Machine Learning Threats and Opportunities for Debian and Free Software

Machine Learning Models and Source Data

Pablo Ariel Duboue

Debian Conference 2012, Managua

Outline

Outline

This Talk in a Nutshell.

- Status Quo is good for now.
- Many threats need to be addressed outside of Debian (e.g., licensing).
- The opportunities can be tackled by multi-distro efforts.

What is Machine Learning.

- Statistical modelling with focus on predictive applications.
- Common case:
  - Training/estimation/"compilation?"
    - input: vectors of features, including target feature (**data**)
    - output: trained **model**
  - Execution/prediction/"interpretation?"
    - input: vector of features (w/o target feature) plus trained model
    - output: predicted target feature

Example.

- Stanford Syntatic Parser.
  - Java, GPL licensed
  - Mature code, surprisingly well-written
- Probabilistic Context Free Grammar (2Mb trained model)
  - Source: Penn Treebank 640Mb (compressed)
- ```
  (S (NP (DT An) (VBG operating) (NN system))
    (VP (VBZ is)
      (NP
        (NP (DT the) (NN set))
        (PP (IN of)
          (NP (JJ basic) (NNS programs)
          (CC and)
          (NNS utilities)))
        (SBAR (WHNP (WDT that)) (S (VP (VBP make) (NP (PRP$ your) (NN computer) (NN run)))))))
  ```
  ```

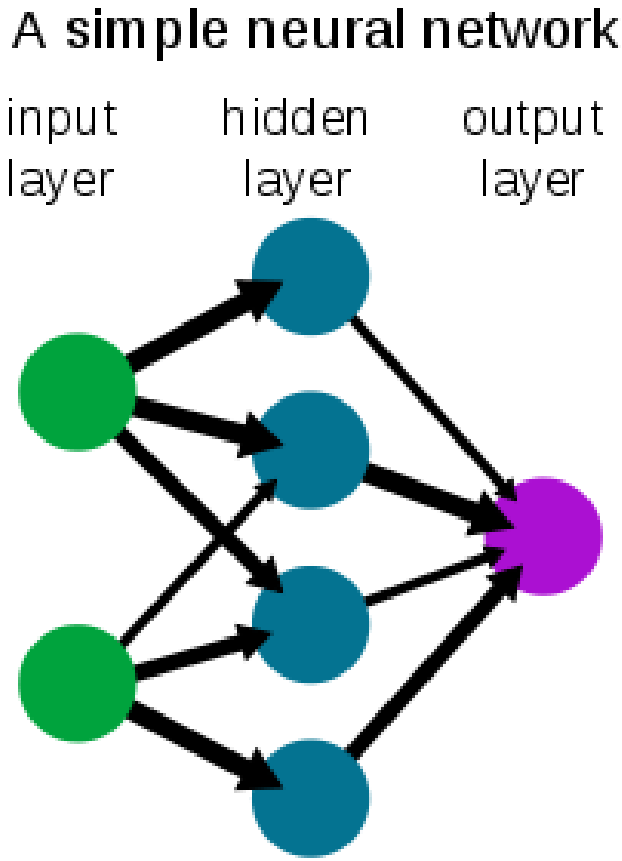
What is a Model?

- Depends on the machine learning methodology employed.
  - Some models **are** easy to understand and modify by hand.
  - *"Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach"* Hatzivassiloglou, Duboue and Rzhetsky (2001)
    - after DEVELOPMENT is present
      after ET is present
      before DATA is NOT present $\Longrightarrow$ class gene [91.7%]
    - before THAT is NOT present
      before FRAGMENT is NOT present
      before ALLELE is present $\Longrightarrow$ class gene [93.9%]
    - after ENCODES is present
      before ENCODES is NOT present $\Longrightarrow$ class gene [96.5%]

Incomprehensible Models.

- Most models being used nowadays are not intendend to be understood as such nor modified by hand
  - Neural networks
  - Support Vector Machines
  - Markov Models
  - Conditional Random Fields

Neural Network.

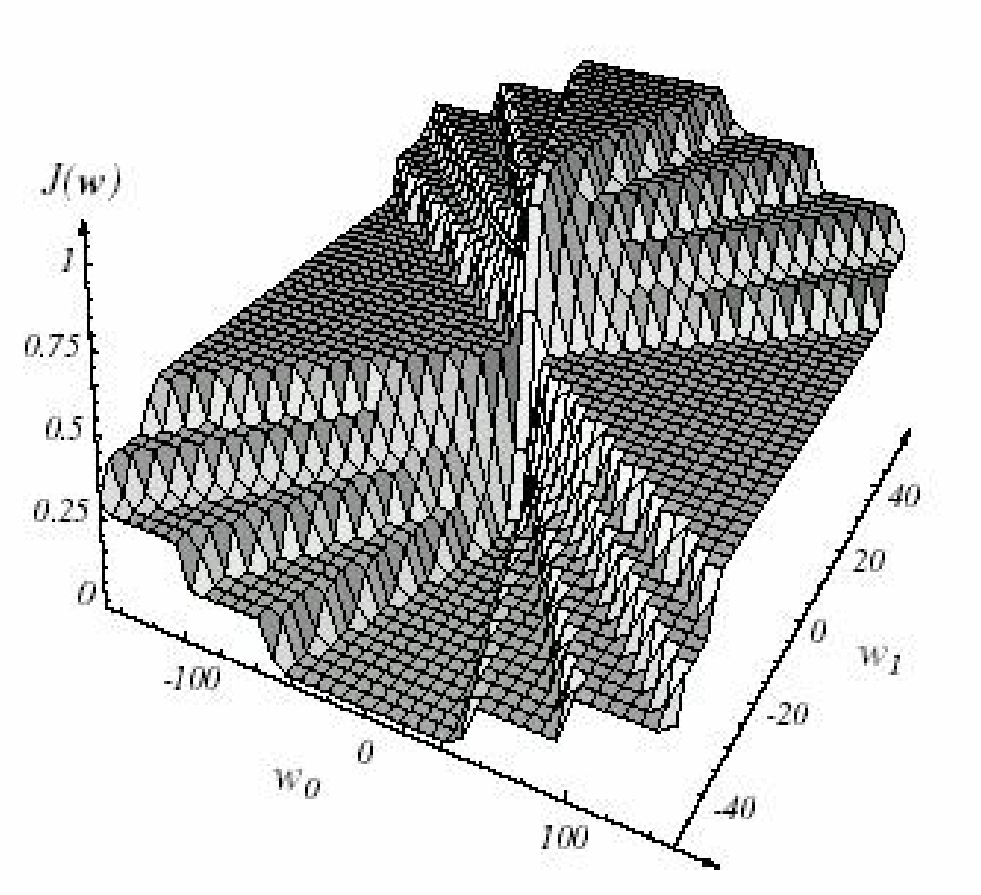


from http://en.wikipedia.org/wiki/Neural_network

- For a three layer network with n,m,l nodes per layer, the model

Weight Space Representation

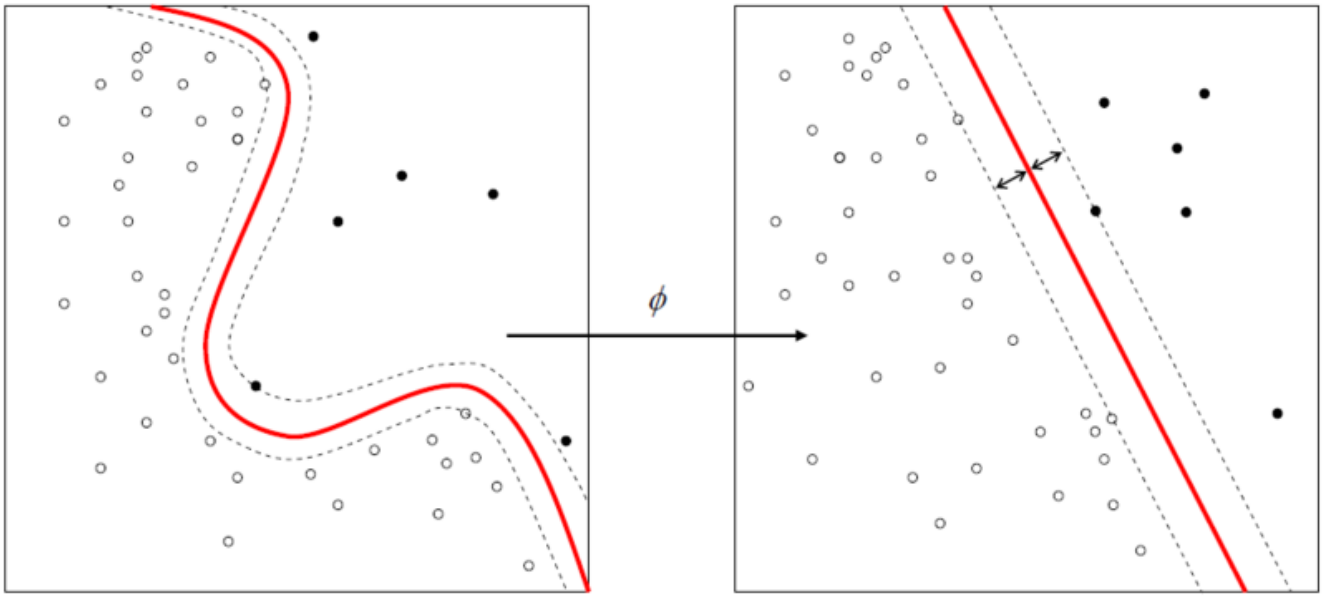
- 1-bit neural network



from http://www.byclb.com/TR/Tutorials/neural_networks/ch10_1.htm

Support Vector Machines

- Find a hyperplane that divides positive from negative training inputs
- Kernel trick:
  - Map input features into a higher dimension feature space
  - Find a hyperplane in the higher dimension



from http://en.wikipedia.org/wiki/Support_vector_machine

Training Data vs. Features

- Feature vectors are not unlike generated YACC (or Bison) C files.
- Examples
  - Speech
    - Training data: transcribed speech
    - Feature data: wave segments with associated transcription
  - Spelling correction
    - Training data: Wikipedia history
    - Feature data: edits that modify a word with less than 3 characters total edit
  - Syntactic Parsing
    - Training data: newspaper articles bracketed and annotated with syntactic categories
    - Feature data: trees of height one, with the most important word of it ("lexical head")

Outline

Threats to Freedom.

- The main threat is **obsolescence**
  - What are we going to do if type of applications users grow to expect and enjoy in privative platforms rely on large train sets?
  - Not unlike the threat posed by cloud services being addressed by the FreedomBox foundation.
- Applications such as
  - OCR (book scanning)
  - Speech Recognition (dictation)
  - Computer Vision (automatically tag your friends on photos)
  - Question Answering (Siri / Watson)

Diminishing Value Behind Source Code

- Value on the data
  - Facebook
  - LinkedIn
  - Google+
  - Flickr
- Data vendors
  - http://www.infochimps.com/marketplace (general data, including Twitter data)
  - http://www.ldc.upenn.edu (linguistic data)

Yet-another-clever-GPL-circumvention trick?

- Vendor releases the source code but keeps the data behind the trained model closed.
- Not unlike firmware binary blobs?
  - To me, the firmware binary blobs are a much better analogy to machine learning models than video game assets.

Threats to Practicality.

- Training machine learning models takes a whole different type of build-machine
  - 64Gb of RAM for 3 days, sure!
  - Why? Oh my, why?
- Distributing training data involves order of magnitude more space and bandwidth
  - Comparable to wikimedia mirroring (or more)

debian-legal circa 2009.

- Original message: http://lists.debian.org/debian-legal/2009/05/msg00028.html
- Mathieu Blondel asked two questions:
  - Can Debian ship models in main without distributing the original data?
    - Yes, because the model is considered the preferred form for modification.
    - The reasoning followed a pre-existing decision from 2D rendered images for games (rendered from an underlining 3D model).
  - Can violations of data licensing be detected? (Debian off-topic)
    - Artificially introduced errors for fingerprinting

Some Quotes

- "Free data is important for the very same reason that free programs are!"
  - Mark Weyer (Wed, 27 May 2009 11:36:55 +0200) <20090527093654.GF24759@athen.informatik.hu-berlin.de>
- "[then do not ship] pictures that are initially photographs of an object (the preferred form of modification is the original object; if you want to see it at another angle, you need to take another photograph)"
  - Josselin Mouette (Wed, 27 May 2009 10:33:52 +0200) <1243413232.14420.49.camel@shizuru>

Real Issues in the Debian Archive.

- A cursory search did not reveal anything immediate
- Possible leads:
  - rdkit
  - opencv
  - UIMA sandbox

Outline

Opportunities for Debian.

- Main challenge for Debian IMO is to change users into contributors
- Contributors volunteering new training data can follow the success case of Translators
- Data contributors can
  - Annotate more data to fix a bug
  - Bugs with "data patches"

Opportunities for Debian Collaboration with other Distros.

- Inter-distro collaboration opportunities.
  - Sharing data is easier than sharing code as its format seldom changes.
    - Think object-orientation.
    - All syntactic parsers in the last 15 years of work in the field have used the same Penn Treebank data set.
  - Sharing annotation work is easier than sharing data patches.
    - Think work on i10n

Questions.

- How can we acquire the data?
  - Maybe build a Free Software-volunteer driven Mechanical Turk-like tool?
  - Not unlike BOINC.
  - Build on the success of initiatives like LibriVox (http://librivox.org)
- How can we assure the data is kept Free?
  - CC-SA and derivatives?
  - Is GPL enough?

Mechanical Turk

- An Amazon Web Services offering.
- Write a task that is easy to do by humans (e.g., "is there a person in this picture") but difficult for computers.
- Have paid workers ("turkers") do these tasks for tiny wages.
- Plenty of ethical issues
  - http://chronicle.com/blogs/profhacker/the-ethics-of-amazons-mechanical-turk/23010
  - Is exploitative towards turkers?
  - Are turkers helping anonymously projects against their own moral values?

VoxForge.

- GPL transcribed sound samples.
  - Acoustic models then have to remain Free.
  - I recently found out about it, still not sure whether GPL would really do it.
- Currently, the tool to train acoustic models is proprietary (HTK) but the speech recognition engine is Free software (Julius).
- If we want to package VoxForge models will we have to also distribute the source data?

Outline

To Sum Up

- Don't shoot the messenger
  - Not doing anything is an option... for now
  - Any pointers for licensing?
    - VoxForge uses GPL for the data
  - Any pointers for an inter-distro model training project?

Where to go from here.

- Revisit current policy?
  - How can we transition away from considering trained models as the "preferrable form for modification"?
  - Agree to host such packages in contrib?
  - Differentiate between Debian archive hosting all the code and assets vs. source data?
- Discuss with archive.org for hosting data?
  - Or any other large scale archiving service... suggestions?
- Discuss with Grid 5000 or Amazon Cloud for training?
  - Grid 5000 was used for the clang 2.9 and 3.0 Debian archive rebuilds (http://clang.debian.net/)