Maximum entropy (MAXENT) competes with Maximum Likelihood (ML)

Armen Allahverdyan

MAXENT originates in statistical physics

MAXENT in data science

Deciding on the quality of MAXENT via Bayesian theory

MAXENT vs ML (optimally regularized ML)

Co-authors: Narek Martirosyan and Edvard Khalafyan

https://ucsb.zoom.us/j/81193221889

MAXENT in statistical physics

Find unknown probabilities $q = \{q(z_k)\}_{k=1}^n,$

of a random variable $Z = (z_1, ..., z_n)$

via maximization of entropy $S[q] = -\sum_{k=1}^{n} q(z_k) \ln q(z_k),$

→ $q = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$

If constraints are precisely known, e.g. $E(Z) = \sum_{k=1}^{n} q(z_k) z_k.$

Take conditional maximization

$$q^{[1]}(z_k) = \frac{e^{-\beta z_k}}{\sum_{l=1}^{n} e^{-\beta z_l}},$$

Physical example: fixed average energy; Gibbs distribution

But why?

Botzmann, Gibbs (1880-90): second law of thermodynamics for an isolated system

Jaynes (1957): entropy is an uncertainty measure + we must not cheat in inference

Many schemes for deducing entropy. Three axioms by *Chaundy & McLeod, 1960*

1. Only different probabilities matter

$(p_1, ..., p_n)$ versus $(q_1, ..., q_n)$

$$S[p_1, ..., p_n] = \sum_{k=1}^{n} \psi(p_k),$$

2. Concavity

$\psi(x)$ is concave: $\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} \leq 0.$

$$S\left(\frac{1}{n}, ..., \frac{1}{n}\right) \geq S(p_1, ..., p_n) \geq S(1, 0, ...0)$$

3. Additivity:

$\{p_k p'_l\}_{k=1}^{n} {}_{l=1}^{n'}$.

$$S[p_1 p'_1, ..., p_n p'_{n'}] = S[p_1, ..., p_n] + S[p'_1, ..., p'_{n'}].$$

MAXENT in data science.
Huge number of applications

Data

i.i.d. sample of length M:

$$\mathcal{S} = (z_{i_1}, ..., z_{i_M}).$$

m_k is the number of appearances of z_k $\quad Z = (z_1, ..., z_n)$

Sparse data: *M<n*

MAXENT advantages:

-- non-parametric estimate
-- does not report zero probabilities if $m_k = 0$

$$\mu_1 = \frac{1}{M}\sum_{u=1}^{M} z_{i_u} = \frac{1}{M}\sum_{k=1}^{n} z_k m_k \overset{?}{=} \sum_{k=1}^{n} q_k z_k.$$

$$q^{[1]}(z_k) = \frac{e^{-\beta z_k}}{\sum_{l=1}^{n} e^{-\beta z_l}},$$

What to take as constrain(s) for MAXENT ?

Or the second moment?
Or both ? Or none ?

Prone to overfitting ? Noise !

$$\mu_2 = \frac{1}{M}\sum_{u=1}^{M} z_{i_u}^2 = \frac{1}{M}\sum_{k=1}^{n} z_k^2 m_k \overset{?}{=} \sum_{k=1}^{n} q_k z_k^2,$$

How to compare with other estimators?

regularized maximum likelihood (ML)

$$p_{\mathrm{ML}}(z_k) = \frac{m_k + b}{M + nb}, \qquad M \equiv \sum_{k=1}^{n} m_k, \qquad b \geq 0,$$

b=0: usual ML
b>0: does not report zero probabilities

Bayes and Laplace said *b=1* is fine (No)

We need decision and comparison criteria for MAXENT

Bayesian decision theory

$q = (q_1, \ldots, q_n)$ unknown probabilities come from a known prior density $\mathcal{D}(q)$

$(z_{i_1}, \ldots, z_{i_M})$, sample multinomial density $m = \{m_k\}_{k=1}^n$, $P(m|q) = M! \prod_{k=1}^n \frac{q_k^{m_k}}{m_k!}$

Quality of the estimator → KL-distance.
Several convenient features.

$$K[q, \hat{q}(m)] = \sum_{k=1}^n q_k \ln \frac{q_k}{\hat{q}_k(m)}.$$

Mean distance for fixed data (risk)

$$\overline{K} = \int \mathrm{d}q \, P(q|m) K[q, \hat{q}(m)]$$

The optimal, parametric estimator for KL distance.

We do not focus on this estimator: too *ad hoc*

$$\mathrm{argmin}_{\hat{q}} \left[\overline{K} \right] = \int \mathrm{d}q \, q \, P(q|m).$$

double-average: over the prior AND data

Our main quantity

$$\langle \overline{K} \rangle = \int \mathrm{d}q \, \mathcal{D}(q) \sum_m P(m|q) K[q, \hat{q}(m)]$$

Non-informative prior
= independence + covariance
= Dirichlet density

$$\mathcal{D}(q) \propto \delta(\sum_{k=1}^{n} q_k - 1) \prod_{k=1}^{n} q_k^{\alpha_k - 1}$$

independence

Dirichlet is covariant under various aggregations $Z = (z_1, z_2, z_3) \to Z' = (z_1 + z_2, z_3)$

Dirichlet is sampled via i.i.d. $\{\zeta_k\}_{k=1}^{n} \quad \zeta_k \sim \rho(\zeta_k). \qquad q_k = \frac{\zeta_k}{\sum_{l=1}^{n} \zeta_l},$

Dirichlet is the only density that has independence in the two senses

Most probable values vs average:

$\alpha_k = \alpha < 1 \quad \text{m.p.}[\mathcal{D}(q)] = \{ (1, 0, ..., 0), ..., (0, 0, ..., 1) \}$

$\langle q \rangle = \frac{1}{n}$

$\alpha > 1 \quad \text{m.p.}[\mathcal{D}(q)] = \left(\frac{1}{n}, ..., \frac{1}{n} \right)$

MAXENT is meaningful if for some constraints $\langle \overline{K} \rangle < \langle \overline{K} \rangle_{q=1/n}$ no constraints

For synthetic data we take: *Z=(1,..,n)*

For sparse data *M<n* and $\alpha > 1$ MAXENT is meaningless for all constraints

For sparse data *M<n* and $\alpha < 1$ MAXENT is meaningful for *M>10*

You need prior knowledge to apply MAXENT

Overfitting !
Smaller M → less constraints

$$\langle \overline{K} \rangle_1 < \langle \overline{K} \rangle_{1+2}$$

$$\mu_2 = \frac{1}{M} \sum_{u=1}^{M} z_{i_u}^2 = \sum_{k=1}^{n} q_k z_k^2,$$

Optimally regularized ML beats MAXENT,

because for a single Dirichlet prior this ML is the globally optimal estimator

$$p_{\mathrm{ML}}(z_k) = \frac{m_k + b}{M + nb},$$

$$b = \alpha$$

Prior information: Mixture of Dirichlet priors

$$\mathcal{D}(q) = \frac{1}{2}\mathcal{D}(q_1, ..., q_n | \alpha_1 < ... < \alpha_n) + \frac{1}{2}\mathcal{D}(q_1, ..., q_n | \alpha_n > ... > \alpha_1)$$

$$\langle q \rangle = \frac{1}{n}$$

Prior information $q_1, \cdots, q_n$ tend to be ordered or anti-ordered

Facilitates the applicability of MAXENT
Z=(1<2<...<n)

$$q^{[1]}(z_k) = \frac{e^{-\beta z_k}}{\sum_{l=1}^{n} e^{-\beta z_l}},$$

$$\langle \overline{K}_1 \rangle < \langle \overline{K}_{\mathrm{ML}} \rangle_{b=b_{\mathrm{opt}}}.$$

The optimally regularized ML is beaten for *M>10*

Conclusions

MAXENT looks non-parametric, but even its meaningfulness need prior information

In contrast to ML, MAXENT needs numeric *Z*. **Categorical data is an open**.

MAXENT applies to sparse data, but not to very short samples

Overfitting: more constraints lead to worse results.

The number of constraints should grow with M. **More or less precise rule is open**.

Affine symmetry

Affine transformation

$$\widetilde{z}_k = \mathcal{F}(z_k), \qquad \mathcal{F}(z) = gz + h,$$

$$\mu_1 \to \widetilde{\mu}_1 = g\mu_1 + h$$

$$q^{[1]}(z_k) = \frac{e^{-\beta z_k}}{\sum_{l=1}^{n} e^{-\beta z_l}},$$

MAXENT probabilities are affine invariant if we fix moments: *1* or *1+2 or 1+...+K*

Conjecture: For any constraint that does not respect the full affine asymmetry there is a better constraint that does respect it.

$$\langle \overline{K}_1 \rangle \lesssim \langle \overline{K}_2 \rangle < \langle \overline{K}_{1/2} \rangle < \langle \overline{K}_3 \rangle$$

Explains why you need take integer-order constraints sequentially