18.05.22

OPENCAMPUS.sh

Transformers for Natural Language Processing and Beyond

THE DATASETS LIBRARY

- Quiz
- Dataset Characteristics
- Coding Examples
- Baseline Models

QUIZ

https://forms.office.com/r/F3pb9AnYik

CODING EXAMPLES

DATASET CHARACTERISTICS

- Is your collection of samples possibly biased?
- How must the data be collected to be used with your model?
- For classification problems:
  - Is your sample balanced across all classes?
  - If not, how will you deal with it?

SHORTCOMINGS OF LANGUAGE MODELING

- Human Reporting bias (Gordon and Van Durme, 2013):
  - Not stating the obvious
  - Common sense isn‘t written down
- Facts about named entities
- No grounding to other modalities

Possible Solutions:

- Incorporate structured knowledge (e.g. databases; Zhang et. A. 2019)
- Multimodal learning (e.g. visual representations; Sun et al. 2019)

PROJECT MILESTONES

- 11.05. Form project groups
- 18.05. Literature review
- 25.05. Dataset characteristics
- 01.06. Baseline model
- 08.06. Model & model evaluation (Joint Coding)
- 15.06. Project presentations

CHARACTERISTICS OF BASELINE MODELS

- Should be simple to setup, with a reasonable chance of providing decent results, and very unlikely to overfit.

- Should be interpretable, which can help your understanding of the data and guide your feature engineering.

Ameisen, E. (2018, March 6). *Always start with a stupid model, no exceptions.* Medium. https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa

POSSIBLE TYPES OF BASELINE MODELS

- **Linear Regression**
  A solid first approach for predicting continuous values (prices, age, ...) from a set of features.

- **Logistic Regression**
  When trying to classify structured data or natural language, logistic regressions will usually give you quick, solid results.

- **Gradient Boosted Trees**
  A Kaggle classic! For time series predictions and general structured data, it is hard to beat Gradient Boosted Trees. While slightly harder to interpret than other baselines, they usually perform very well.

- **Simple Convolutional Architecture**
  Fine tuning VGG or re-training some variant of a U-net is usually a great start for most image classification, detection, or segmentation problems.

BASELINE MODEL RESULTS

Help you understand your data:

- **Which classes are harder to separate?**
- **What type of signal picks your model up on?**
  How is your model making decisions?
- **What signal is your model missing?**
  Is it possible to engineer an additional feature?

IF THERE IS NO EFFECTIVE BASELINE

- **Instead of simplifying the model, simplify the data.**

- **Try to get your complex model to overfit to a very small subset of your data.**

TODOS UNTIL NEXT WEEK

- Complete at least two sections from chapter 7 (Main NLP Tasks) of the Hugging Face course

- Calibrate a First Baseline Model