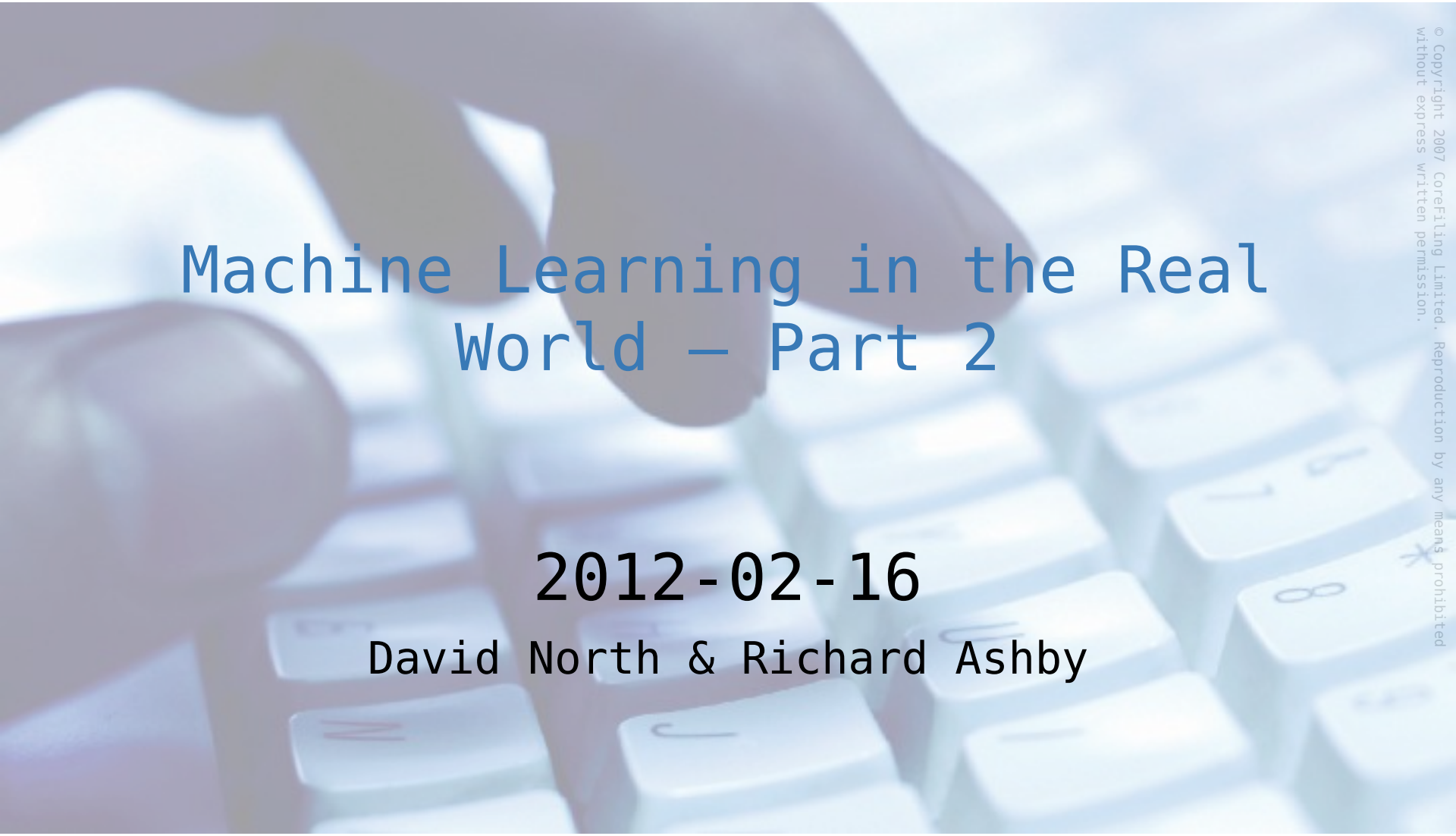

Machine Learning in the Real World – Part 2

2012-02-16

David North & Richard Ashby

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

CoreFiling

Contents

- Improving accuracy & performance:
  - Tweaking the library
  - Processing the inputs
- Deploying to the real world
- Scaling with the data
- Monitoring
- Summary

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Making Improvements

- Crucial to have tests in place first
- Improving accuracy vs improving performance
- Types of classifier:
- Decision Tree, Naive Bayes, Max Ent
  - Is your classifier a black/white box?
  - A DT can be debugged more easily

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Explore the Library

- Variants on theme of MaxEnt:
- L1 and L2 regularization
- Number of corrections
- What is the tradeoff performance vs accuracy?
- -> Must have tests in place first

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Modify the Inputs?

- Input is text, human readable
- Temptation is to process, use our knowledge of domain/language to give the classifier a helping hand
- Resist this – the mind is an imperfect classifier
- Aim to model how the world is, not how we think of it
- Do however look at regularisation of inputs

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Process the Inputs

- Remove irrelevant differences
  - Capitalisation, punctuation
  - Stemming - “accounts”, “accounting”, “accounted”
  - Spelling conventions - “license” vs “licence”
- Discard “invalid” data
- Synonyms?
  - “expense” == “costs”? In all contexts?
- Antonyms?
  - Profit/Loss can be the same thing
- Depends on the domain

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Real World Deployment

- Fast, reliable, adaptive
- Must ship with built-in data
- Classifier slow to build, quick to respond
- Need to be able to serialize results
  - Gives fast startup
  - Custom serialization vs java Serializable

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Real World Deployment

- Other real world concerns
- Sharing data
- Migrating data
- Fallback
  - What if you have little or no data?
  - What if confidence is low
  - Our tactic, use search / composite

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Scaling - Problems

- Data increasing constantly, over 100k items
- Diversity of data increasing, both number of features (f) and size of universe (u)
- Critical value is product of f * u.
- Reached 15.6 million in live service
- Problems
  - Classifier build prohibitively slow (4:35hrs)
  - Classifier build uses > 2GB memory
  - Serialized classifiers > 1GB, problems storing to db
  - The whole service is at risk if any of the above goes wrong

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Scaling - Solutions

- Limit classifier building to “quiet” times
- Separate classifier building from web application
- Stemming etc. reduces the number of distinct features in play
- Store classifiers as discrete units not a collection
- Always use compression!
- Limit the amount of training data used – prioritise most recent
- How do you know what effect this is having on accuracy?

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Monitoring Performance

- Realistic data is vital for testing – hard to generate
- Developers won't necessarily have access to this data

Logging

- Record stats, how much data, how long, how much memory
- Assumptions about the shape of your data invalid?
- Have these logs mirrored, convenient to view
- Record the confidential inputs, access on request

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Monitoring Accuracy

- How do you know if you are giving the right answers?
- Much harder to track than performance
- Customer feedback
  - Time consuming, subjective
- Track how many suggestions were accepted
  - self-perpetuating, if users are lazy
- Get the real data and perform x fold tests
  - Not allowed the data
  - Obfuscate the real data and then test?

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Monitoring Accuracy Solution

- Run tests on the live data in original format and only report results
- Report statistics – why is our feature set so large?
- Report accuracy – what is success rate/confidence when run on existing data
- Currently not using x fold tests, just on everything
  - Not as fair, as we are comparing to seen content
  - Much quicker as don't need to rebuild classifiers

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

Summary

- Test first, record performance and accuracy
- Regularize your input data
- Save your work – serialize classifiers
- Perform batch tasks offline
- Monitor what's happening

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.

CoreFiling Ltd

Want to solve interesting problems like this?

...come work with us

www.corefiling.com/careers

© Copyright 2007 CoreFiling Limited. Reproduction by any means prohibited without express written permission.