

III-Logics

Estimating the Parameters of an MLN

Logik für Erklärbare KI: Technische Einführung in das ENEXA Projekt

Maria Schnödt-Fuchs, Alex Goëßmann

Funded by the
European Union



15.+16. July, 2024

Definition

The Shannon entropy of a probability distribution is defined as

$$S[\mathbb{P}] = \sum_{j_0, \dots, j_{d-1} \in \times_{k \in [d]} [m_k]} \mathbb{P}(j_0, \dots, j_{d-1}) \cdot \ln [\mathbb{P}(j_0, \dots, j_{d-1})]$$

Application in

- ▶ **Information Theory:** Minimal averaged code length to communicate samples
- ▶ **Machine Learning:** Measuring the **degree of structure** in a model

Further entropies

- ▶ Cross entropy: Likelihood of data for objective design
- ▶ Relative entropy (Kullback-Leibler divergence): Overhead by cross entropy as a distance measure of distributions

Let there be a dataset of m samples in the form of a mapping

$$D : [m] \rightarrow \bigtimes_{k \in [d]} [2]$$

to observed states. The likelihood is the product

$$\mathbb{P}^{(\mathcal{F}, w)}[D] = \prod_{j \in [m]} \mathbb{P}^{(\mathcal{F}, w)}[D_j] .$$

The averaged negative log-likelihood is the cross entropy

$$\mathcal{L}_D(\mathcal{F}, w) = -\frac{1}{m} \sum_{j \in [m]} \ln \left[\mathbb{P}^{(\mathcal{F}, w)}[D_j] \right] = \mathbb{H} \left[\mathbb{P}^D, \mathbb{P}^{(\mathcal{F}, w)} \right]$$

For each $f \in \mathcal{F}$ we have

$$\frac{\partial}{\partial w^f} \mathcal{L}_D (\mathcal{F}, w) = \mathbb{E}_{\mathbb{P}^D} [f] - \mathbb{E}_{\mathbb{P}^{(\mathcal{F}, w)}} [f]$$

Algorithm: Alternating Weight Optimization

Iterate through $f \in \mathcal{F}$ until convergence and adjust w^f such that

$$\frac{\partial}{\partial w^f} \mathcal{L}_D (\mathcal{F}, w) = 0.$$

Learning Task

Find a distribution with **minimal structure** reproducing **key observed features**.

Formalize

- ▶ **minimal structure by maximal entropy**
- ▶ **key observed features** by a set \mathcal{F} such that

$$\forall f \in \mathcal{F} : \quad \mathbb{E}_{x \sim \mathbb{P}^\theta} [f(x)] = \mathbb{E}_{x \sim \mathbb{P}^D} [f(x)]$$

We state the **Entropy Maximization Problem**

$$\operatorname{argmax}_{\theta \in \Gamma} \mathbb{S}[\theta] \quad \text{subject to} \quad \forall f \in \mathcal{F} : \mathbb{E}_{x \sim \mathbb{P}^\theta} [f(x)] = \mathbb{E}_{x \sim \mathbb{P}^D} [f(x)] \\ (P_{\Gamma, \mathcal{F}, D})$$

By variational calculus one can proof the following:

Theorem

Among the distributions \mathbb{P} satisfying

$$\forall f \in \mathcal{F} : \mathbb{E}_{x \sim \mathbb{P}} [f(x)] = \mathbb{E}_{x \sim \mathbb{P}^D} [f(x)]$$

the one with maximal entropy is the Markov Logic Network to \mathcal{F} and the maximum likelihood weight w .

- ▶ We did not assume that \mathbb{P} is a Markov Logic Network, this is the result of the theorem!
- ▶ Removing the constraint that f is a map to $[2]$, we can generalize to exponential families.