Maximum Entropy Scoring Rules

Sébastien Lahaie

Microeconomics and Social Systems
Yahoo! Research

Problem of Prediction Workshop
December 2, 2011

Motivation

We would like to develop scoring rules to elicit an expert's opinion on certain properties (e.g., mean, variance) of the density over outcomes of a random experiment.

1. Systematic approach for generalizing classic scoring rules for probabilities to more general expectations.
2. A host of scoring rules for different outcome spaces (discrete, continuous).
3. Connections with prediction markets and variational inference methods in machine learning.

Disclaimers

- Large body of related work still being uncovered.
- Assume appropriate regularity conditions with infinite outcomes.

Motivation

We would like to develop scoring rules to elicit an expert's opinion on certain properties (e.g., mean, variance) of the density over outcomes of a random experiment.

1. Systematic approach for generalizing classic scoring rules for probabilities to more general expectations.
2. A host of scoring rules for different outcome spaces (discrete, continuous).
3. Connections with prediction markets and variational inference methods in machine learning.

Disclaimers

- Large body of related work still being uncovered.
- Assume appropriate regularity conditions with infinite outcomes.

Outline

1. Model
2. Scoring Rules
3. Prediction Markets
4. Discussion

Outline

1. Model
2. Scoring Rules
3. Prediction Markets
4. Discussion

Density Properties

- Outcome space $\mathcal{X}$ (discrete or continuous).
- Some σ-algebra (e.g., power set or Borel sets).
- Base measure ν (e.g., counting or Lebesgue).
- $\mathcal{P}$ denotes the set of densities p of distributions $P \ll \nu$.

Definition

Given a subset of densities $\mathcal{D} \subseteq \mathcal{P}$, a property map is a real (possibly vector-valued) function $\Gamma : \mathcal{D} \to \mathbf{R}^k$. Elements of its image $\mathcal{M} = \Gamma(\mathcal{D})$ are called properties and we say that p has property μ when $\Gamma(p) = \mu$.

If Γ is one-to-one then it is a parametrization and elements of its image are parameters.

Scoring Rules

The expert is asked to report $\mu = \Gamma(p)$, where $p \in \mathcal{D}$ is its belief, and is rewarded according to a scoring rule

$$S : \mathcal{M} \times \mathcal{X} \to [-\infty, +\infty)$$

which pays it $S(\hat{\mu}, x)$ if it reports $\hat{\mu}$ and outcome x occurs.

Definition

Given a property map $\Gamma : \mathcal{D} \to \mathbf{R}^k$, a scoring rule S is proper with respect to $\mathcal{D}$ if for all $\mu \in \mathcal{M}$ and $p \in \mathcal{D}$ with property μ, we have

$$\mathrm{E}_p[S(\mu, x)] > \mathrm{E}_p[S(\hat{\mu}, x)]$$

for all $\hat{\mu} \neq \mu$. If the domain is $\mathcal{D} = \mathcal{P}$, we simply say that the scoring rule is proper.

Informational Requirements

Observation

Propriety implies that the expert does not need an opinion of p to be incentivized to report $\mu = \Gamma(p)$; it only needs an opinion of μ.

- If proper w.r.t. $\mathcal{D}$, expert must agree that underlying $p \in \mathcal{D}$.
- If just proper, expert must only agree on support (recall the base measure ν).

[Lambert et al. '08; Lambert '11]

Maximum Likelihood

Lemma

Assume that Γ is a parametrization of $\mathcal{D} \subseteq \mathcal{P}$. The logarithmic scoring rule defined by

$$S(\mu, x) = \log p(x; \mu)$$

is proper w.r.t. $\mathcal{D}$ if and only if the maximum likelihood estimate for μ is consistent.

Proof sketch: Propriety and consistency are both equivalent to the condition that the true parameter μ maximizes $\mathrm{E}_p[\log p(x; \mu)]$.

Example: Pareto Distribution

- Pareto densities $f(x; \alpha) = \alpha / x^{\alpha+1}$, parameter $\alpha > 0$.
- Can be parametrized by the mean $\mu = \frac{\alpha}{\alpha-1}$.
- Support is $[1, +\infty)$.

This leads to the following scoring rule for the mean, proper w.r.t. this domain of densities:

$$S(\mu, x) = \log \frac{\mu}{\mu - 1} - \left(\frac{\mu}{\mu - 1} + 1 \right) \log x.$$

Example: Exponential Distribution

- Exponential densities $f(x; \lambda) = \lambda e^{-\lambda x}$ parameter $\lambda > 0$.
- Can be parametrized by the mean $\mu = 1/\lambda$.
- Support is $[0, +\infty)$.

This leads to the following scoring rule for the mean, proper w.r.t. this domain of densities:

$$S(\mu, x) = -\frac{x}{\mu} - \log \mu.$$

In fact, proper for the mean of any density with support in $\mathbf{R}_+$.

Example: Exponential Distribution

- Exponential densities $f(x; \lambda) = \lambda e^{-\lambda x}$ parameter $\lambda > 0$.
- Can be parametrized by the mean $\mu = 1/\lambda$.
- Support is $[0, +\infty)$.

This leads to the following scoring rule for the mean, proper w.r.t. this domain of densities:

$$S(\mu, x) = -\frac{x}{\mu} - \log \mu.$$

In fact, proper for the mean of any density with support in $\mathbf{R}_+$.

Outline

Expectation Properties

Let $\phi : \mathcal{X} \to \mathbf{R}^k$ be a "feature map" (a random variable), and let $A_\phi : \mathcal{P} \to \mathbf{R}^k$ be an associated linear operator:

$$A_\phi p = \mathrm{E}_p[\phi] = \int_{x \in \mathcal{X}} \phi(x) p(x) \, d\nu(x).$$

Expectations

We consider properties of the form

$$\Gamma(p) = g(A_\phi p)$$

where $g : \mathbf{R}^k \to \mathbf{R}^k$ is a bijection.

Example: Mean and Variance

- $\mathcal{X} = \mathbf{R}$, $\nu =$ Lebesgue measure.
- $\phi(x) = (x, x^2)$.

$$A_\phi p = (\mu_1, \mu_2) = (\mathrm{E}_p[x], \mathrm{E}_p[x^2])$$

The bijection g recovers the mean and variance by setting

$$\begin{aligned} \mu &\leftarrow \mu_1 \\ \sigma^2 &\leftarrow \mu_2 - \mu_1^2 \end{aligned}$$

Maximum Entropy Rule

Let $F : \mathcal{P} \to \mathbf{R}$ be a strictly convex function over densities.
Given expert report μ, formulate the following optimization:

$$G(\mu) = \max \{-F(p) \; : \; A_\phi p = \mu\}\,.$$

- $F^*(q) = \sup_{p \in \mathcal{P}} \{\langle q, p \rangle - F(p)\}$, convex conjugate.
- A_ϕ^* is adjoint of A_ϕ (like transpose).
- $\theta(\mu) \in \mathbf{R}^k$ is Lagrange multiplier for equality constraints.

Theorem

The following scoring rule is proper for eliciting $\mu = \mathrm{E}_p[\phi]$:

$$S(\mu, x) = \langle \theta(\mu), \phi(x) \rangle - F^*(A_\phi^* \theta(\mu))$$

Proof Sketch

Assume that the expert believes $\mathsf{E}_p[\phi] = \mu$ and let $\mu' \neq \mu$.

$$\begin{aligned} & \mathsf{E}_p[S(\mu, x)] - \mathsf{E}_p[S(\mu', x)] \\ = \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle \theta(\mu') - \theta(\mu), \mu \rangle \right] \end{aligned}$$

Proof Sketch

Assume that the expert believes $\mathrm{E}_p[\phi] = \mu$ and let $\mu' \neq \mu$.

$$\begin{aligned} & \mathrm{E}_p[S(\mu, x)] - \mathrm{E}_p[S(\mu', x)] \\ = \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle \theta(\mu') - \theta(\mu), A_\phi p \rangle \right] \end{aligned}$$

Proof Sketch

Assume that the expert believes $\mathrm{E}_p[\phi] = \mu$ and let $\mu' \neq \mu$.

$$
\begin{aligned}
& \mathrm{E}_p[S(\mu, x)] - \mathrm{E}_p[S(\mu', x)] \\
= \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle \theta(\mu') - \theta(\mu), A_\phi p^* \rangle \right]
\end{aligned}
$$

Proof Sketch

Assume that the expert believes $\mathsf{E}_p[\phi] = \mu$ and let $\mu' \neq \mu$.

$$\begin{aligned}
& \mathsf{E}_p[S(\mu, x)] - \mathsf{E}_p[S(\mu', x)] \\
= \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle \theta(\mu') - \theta(\mu), A_\phi p^* \rangle\right] \\
= \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle A_\phi^* \theta(\mu') - A_\phi^* \theta(\mu), p^* \rangle\right]
\end{aligned}$$

Proof Sketch

Assume that the expert believes $\mathrm{E}_p[\phi] = \mu$ and let $\mu' \neq \mu$.

$$\begin{aligned}
& \mathrm{E}_p[S(\mu, x)] - \mathrm{E}_p[S(\mu', x)] \\
= \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle \theta(\mu') - \theta(\mu), A_\phi p^* \rangle\right] \\
= \quad & F^*(A_\phi^* \theta(\mu')) - \left[F^*(A_\phi^* \theta(\mu)) + \langle A_\phi^* \theta(\mu') - A_\phi^* \theta(\mu), p^* \rangle\right] \\
> \quad & 0
\end{aligned}$$

Inequality follows from the fact that $\nabla F^*(A_\phi^* \theta(\mu)) = p^*$.

Example: Eliciting Probabilities

$$\mathcal{X} = \{1, 2, 3\} \quad \phi(x) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathrm{E}_p[\phi] = \begin{bmatrix} p(1) \\ p(2) \\ p(3) \end{bmatrix}$$

Score	$F(p)$	$S(p, x)$
Logarithmic	$-$entropy of p	$\log\langle p, \phi(x)\rangle$
Brier	squared norm of p	$\langle p, \phi(x)\rangle - \frac{1}{2}\lvert\lvert p\rvert\rvert^2$
Spherical	norm of p	$\langle p, \phi(x)\rangle \,/\, \lvert\lvert p\rvert\rvert$

Information function F recovered via $F(p) = \mathrm{E}_p[S(p, x)]$.
[Dawid '98; Grünwald & Dawid '04; Gneiting & Raftery '07]

Example: Eliciting Probabilities

$$\mathcal{X} = \{1,2,3\} \quad \phi(x) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathrm{E}_p[\phi] = \begin{bmatrix} p(1) \\ p(2) \\ p(3) \end{bmatrix}$$

Score	$F(p)$	$S(p,x)$
Logarithmic	$-$entropy of p	$\log\langle p, \phi(x)\rangle$
Brier	squared norm of p	$\langle p, \phi(x)\rangle - \frac{1}{2}\|\|p\|\|^2$
Spherical	norm of p	$\langle p, \phi(x)\rangle \,/\, \|\|p\|\|$

Information function F recovered via $F(p) = \mathrm{E}_p[S(p,x)]$.
[Dawid '98; Grünwald & Dawid '04; Gneiting & Raftery '07]

Example: Eliciting Mean

$$\phi(x) = x$$

$$F = -\text{entropy}$$

$\mathcal{X}$	Distribution	$S(\mu, x)$
$\{1, 2, \ldots\}$	Poisson	$x \log \mu - \mu$
$\mathbf{R}_+$	Exponential	$-\frac{x}{\mu} - \log \mu$
$\mathbf{R}$	Normal	$\mu x - \frac{1}{2}\mu^2$

Example: Eliciting Mean and Variance

$$\phi(x) = (x, x^2)$$

$$F = -\text{entropy}$$

$\mathcal{X}$	Distribution	$S((\text{mean}, \text{variance}), x)$
R	Normal	$-\frac{(x-\mu)^2}{\sigma^2} - \log \sigma^2$
$\mathbf{R}^k$	Multivariate Normal	$-(x-\mu)'\Sigma^{-1}(x-\mu) - \log\det\Sigma$

[Dawid '98; Dawid & Sebastiani '99]

Other Expectation Properties

- harmonic mean, geometric mean (closed form)
- skewness, kurtosis (variational formulation)

Open Question

What is the smallest number of expectation properties needed to elicit a general property?

Answer has implications for the *elicitation complexity* of a property. [Lambert et al. '08]

Outline

Savage Representation

Let $C = F^* \cdot A^*_\phi$ be a convex function over $\Theta = \mathcal{M}^*$ (dual of feasible properties), and let $G = C^*$. We get the following representation for the scoring rule.

$$\begin{aligned} S(\mu, x) &= \langle \theta(\mu), \phi(x) \rangle - F^*(A^*_\phi \theta(\mu)) \\ &= \langle \theta(\mu), \phi(x) \rangle - C(\theta(\mu)) \\ &= \langle \nabla C^{-1}(\mu), \phi(x) \rangle - C(\nabla C^{-1}(\mu)) \\ &= G(\mu) - \langle \nabla G(\mu), \mu - \phi(x) \rangle. \end{aligned}$$

[Savage '71; Gneiting & Raftery '07]

Savage Representation

Let $C = F^* \cdot A_\phi^*$ be a convex function over $\Theta = \mathcal{M}^*$ (dual of feasible properties), and let $G = C^*$. We get the following representation for the scoring rule.

$$
\begin{aligned}
S(\mu, x) &= \langle \theta(\mu), \phi(x) \rangle - F^*(A_\phi^* \theta(\mu)) \\
&= \langle \theta(\mu), \phi(x) \rangle - C(\theta(\mu)) \\
&= \langle \nabla C^{-1}(\mu), \phi(x) \rangle - C(\nabla C^{-1}(\mu)) \\
&= G(\mu) - \langle \nabla G(\mu), \mu - \phi(x) \rangle.
\end{aligned}
$$

[Savage '71; Gneiting & Raftery '07]

Cost Function Market Making

1. There are k securities and $\phi(x) \in \mathbf{R}^k$ gives the payoffs for each if outcome $x \in \mathcal{X}$ occurs. If an agent holds a share vector $\theta \in \mathbf{R}^k$ and x occurs, the payoff is

$$\langle \theta, \phi(x) \rangle.$$

2. The market maker commits to a cost function C such that buying shares θ costs $C(\theta)$. If $x \in \mathcal{X}$ occurs, the agent's profit is

$$\langle \theta, \phi(x) \rangle - C(\theta).$$

3. An agent with opinion μ will acquire shares $\theta(\mu)$ and its expected profit will be $\mathrm{E}[S(\mu, x)]$.

[Chen & Wortman '10]

Concept Mapping

Concept	Machine Learning	Prediction Markets
$\mathcal{X}$	states	outcomes
μ	mean parameter	expert opinion
θ	natural parameter	share vector
ϕ	sufficient statistic	payoff function
C	log partition function	cost function

- Share vector in an LMSR market is the natural parameter of exponential family with sufficient statistic ϕ.
- Variational inference methods for C can be used to run prediction markets [Wainwright & Jordan '08].
- Other cost functions C can be applied [Abernethy et al. '11].

Concept Mapping

Concept	Machine Learning	Prediction Markets
$\mathcal{X}$	states	outcomes
μ	mean parameter	expert opinion
θ	natural parameter	share vector
ϕ	sufficient statistic	payoff function
C	log partition function	cost function

- Share vector in an LMSR market is the natural parameter of exponential family with sufficient statistic ϕ.
- Variational inference methods for C can be used to run prediction markets [Wainwright & Jordan '08].
- Other cost functions C can be applied [Abernethy et al. '11].

Outline

Extending Properties

We considered property maps Γ such that $\Gamma(p) = \mu$ if p is a solution to

$$\mathsf{E}_p[\phi(x)] = A_\phi p = \mu.$$

- Allows one to capture boundaries of support $[a, b]$.
- Allows one to capture median, quantiles.

Open Question

Is there a variational framework for eliciting the median of a density?

Extending Properties

We considered property maps Γ such that $\Gamma(p) = \mu$ if p is a solution to

$$\mathsf{E}_p[\phi(x)] = A_\phi p \leq \mu.$$

- Allows one to capture boundaries of support $[a, b]$.
- Allows one to capture median, quantiles.

Open Question

Is there a variational framework for eliciting the median of a density?

Extending Properties

We considered property maps Γ such that $\Gamma(p) = \mu$ if p is a solution to

$$\mathrm{E}_p[\phi(x)] = A_\phi p \leq \mu.$$

- Allows one to capture boundaries of support $[a, b]$.
- Allows one to capture median, quantiles.

Open Question

Is there a variational framework for eliciting the median of a density?

Extending Properties

We considered property maps Γ such that $\Gamma(p) = \mu$ if p is a solution to

$$\mathrm{E}_p[\phi(x)] = A_\phi p \leq \mu.$$

- Allows one to capture boundaries of support $[a, b]$.
- Allows one to capture median, quantiles.

Open Question

Is there a variational framework for eliciting the median of a density?

Nonparametric Elicitation

- We saw elicitation analogs (scoring rules, prediction markets) for parametric statistics.
- We could take a nonparametric approach: let there be a security for every outcome $x \in \mathcal{X}$.
- Define a kernel function $k(x, x)$ over $\mathcal{X}$ (a similarity measure)—for each kernel function there is a mapping ϕ such that

$$k(x, x) = \langle \phi(x), \phi(x) \rangle.$$

- If the agent holds α_1 of security x_1, α_2 of security x_2... the payoff becomes

$$\alpha_1 k(x_1, x) + \alpha_2 k(x_2, x) + \dots$$

Conclusions

- Analogs of classical scoring rules for expectation properties.
- New scoring rules for wide variety of outcome spaces.
- Connections with variational inference in machine learning.

- Implications for elicitation complexity.
- Variational framework for eliciting quantiles.
- Nonparametric scoring rules / prediction markets.