Overview

I Domain.

- Discriminative models: a class of machine learning methods where the focus is on learning class memberships, as opposed to Generative models, where the interest is in full class densities.
- Maximum Entropy models: discriminative methods widely used e.g. in NLP.
- Maximum Entropy models encode relevant statistics about the training data through features.

II Problem.

- In general, the quality of the model grows with the number and scope of features.
- Unfortunately, computational and memory resources needed to manipulate features also grow accordingly, often to an unmanageable extent.

III Approach.

- We investigate the possibility of representing features using randomised techniques.
- We explore a class of one-sided error randomised data structures derived from the Bloom Filter (Bloom, 1970).
- We concentrate on the log-frequency Bloom Filter and on the Bloom Map [Talbot et al., 2008].

IV Results.

We provide details of a Randomised Maximum Entropy Framework, where we examine integration possibilities and propose an integration scenario. A number of interesting findings lead us to argue that the Randomised Maximum Entropy model and the Exact Maximum Entropy model deliver comparable performances, with remarkable space savings introduced by our model.

1. The Maximum Entropy Principle

Model all that is known and assume nothing about what is unknown.

Models with high entropy are more uniform, and correspond to assuming less about the world: nothing more than what is evidenced by the data.

We choose the distribution \( p \) that maximises

\[
H(p) = -\sum_{\xi \in \mathcal{E}} p(\xi) \log p(\xi).
\]

Features.

A feature [Berger et al., 1996]

\[
f : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}
\]

is a binary function that takes the form

\[
f_{p,y}(x, y) = \begin{cases} 
1 & \text{if } y = y' \text{ and } cp(x) = \text{true} \\
0 & \text{otherwise},
\end{cases}
\]

A feature has the purpose of checking the co-occurrence of:

- some desirable linguistic information in a context \( x \);
- some prediction \( y' \).

Several features are needed in a modelling problem. The human modeller decides their number, scope, complexity.

Example

\[
f_{p,y}(x, y) = \begin{cases} 
1 & \text{if } y = \text{sci-fi} \text{ and } \\
\text{doccontains}_{\text{VENUSIAN}}(x) = \text{true} \\
0 & \text{otherwise}
\end{cases}
\]

2. Randomised Data Structures

A huge number of features means more nuisances of the data are captured. It also means huge amounts of data to deal with.

The Idea.

Internally encode features through RDSs. Trade-off:

- \( \{\} \) introduce a (tunable) False Positive probability;
- \( \{\} \) obtain remarkable space savings / deal with more data at once.

3. Case Study

The Truecasing Problem.

Given we want our classifier to output.

\[
\begin{align*}
\text{we provide details of a Randomised Maximum Entropy Framework, where we examine integration possibilities and propose an integration scenario. A number of interesting findings lead us to argue that the Randomised Maximum Entropy model and the Exact Maximum Entropy model deliver comparable performances, with remarkable space savings introduced by our model.}
\end{align*}
\]

4. Experiments & Results

Classification Accuracy compared when using our new approach.

4.1 Results.

Verification (A)

- Verified feasibility of using novel RDSs in a discriminative machine learning context.
- Optimal FP/Quantiﬁcation values allow for accuracy ﬁgures comparable to non-randomised framework.
- The combination of even bigger datasets and deeper randomisation might lead to improvements upon current accuracy ﬁgures.

References


Acknowledgements

Miles Osborne, David Talbot, Andy Loveland. SAAS (Student Awards Agency for Scotland).