Generalised max entropy classifiers

Fabio Cuzzolin

Visual Artificial Intelligence Laboratory
School of Engineering, Computing and Mathematics
Oxford Brookes University

SMPS-BELIEF 2018 - Compiegne, France
September 2018

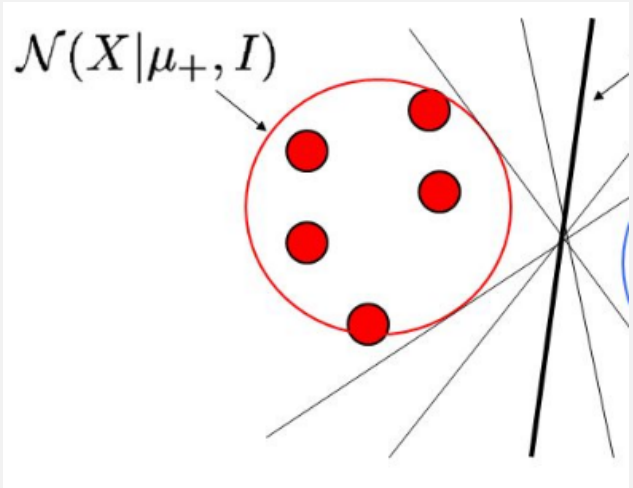
Rationale

- serious issues with current approaches to model adaptation in machine learning
- the crucial question is: what exactly can one infer from a training set?
- max entropy classifiers provide a significant example, simple and widespread
- the entropy of the sought joint (or conditional) probability distribution of data and class is maximised, following the *maximum entropy principle* that the least informative distribution which matches the available evidence should be chosen
- two strong assumptions:
  - training and test data come from the same probability distribution
  - the empirical expectation of the training data is correct, and the model expectation should match it
- we make a step in that direction by generalising the max entropy classification framework

Contributions

- we assume that a training set does not provide sufficient information to precisely estimate the joint probability distribution of class and data
- we assume instead that a belief measure can be estimated, providing lower and upper bounds on the joint probability of data and class
- an appropriate measure of entropy for belief measures is maximised
- the empirical expectation of the chosen feature functions is assumed to be *compatible* with lower and upper bounds associated with the sought belief measure
- leads to a constrained optimisation problem with inequality constraints, rather than equality ones
- needs to be solved by looking at the Karush-Kuhn-Tucker (KKT) conditions
- due to the concavity of the objective function and the convexity of the constraints, KKT conditions are both necessary and sufficient

Outline



1. Maximum entropy classification
2. Belief functions
3. Measures of entropy
4. Generalised max entropy
5. Conclusions and future work

Maximum entropy classification

Setting

- **maximum entropy classifiers** maximise the Shannon entropy of the conditional classification distribution $p(C_k|x)$
- $x \in X$ is the observable and $C_k \in \mathcal{C} = \{C_1, ..., C_K\}$ is the class
- given a training set $\mathcal{D} = \{(x_i, y_i), i = 1, ..., N | x_i \in X, y_i \in \mathcal{C}\}$, ..
- .. a set M of **feature maps** is designed,

$$\phi(x, C_k) = [\phi_1(x, C_k), \cdots, \phi_M(x, C_k)]'$$

whose values depend on both the object observed and its class

- each feature map $\phi_m : X \times \mathcal{C} \to \mathbb{R}$ is then a random variable whose expectation is:

$$E[\phi_m] = \sum_{x,k} p(x, C_k)\phi_m(x, C_k)$$

Maximum entropy classification

Setting

- the **empirical expectation** of ϕ_m is:

$$\hat{E}[\phi_m] = \sum_{x,k} \hat{p}(x, C_k)\phi_m(x, C_k),$$

where $\hat{p}$ is a histogram constructed by counting occurrences of the pair (x, C_k) in the training set:

$$\hat{p}(x, C_k) = \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \delta(x_i = x \wedge y_i = C_k).$$

- theoretical expectation $E[\phi_m]$ can be approximated as

$$\tilde{E}[\phi_m] = \sum_{x,k} \hat{p}(x)p(C_k|x)\phi_m(x, C_k)$$

Maximum entropy classification

Optimisation problem

- the **max entropy classifier** is the conditional probability $p^*(C_k|x)$ such that:

$$p^*(C_k|x) \doteq \arg \max_{p(C_k|x)} H_s(P),$$

where H_s is the traditional Shannon entropy, subject to:

$$\tilde{E}_p[\phi_m] = \hat{E}[\phi_m] \quad \forall m = 1, ..., M \tag{1}$$

- the classifier is to be consistent with the empirical frequencies of the features in the training set ...
- .. whilst seeking the least informative probability distribution that does so

Maximum entropy classification

Solution: the log-linear model

- the solution is the so-called **log-linear model**:

$$p^*(C_k|x) = \frac{1}{Z_\lambda(x)} e^{\sum_m \lambda_m \phi_m(x, C_k)},$$

where $\lambda = [\lambda_1, ..., \lambda_M]'$ are the Lagrange multipliers associated with the linear constraints (1) and $Z_\lambda(x)$ is a normalisation factor

- the related classification function is:

$$y(x) = \arg\max_k \sum_m \lambda_m \phi_m(x, C_k),$$

i.e., x is assigned the class which maximises the linear combination of the feature functions with coefficients λ

Belief functions in the credal interpretation

Lower and upper expectations

- belief functions are mathematically equivalent to a special class of credal set

$$\mathcal{P}[Bel] = \{P : P(A) \geq Bel(A)\}$$

- given a function $f : \Theta \to \mathbb{R}$, the *lower expectation* and *upper expectation* of f w.r.t. Bel are, respectively:

$$E_{Bel*}[f] \doteq \inf_{P \in \mathcal{P}[Bel]} E_P[f] = \sum_{A \subseteq \Theta} m(A) \inf_{x \in A} f(x),$$
$$E^*_{Bel}[f] \doteq \sup_{P \in \mathcal{P}[Bel]} E_P[f] = \sum_{A \subseteq \Theta} m(A) \sup_{x \in A} f(x)$$

Measures of entropy

for belief functions

- we wish to replace Shannon's entropy for conditional probabilities with a measure of entropy for belief functions
- the issue of how to assess the level of uncertainty associated with a belief function [Jirousek'16] is not trivial
- there are various proposals, including generalisations of classical entropy, specificity measures, composite and credal measures
- Jirousek and Shenoy [Jirousek'16] assessed all these proposals versus a number of significant properties, concluding that only the Maeda-Ichihashi proposal meets all these properties
- while the issue remains unsettled, here we merely adopt a straighforward generalisation of Shannon's entropy, and a few selected proposals based on their concavity property

Measures of entropy

Some significant proposals

- Nguyen $H_n[m] = -\sum_{A \in \mathcal{F}} m(A) \log m(A)$,
- Yager $H_y[m] = -\sum_{A \in \mathcal{F}} m(A) \log Pl(A)$.
- Dubois and Prade (a generalization of Hartley's entropy $H = \log(|\Theta|)$) $H_d[m] = \sum_{A \in \mathcal{F}} m(A) \log |A|$,
- Pal $H_a[m] = \sum_{A \in \mathcal{F}} m(A)/|A|$,
- Smets: $H_t = \sum_{A \in \mathcal{F}} \log(\frac{1}{Q(A)})$.
- **composite** measures: Klir and Ramer $H_k[m] = D[m] + H_d[m]$, where: $D(m) = -\sum_{A \in \mathcal{F}} m(A) \log[\sum_{B \in \mathcal{F}} m(B) \frac{|A \cap B|}{|B|}]$.
- **credal** measures: Harmanec $H_h[m] = \max_{P \in \mathcal{P}[Bel]}\{H_s[P]\}$.
- Maeda: $H_i[m] = H_h[m] + H_d[m]$

Generalised max entropy

Problem formulation

- to generalise the max-entropy optimisation problem we need to:
  1. choose an appropriate measure of entropy for belief function as o.f.
  2. revisit the constraints that the (theoretical) expectations of the feature maps are equal to the empirical ones computed over the training set
- sensible to require that the empirical expectation of the feature functions is bracketed by the lower and upper expectations associated with the sought belief function $Bel : 2^{\mathcal{X} \times \mathcal{C}} \rightarrow [0, 1]$
- here, we only make use of the 2-monotonicity of belief functions:

$$\sum_{(x, C_k)} Bel(x, C_k) \phi_m(x, C_k) \leq \hat{E}[\phi_m] \leq \sum_{(x, C_k)} Pl(x, C_k) \phi_m(x, C_k) \; \forall m$$

  i.e., we only consider probability intervals on $(x, C_k) \in \mathcal{X} \times \mathcal{C}$
- fully fledged lower and upper expectations, which express the full monotonicity of BFs, will be considered in future work

Maximum belief entropy classifier

- **maximum belief entropy classifier** is the joint belief measure $Bel^*(x, C_k) : 2^{X \times \mathcal{C}} \rightarrow [0, 1]$ which solves the optimisation problem:

$$Bel^*(x, C_k) \doteq \arg \max_{Bel(x, C_k)} H(Bel)$$

subject to the inequality constraints

$$\sum_{(x, C_k)} Bel(x, C_k) \phi_m(x, C_k) \leq \hat{E}[\phi_m] \leq \sum_{(x, C_k)} Pl(x, C_k) \phi_m(x, C_k) \; \forall m \tag{2}$$

where H is an appropriate measure of entropy for belief measures

- involves inequality constraints (2), as opposed to the equality constraints of traditional max entropy classifiers
- we need to analyse the Karush-Kuhn-Tucker (KKT) [Karush'39] necessary conditions for Bel to be an optimal solution

Karush-Kuhn-Tucker (KKT) conditions

in contrained optimisation

- suppose that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ of a nonlinear optimisation problem $\arg\max_x f(x)$ subject to:

$$g_i(x) \leq 0 \; i = 1, ..., m, \quad h_j(x) = 0 \; j = 1, ..., l$$

  are continuously differentiable at a point x^*
- if x^* is a local optimum, under appropriate regularity conditions then there exist constants μ_i, $(i = 1, \ldots, m)$ and λ_j $(j = 1, \ldots, l)$, called *KKT multipliers*, such that the following **KKT conditions** hold:
  1. *Stationarity*: $\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*)$;
  2. *Primal feasibility*: $g_i(x^*) \leq 0 \; \forall i = 1, \ldots, m$, and $h_j(x^*) = 0$, $\forall j$;
  3. *Dual feasibility*: $\mu_i \geq 0$ for all $i = 1, \ldots, m$;
  4. *Complementary slackness*: $\mu_i g_i(x^*) = 0$ for all $i = 1, \ldots, m$.

KKT conditions for the max belief entropy problem

- the KKT conditions are also sufficient whenever the objective function f is concave, the inequality constraints g_i are continuously differentiable convex functions, and the equality constraints h_j are affine
- more general sufficient conditions can be given in terms of *invexity* [Ben-Israel'86] requirements

Theorem 1

If either $H_t, H_n, H_d, H_s[Bel]$ or $H_s[Pl]$ is adopted as measure of entropy, the generalised max entropy optimisation problem has concave objective function and convex constraints. Therefore, the KKT conditions are sufficient for the optimality of its solution(s)

KKT conditions for Shannon-like entropy

- the KKT stationarity conditions for the generalised, belief-theoretical maximum entropy problem amount to, for all $\overline{B} \subset X \times \mathcal{C}$:

$$\begin{cases} -\sum_{A \supseteq \overline{B}} [1 + \log Bel(A)] = \sum_{m=1}^{M} \phi_m(\overline{x}, \overline{C}_k)[\mu_m^1 - \mu_m^2], |\overline{B}| = 1, \\ -\sum_{A \supseteq \overline{B}} [1 + \log Bel(A)] = \sum_{m=1}^{M} \mu_m^2 \sum_{(x, C_k) \in \overline{B}} \phi_m(x, C_k), |\overline{B}| > 1. \end{cases} \tag{3}$$

- The other conditions are, $\forall m = 1, ..., M$, (2) (primal feasibility), $\mu_m^1, \mu_m^2 \geq 0$ (dual feasibility), and complementary slackness:

$$\mu_m^1 \sum_{(x, C_k) \in \Theta} Bel(x, C_k)\phi_m(x, C_k) - \hat{E}[\phi_m] = 0,$$
$$\mu_m^2 \sum_{(x, C_k) \in \Theta} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)] = 0$$

Conclusions

- we proposed a generalisation of the max entropy classifier entropy
- the assumptions that test and training data are sampled by a same probability distribution, and that the empirical expectation of the feature functions is 'correct' are relaxed in the formalism of belief theory
- we studied the conditions under which the associated KKT conditions are necessary and sufficient for the optimality of the solution

What's next

- much interesting work remains to be done:
  1. providing analytical model expressions, similar to log-linear models, for the Shannon-like and other major entropy measures for belief functions
  2. analysing the case of full lower and upper expectations
  3. comparing the resulting classifiers
  4. analysing a formulation based on the least commitment principle, rather than max entropy, for the objective function to optimise
  5. finally, relaxing the constraint that feature functions be defined on singleton pairs (x, C_k)
- should constraints of the form (2) be enforced on all possible subsets $A \subset X \times \mathcal{C}$, rather than just singleton pairs (x, C_k)?
- basic question: what information does a training set actually carry?
- more general constraints would require extending the domain of feature functions to set values