MLE Voting Rules via Bregman Divergence

June 5, 2013

1 Exponential Families

To begin we have

- $\mathcal{X}$ is the set of outcomes (in general, can be discrete or continuous).
- ν is a base measure over the outcomes (usually counting or Lebesgue).
- $\phi : \mathcal{X} \to \mathcal{R}^k$ is a "sufficient statistic"—a vector-valued random variable.

An *exponential family* is a collection of probability densities over $\mathcal{X}$, with respect to ν, that take the form

$$p(x;\theta) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}. \tag{1}$$

Here $\theta \in \mathcal{R}^k$ is the *natural parameter*. The *cumulant* function is a normalizer:

$$A(\theta) = \log \int_{\mathcal{X}} \exp\{\langle \theta, \phi(x) \rangle\} \, d\nu(x). \tag{2}$$

The domain of A is $\Theta = \{\theta \in \mathcal{R}^k : A(\theta) < +\infty\}$. Both A and its domain Θ are convex. It is well-known that exponential families correspond to maximum entropy distributions. The entropy of a density is

$$H(p) = \int_{\mathcal{X}} p(x) \log p(x) \, d\nu(x) \tag{3}$$

Now suppose we maximize entropy subject to the constraint that the average sufficient statistic is $\mu \in \mathcal{R}^k$:

$$\int_{\mathcal{X}} \phi(x) \, p(x) d\nu(x) = \mu. \tag{4}$$

Then letting $\theta \in \mathcal{R}^k$ be the Lagrange multiplier vector for constraint (4), the solution takes the form (1). This shows that an exponential family has two parametrizations: the *natural* parametrization in terms of θ, and the *mean* parametrization in terms of μ.

2 Bregman Divergence

Associated with the convex function A is its *convex conjugate*

$$A^*(\mu) = \sup_{\theta' \in \Theta} \{\langle \theta', \mu \rangle - A(\theta')\}, \tag{5}$$

which is also a convex function. The *Bregman divergence* associated with a differentiable convex function f is $D(y, z) = f(y) - f(z) - \langle \nabla f(z), y - z \rangle$. Note that this distance function is convex in its first argument (but not necessarily its second), non-negative, and is zero only if $y = z$ when f is strictly convex. It is not necessarily symmetric, but it does satisfy some generalized version of the triangle inequality.

Now suppose θ and μ and the natural and mean parameters for a density from an exponential family. Assuming some unrestrictive conditions on ϕ (that it is a *minimal* statistic), it turns out we have $\mu = \nabla A(\theta)$ and $\theta = \nabla A^*(\mu)$. In particular the maximum in (5) is achieved at $\theta' = \theta$. Therefore, we have

$$\begin{aligned} \log p(x;\theta) &= \langle \theta, \phi(x) \rangle - A(\theta) \\ &= \langle \theta, \phi(x) - \mu \rangle + \langle \theta, \mu \rangle - A(\theta) \\ &= \langle \nabla A^*(\mu), \phi(x) - \mu \rangle + A^*(\mu) \\ &= -D_{A^*}(\phi(x), \mu) + A^*(\phi(x)) \end{aligned}$$

Therefore, the mean parametrization of the density takes the form

$$p(x;\mu) = \exp\{-D_{A^*}(\phi(x), \mu) + A^*(\phi(x)).\} \tag{6}$$

A voting rule according to this kind of parametrized distribution would proceed by first finding the MLE of μ (or θ) given the agents' reported rankings, and then computing the mode of the associated distribution.

3 MLE Voting

Recall the Mallows model for the distribution over permutations given a mode permutation π_0 and scalar parameter θ.

$$P(\pi;\theta,\pi_0) = \exp\{-\theta d(\pi, \pi_0)\}/\psi(\theta, \pi_0). \tag{7}$$

Here d is a distance function, which should satisfy $d(\pi, \sigma) \geq 0$ with equality if and only if $\pi = \sigma$. The remaining properties assumed of d vary in the literature. Sometimes it satisfies symmetry and the triangle inequality (so that it becomes a metric), but not always. Given this distribution, votes from agents are treated as sample points and the associated voting rule selects the pi_0 that is an MLE with respect to the sample.

There are many similarities in comparing (7) with (1) and (6) but also important differences. In (7) θ is a scalar parameter that affects the "dispersion" of the distribution, and it should not be confused with the vector natural parameter θ in (1). (It is possible to generalize exponential families by adding a dispersion parameter, but I'm not sure it's worth looking into because there is already a lot of flexibility with just the natural/mean parameter.) The functions $\psi(\theta, \pi_0)$ and $A(\theta)$ both serve as normalizers and are basically equivalent up to a log transformation.

A key difference is that (7) has a finite, discrete parameter set (the set of permutations) while (1) and (6) have convex parameter sets. This is crucial because for an exponential family, the MLE of the μ parameter given a set of samples $x_1, \ldots, x_n$ is just $\hat{\mu} = \sum_{i=1}^{n} \phi(x_i)$. Deriving the natural parameter from the mean parameter is a nontrivial computational problem. However, the natural parametrization is convenient for then computing the mode, because this reduces to the linear programming problem of maximizing $\langle \theta, y \rangle$ where $y \in \mathcal{M} = \text{co}\{\phi(x) : x \in X\}$, where here X is the set of permutations of the candidates. There is a very large amount of work in the machine learning literature on graphical models on methods to efficiently compute the natural parametrization from the mean parametrization, and vice-versa, which we could draw on the implement our version of MLE voting with exponential families and Bregman divergences (i.e., compute MLE and then them mode).

Another important differences is that (1) and (6) are both defined with respect to a certain representation $\phi(x)$ of rankings in Euclidean space $\mathcal{R}^k$, while (7) is defined directly on rankings. The exponential family approach opens the door to associating properties of the voting rule with the structure of ϕ.

4 Research Questions

1. Distance functions between rankings usually satisfy the natural property of right-invariance. What kind of Bregman divergences satisfy right-invariance? How does the property restrict the structure of ϕ?

2. How does the IIA condition restrict the structure of ϕ? Does it imply ϕ can only provide pairwise sufficient statistics?

3. If we can characterize the structure of the ϕ given IIA, can we get a simple proof of Arrow's theorem for our more restricted class of voting rules?

4. A positional voting rule induces a natural ϕ mapping into R^k, where k is the number of candidates. Does our MLE voting rule approach result in the associated positional voting rule this way?

A Generalized Entropy

The concepts above were specifically linked to the entropy function (3), but they can be generalized to any convex function over the set of probability densities, which is then construed as a generalized (negative) entropy function.

Let $\mathcal{P}$ denote the set of probability densities over $\mathcal{X}$ (with respect to base measure ν). Let H be a strictly convex function over the convex domain $\mathcal{P}$. To this convex function we can associate a *scoring rule* (not to be confused with position scores in a voting context) defined over $\mathcal{X} \times \mathcal{P}$:

$$S(x, p) = H(p) + \langle \nabla H(p), \mathbf{1}_x - p \rangle,$$

where $\mathbf{1}_x$ is the probability density that puts mass 1 on outcome x. It can be shown that if an agent is compensated with $S(x, q)$ when it reports $q \in \mathcal{P}$ and $x \in \mathcal{X}$ occurs, then its optimal strategy is to report its actual belief (i.e., the actual probability density it has in mind). Hence the term "scoring rule". For instance, if H is negative entropy then $S(x, p) = \log p(x)$.

To formulate the generalized exponential family, we then take the family of densities that satisfy

$$S(x, p) = \alpha + \langle \theta, \phi(x) \rangle. \tag{8}$$

Namely, there are $\alpha \in \mathcal{R}$ and $\theta \in \mathcal{R}^k$ so that the score $S(x, p)$ is a linear function of $\phi(x)$ as above. This gives the "natural" parametrization of p. Taking H as negative entropy, this condition becomes $\log p(x) = \alpha + \langle \theta, \phi(x) \rangle$, so that p takes the form of an exponential family.

The density p has an alternate characterization as the distribution that minimizes $H(p)$, subject to the constraint that the expectation of $\phi(x)$ under p is μ. This is the "mean" parametrization. More explicitly, let p_μ be the density with mean μ that minimizes $H(p)$, and define the convex function $F(\mu) = H(p_\mu)$ over $\mathcal{M} = \mathrm{co}\{\phi(x) : x \in \mathcal{X}\}$, where co denotes the convex hull. We take the family of densities that satisfy

$$S(x, p) = -D_F(\phi(x), \mu) + F(\phi(x)). \tag{9}$$

Note: still need to prove that (8) and (9) are equivalent characterizations. But I'm not sure that's true for any convex H.