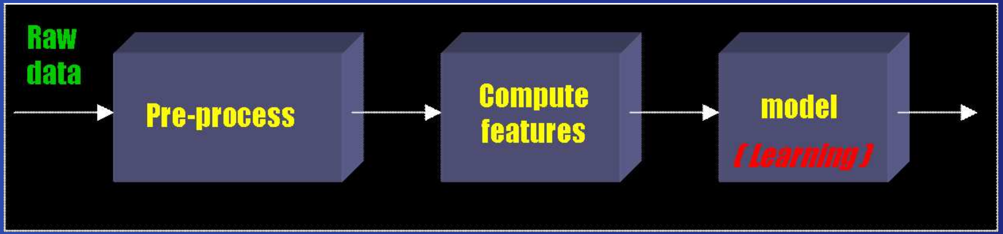


  New Approach:
  - push learning into earlier stages

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

Traditional Approach:



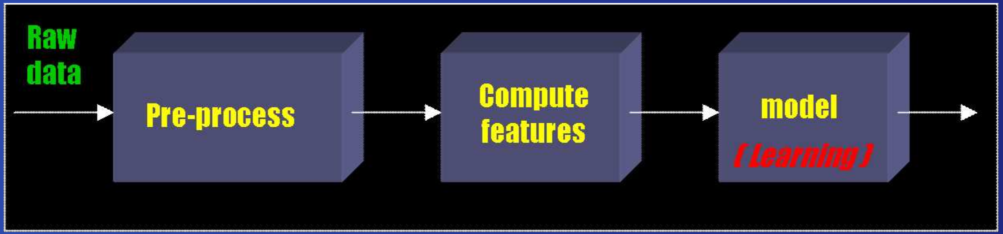
New Approach:

- push learning into earlier stages
- collapse last two stages into one using model classes that are

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

  Traditional Approach:



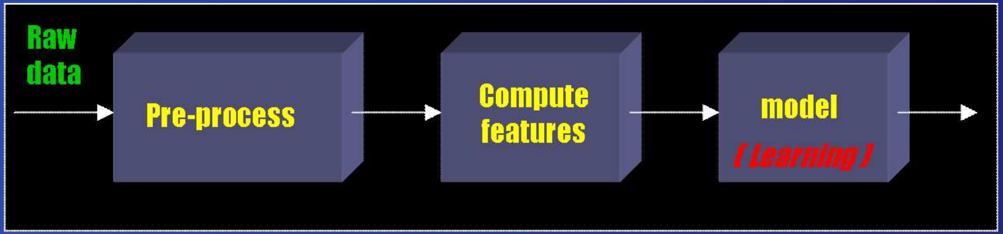
  New Approach:

  - push learning into earlier stages
  - collapse last two stages into one using model classes that are
    - richer (e.g. higher dimensions)

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

  Traditional Approach:



  New Approach:

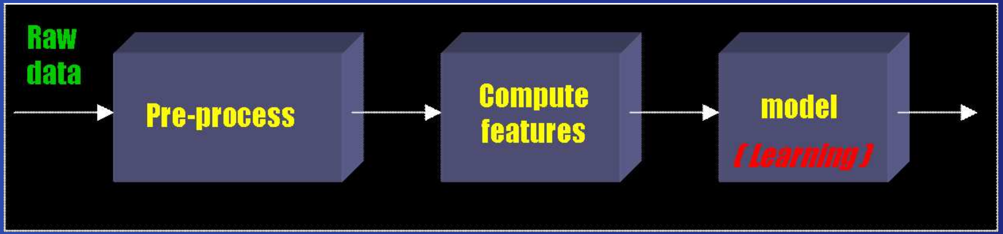
  - push learning into earlier stages
  - collapse last two stages into one using model classes that are
    - richer (e.g. higher dimensions)

      Myth: *dimensionality must be reduced to achieve good performance.* example

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

  Traditional Approach:



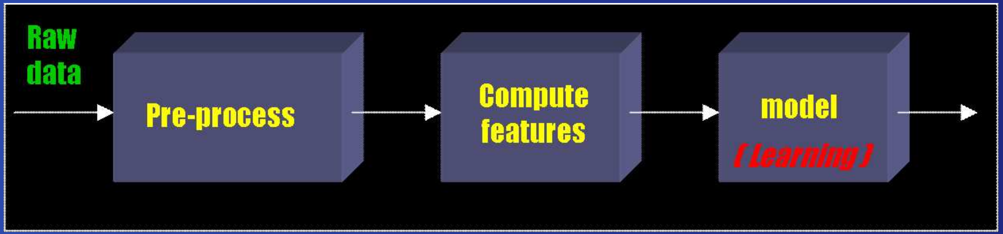
  New Approach:

  - push learning into earlier stages
  - collapse last two stages into one using model classes that are
    - richer (e.g. higher dimensions)
    - more flexible (e.g. accommodates different data types)

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

  Traditional Approach:



  New Approach:

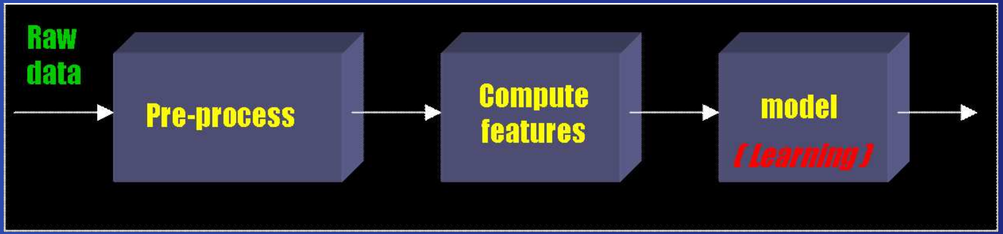
  - push learning into earlier stages
  - collapse last two stages into one using model classes that are
    - richer (e.g. higher dimensions)
    - more flexible (e.g. accommodates different data types)

    Myth: *data must be mapped to $\mathbb{R}^d$ before we can build a model.* example

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

  Traditional Approach:



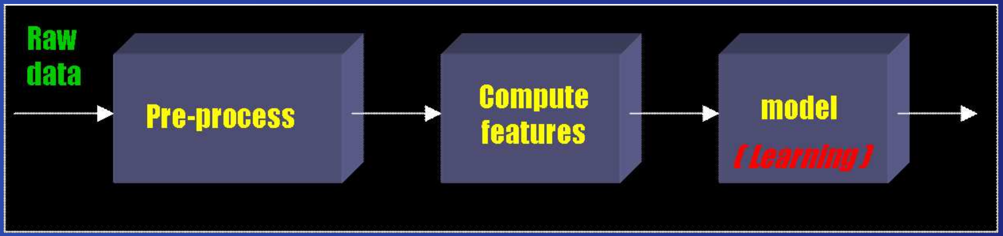
  New Approach:

  - push learning into earlier stages
  - collapse last two stages into one using model classes that are
    - richer (e.g. higher dimensions)
    - more flexible (e.g. accommodates different data types)
    - simply parameterized

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

Traditional Approach:



New Approach:

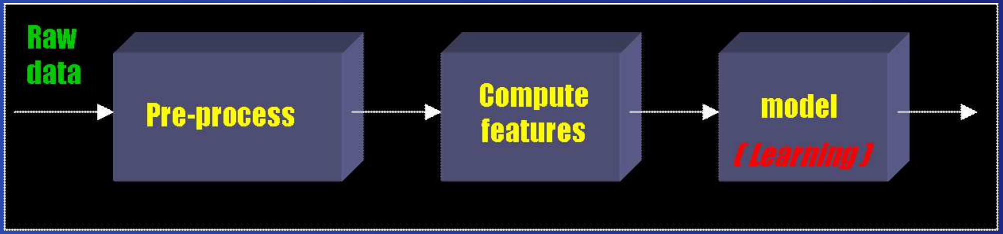
- push learning into earlier stages
- collapse last two stages into one using model classes that are
  - richer (e.g. higher dimensions)
  - more flexible (e.g. accommodates different data types)
  - simply parameterized

  Example: *Kernel Machines*

Impacts of ML^* on DDM

- It has focused attention on Direct Solution Methods
- It has started a movement towards end–to–end learning

Traditional Approach:



New Approach:

- push learning into earlier stages
- collapse last two stages into one using model classes that are
  - richer (e.g. higher dimensions)
  - more flexible (e.g. accommodates different data types)
  - simply parameterized

Paradigm Shift?: replace *feature design* with *kernel design*

Applying ML^* to Anomaly Detection

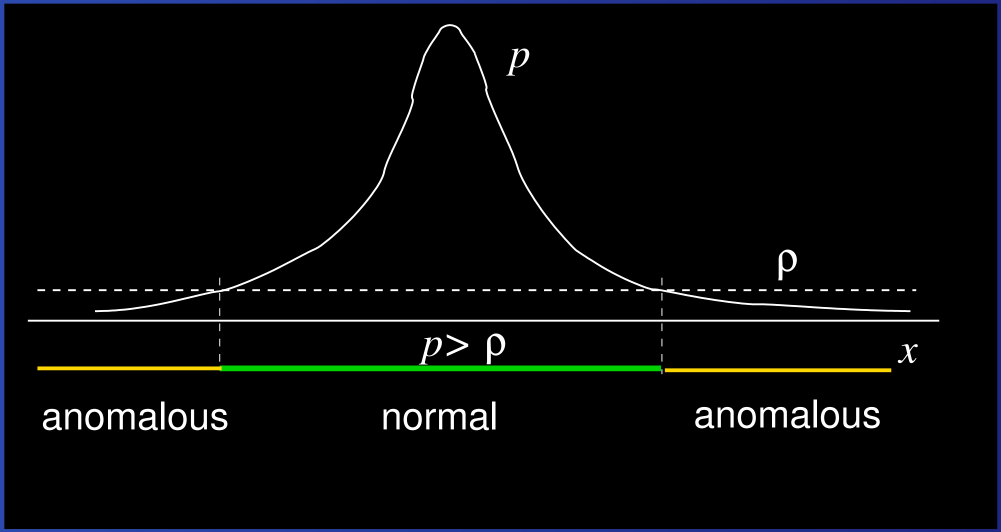
ML^* + Anomaly Detection

ML^* + Anomaly Detection

- Definition of Anomaly: x is anomalous if its density values falls below a threshold, i.e. $p(x) \leq \rho$.

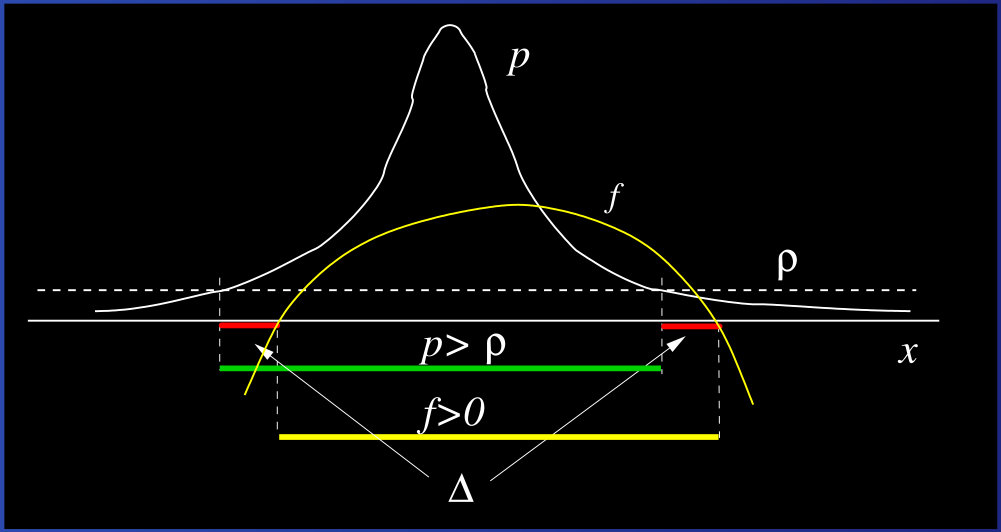
ML^* + Anomaly Detection

- Definition of Anomaly: x is anomalous if its density values falls below a threshold, i.e. $p(x) \leq \rho$.



ML^* + Anomaly Detection

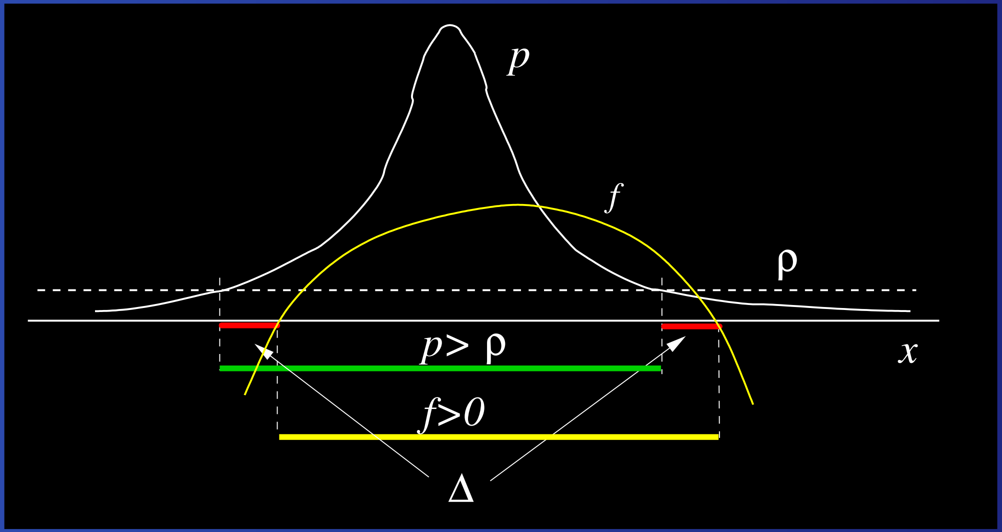
- Definition of Anomaly: x is anomalous if its density values falls below a threshold, i.e. $p(x) \leq \rho$.



- A detector f predicts an anomaly when $f(x) < 0$.

ML^* + Anomaly Detection

- Definition of Anomaly: x is anomalous if its density values falls below a threshold, i.e. $p(x) \leq \rho$.



- A detector f predicts an anomaly when $f(x) < 0$.
- Criterion Function: the labeling error rate, $e(f) = P(\Delta)$

ML^* + Anomaly Detection

- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (unlabeled) training data is sampled from that process

ML^* + Anomaly Detection

- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (unlabeled) training data is sampled from that process
- Validation:
  - Empirical: *no reliable method for estimating $e(f)$!*

ML^* + Anomaly Detection

- Information:
  - First Principles: the operating environment is characterized by a stationary random process
  - Empirical: (unlabeled) training data is sampled from that process
- Validation:
  - Empirical: *no reliable method for estimating $e(f)$!*
  - Theoretical: Substantial work on the accuracy of density estimation methods, *but little work on their accuracy with respect to $e(f)$!*

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is
  - calibrated with respect to e, and

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is
  - calibrated with respect to e, and
  - can be reliably estimated from sample data

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is
  - calibrated with respect to e, and
  - can be reliably estimated from sample data
- Consequences:

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is
  - calibrated with respect to e, and
  - can be reliably estimated from sample data
- Consequences:
  - *empirical validation is now possible!*

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is
  - calibrated with respect to e, and
  - can be reliably estimated from sample data
- Consequences:
  - *empirical validation is now possible!*
  - direct solution methods can now be developed for AD

ML^* + AD = Recent Discovery

- LANL discovered a function $\bar{e}$ that, with a mild assumption on the distribution, is
  - calibrated with respect to e, and
  - can be reliably estimated from sample data
- Consequences:
  - *empirical validation is now possible!*
  - direct solution methods can now be developed for AD
  - LANL has developed a direct solution method with properties similar to SVMs for supervised classification

ML^* Philosophy

ML^* Philosophy

- Without a performance criterion model design is trivial.

ML^* Philosophy

- Without a performance criterion model design is trivial.
- Validation is the cornerstone of the scientific method.

ML^* Philosophy

- Without a performance criterion model design is trivial.
- Validation is the cornerstone of the scientific method.
- Empirical and theoretical validation have different strengths and weaknesses, and having both provides a complete picture.

ML^* **Philosophy**

- Without a performance criterion model design is trivial.
- Validation is the cornerstone of the scientific method.
- Empirical and theoretical validation have different strengths and weaknesses, and having both provides a complete picture.
- FP and EMP information are both critical for success.
  Myth: *All you need is data.*

ML^* Philosophy

Why there will never be a Nobel Prize in ML^*

ML^* Philosophy

Why there will never be a Nobel Prize in ML^*

- *Performance matters, but a first principles interpretation of the model does not.*

ML^* Philosophy

Why there will never be a Nobel Prize in ML^*

- *Performance matters, but a first principles interpretation of the model does not.*
- Myth: A FP model is necessary to achieve good performance. example.

Challenges of Modern Data

Biggest challenge is *not* large amounts of data, but rather

- the lack of relevant information

Challenges of Modern Data

Biggest challenge is *not* large amounts of data, but rather

- the lack of relevant information

  $\boxed{\text{large amount of data} \;\not\Rightarrow\; \text{large amount of information}}$

Challenges of Modern Data

Biggest challenge is *not* large amounts of data, but rather

- the lack of relevant information

  $\boxed{\text{large amount of data} \;\not\Rightarrow\; \text{large amount of information}}$

- the mis–match between the natural structure of the data and the objects of interest

Challenges of Modern Data

Biggest challenge is *not* large amounts of data, but rather

- the lack of relevant information

  $\boxed{\text{large amount of data} \not\Rightarrow \text{large amount of information}}$

- the mis–match between the natural structure of the data and the objects of interest
- the (increasing) gap between existing tools and the problems we want to solve

Challenges of Modern Data

Biggest challenge is *not* large amounts of data, but rather

- the lack of relevant information

  $\boxed{\text{large amount of data} \;\not\Rightarrow\; \text{large amount of information}}$

- the mis–match between the natural structure of the data and the objects of interest
- the (increasing) gap between existing tools and the problems we want to solve
- and last but not least ...

Challenges of Modern Data

The lack of a scientific problem formulations !

Challenges of Modern Data

The lack of a scientific problem formulations !

- How well does Google work?
- What is a meaningful performance criterion?
- How can it be validated?

THE END

Thank You!

SVMs and Curse of Dimensionality

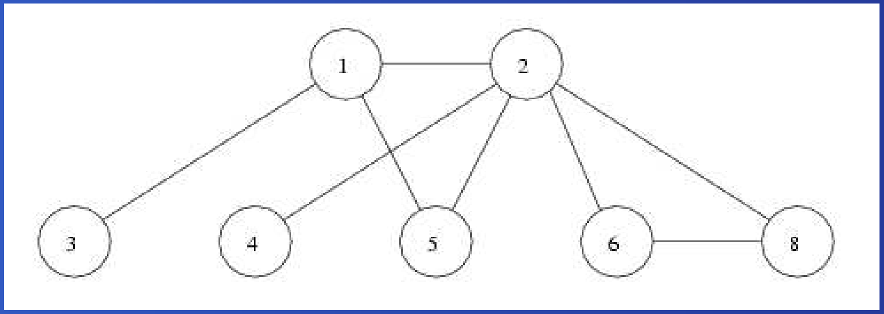
DARPA Intrusion Detection Data

Dimension	Error Rate (%)
27	0.47
4×10^2	0.18
2×10^7	0.14

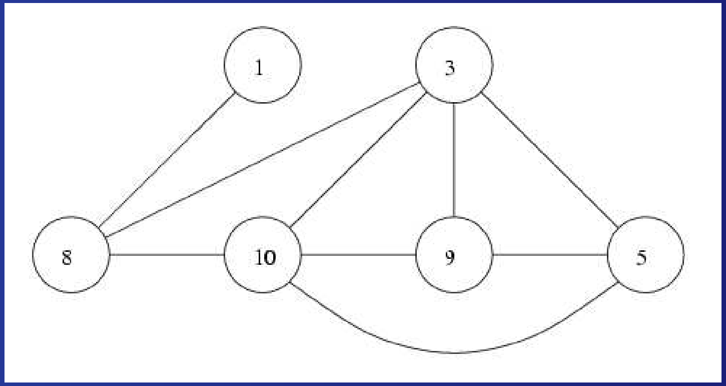
return

Anomaly Detection on Graphs

- Individual graphs represent the interaction between people in a text unit (e.g book, magazine, newspaper, report, or *sections* of these types of documents).
- People (vertices) are labeled by their rank (1 = most important).



Normal Graph

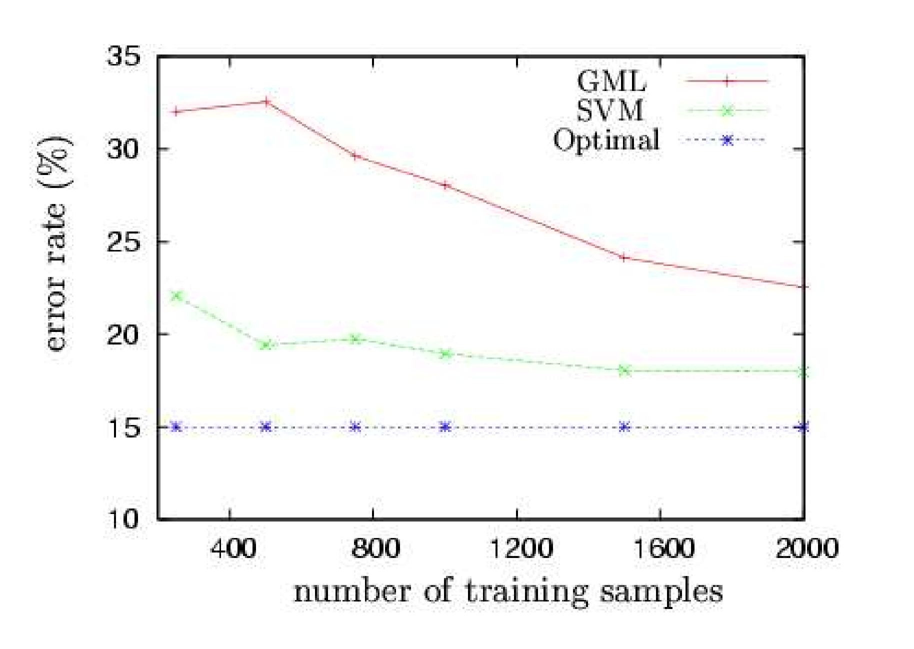


Anomalous Graph

return

Gaussian Benchmark Problem

- The data is Gaussian
- The Gaussian Maximum Likelihood (GML) method uses a first principles model and the SVM uses a universal model.



return