GOOD FEATURES

ENGINEERING TRICKS

martin @ reddragon.ai
sam @ reddragon.ai

27 September 2017

`WiFi : SG-Guest`

Problems with Installation? **ASK!**

RED DRAGON AI

PLAN OF ACTION

TODAY

- Math (slower)
- Creating Good Features
- Intros and Final Project scope discussion
- Play with Project 01

PLAN OF ACTION

NEXT MONDAY

- Start with project wrap-up
- CNNs - the basics

GOOD FEATURES

- Magic of DL :
  - hand-crafted features not essential
- But : Does no harm to provide something sensible

OVERVIEW

- Numerical inputs
- Categorical inputs
- 'Going deeper'
- Applications

NUMERICAL INPUTS

- Going to multiply by weights
- Helpful to have them similar in size
- Easy choice is identically distributed :
  - [0 ... 1]
  - [-1 ... 1]
  - ~NormDist(mean=0, sigma=1)
- Forcefully rescale them :
  - By known factor (or rule-of-thumb)
  - By learned factor, eg: `BatchNorm()`

NUMERICAL INPUTS

- Also helpful to have them non-colinear :
  - Ideal : Independent, identically distributed (IID)

CATEGORICAL VARIABLES

- True or False
- spring / summer / autumn / winter
- ** or **** reviews
- 2 or 4
- 18 or 34
- country:usa or country:singapore
- country x age

BOOLEAN

- False → `0`, True → `1`

  or

- False → `10`, True → `01`

ONE-HOT

- spring → `1000`
- summer → `0100`
- autumn → `0010`
- winter → `0001`

ORDERED

- '*' → 0000
- '**' → 1000
- '***' → 1100
- '****' → 1110
- '*****' → 1111

LINEAR-LIKE

- Use as-is; or
- 3 → `000100` (one-hot, sparse)
- 3 → `111100` (as before)

FLOAT-LIKE

- Bucketing : convert to categorical
  - 0-12 → `100000`
  - 12-15 → `010000`
  - 15-18 → `001000`
  - 18-21 → `000100`
  - 21-25 → `000010`
  - 25-99 → `000001`
- ... and use original too

MANY CLASSES

- one-hot is Ok
  - eg: 100s
- But are you thowing away known structure?

VERY MANY CLASSES

- Hashing trick (for, say, 1000 buckets):
  - Convert label into a 'hash'
  - ... something like `MD5sum`
  - One-hot category is `( hash % 1000 )`
- 'infinitely extensible'
- Is collision a problem?

COMBINATIONS OF CATEGORICALS

- Apply hashing trick to cojoined labels :
  - country x age →

    ```
    hash("country:usa~age:18-21") % 1000
    ```
- Again : 'infinitely extensible'
- Can store lots of coincidence information 'linearly'

NOW WHAT?

- Each feature position ...
  - ... contributes (or not) the corresponding weight to several nodes in the next layer
- Can't do many layers :
  - 1000 categories ⇒ ~1MM parameters per dense layer

EMBEDDING

- Instead of {0,1} in a single location (sparse):
  - Use location as index into 'dictionary'
  - That location has a learnable vector (~8d)
  - Use the vector as inputs to next layer
- Can learn arbitrary input features
- All trainable via Backpropagation

WHO CARES?

- Applications would have to have :
  - Lots of kinds of data
  - Lots of interactions between factors
  - Lots of data
  - Need for speed
- Suggestions?

RECOMMENDATION

- Scoring of different recommendations
- Presentation :
  - Most likely next action
  - Second-best alternative
  - People-like-you also-like ...

ADVERTISING

- Predict click-through rates
- Intense area of study
- Lots of this is not published, though

PUBLISHED RESULTS

- Recommender system (Google Play store) :
  - Deep & Wide networks (June-2016)
  - ... implementations available
- Ad Click Predictions (Google) :
  - Deep & Cross networks (August-2017)
  - ... too new

FURTHER READING

- Wide & Deep :
  - TensorFlow tutorial
  - Kitchen-sink notebook
  - Presentation with even more stuff (up to TFserving)
  - Keras code (but advanced feature creation is 'TODO')
- Cross & Deep :
  - ... your company?
  - ... your project?

- QUESTIONS -

MARTIN.ANDREWS @
REDDRAGON.AI

My blog : http://mdda.net/

GitHub : mdda

RED DRAGON AI