Summary about Prediction Market

Zhihao Ruan, Jiayu Yi, Naihao Deng
{ruanzh,yijiayu,dnaihao}@umich.edu

August 5, 2019

Contents

1 Exponential Family

Distributions over $\mathbf{x}$ with *natural parameters* $\boldsymbol{\eta}$ that are of the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})) \tag{1}$$

We need to make sure that $\int p(\mathbf{x}|\boldsymbol{\eta})d\boldsymbol{x} = 1$, therefore

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}d\mathbf{x} = 1 \tag{2}$$

$g(\boldsymbol{\eta})$ is a coefficient that is needed to ensure that the distribution is normalized.

1.1 Sufficient Statistics

If we take the partial derivative with respect to $\boldsymbol{\eta}$ from both sides of equation 2, we have

$$\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}d\mathbf{x} + g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}\mathbf{u}(\mathbf{x})d\mathbf{x} = 0 \tag{3}$$

If we plug in 2 to 3, we have

$$-\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right\}\mathbf{u}(\mathbf{x})d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{4}$$

Since

$$\nabla \ln g(\boldsymbol{\eta}) = \frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta})$$

It can be deduced that

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{5}$$

If we consider a set of independent identically distributed data denoted by $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the likelihood function is given by

$$p(\mathbf{x}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\} \tag{6}$$

To produce the maximum likelihood estimator $\boldsymbol{\eta}_{ML}$, we need to take the partial derivative of 6 and set its value to 0.

$$\nabla\left(\left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\}\right) = 0 \tag{7}$$

By expanding 7 it can be seen that

$$\left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right)\left(Ng(\boldsymbol{\eta})^{N-1}\nabla g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\} + g(\boldsymbol{\eta})^N\exp\left\{\boldsymbol{\eta}^T\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\}\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right) = 0$$

Therefore,

$$-\nabla \ln g(\boldsymbol{\eta}_{ML}) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n) \tag{8}$$

From equation 8 it can be seen that $\boldsymbol{\eta}_{ML}$ can be determined if $\mathbf{u}(\mathbf{x}_n)$ is known, $\mathbf{u}(\mathbf{x}_n)$ is therefore called the *sufficient statistics*. If $N \to \infty$, the right hand side of 8 would become $\mathbb{E}[\mathbf{u}(\mathbf{x})]$ and $\boldsymbol{\eta}_{ML}$ would be the true $\boldsymbol{\eta}$.

1.2 Conjugate Priors

Conjugate priors are priors that lead to their corresponding posterior distributions to have the same form as they do. For exponential family distributions, their conjugate priors are of the form

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\} \tag{9}$$

ν can be interpreted as the number of samples that are observed. χ is the pseudo observation of the sufficient statistics $\mathbf{u}(\mathbf{x})$. Each one of the ν observations take the value of $\mathbf{u}(\mathbf{x})$. The updated posterior can be shown as

$$\begin{aligned} p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) &= \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^{N}\exp\left\{\boldsymbol{\eta}^{T}\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\} \times f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\} \\ &\propto g(\boldsymbol{\eta})^{\nu+N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{u}\left(\mathbf{x}_n\right)+\nu\boldsymbol{\chi}\right)\right\} \end{aligned} \tag{10}$$

It can been seen from here that we can update the posterior with $\nu \leftarrow \nu+N$ and $\nu\boldsymbol{\chi} \leftarrow \sum_{n=1}^{N}\mathbf{u}\left(\mathbf{x}_n\right)+\nu\boldsymbol{\chi}$

1.2.1 Bernoulli Distribution

The conjugate prior for Bernoulli distribution is the Beta distribution. The probability density function for Beta distribution is

$$p(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{\mathbf{B}(a,b)} \tag{11}$$

When it is written in the form of the conjugate prior

$$\frac{\theta^{a-1}(1-\theta)^{b-1}}{\mathbf{B}(a,b)} = \frac{1}{\mathbf{B}(a,b)}(1-\theta)^{a+b-2}\exp\left[(a+b-2)\ln\left(\frac{\theta}{1-\theta}\right)\frac{a-1}{a+b-2}\right] \tag{12}$$

From equation 12 we can see that ν corresponds to $a+b-2$ here and $\boldsymbol{\chi}$ corresponds to $\frac{a-1}{a+b-2}$.

1.2.2 Normal Distribution

Given that the conjugate priors for exponential family distributions are of the form $p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}$, we were trying to find the ν and χ for univariate Gaussian distribution given that the variance σ^2 is fixed.

The probability density function of univariate Gaussian distribution is given by

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{13}$$

First we can start by proving that the conjugate prior for Gaussian distribution is also a Gaussian distribution (when the variance is fixed).

$$\begin{aligned} p(x|\mu) &\sim N\left(\mu,\sigma^2\right) \\ p(\mu) &\sim N\left(\mu_0,\sigma_0^2\right) \end{aligned}$$

Suppose D is the set of data points that we have observed. We have

$$\begin{aligned} p(\mu|\mathcal{D}) &\propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\right\} \times \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\sum_{i=1}^{N} x^{(i)}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\} \end{aligned} \tag{14}$$

We can see that equation 14 can be easily written in the form of a Gaussian distribution if we extract multiple from the coefficient in the front.

Can all this be still written in the form of exponential family distribution and exponential family distribution's conjugate priors?

We want to write the prior

$$\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$$

in the form of $p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu}\exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}$, and the posterior

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\right\}$$

in the form of $p(\mathbf{x}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}$.

We can first start with priors.

$$\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\} = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{\mu^2 - 2\mu_0\mu + {\mu_0}^2}{2{\sigma_0}^2}\right\} \tag{15}$$

From the equation above we can deduce that the $\boldsymbol{\eta}$ should be $\begin{bmatrix}\mu \\ \mu^2\end{bmatrix}$. Is this OK? Continue with Eq. 15,

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{\mu^2 - 2\mu_0\mu + {\mu_0}^2}{2{\sigma_0}^2}\right\} &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{\boldsymbol{\eta}^T \begin{bmatrix}\frac{\mu_0}{\sigma_0^2} \\ -\frac{1}{2\sigma_0^2}\end{bmatrix} - \frac{{\mu_0}^2}{2{\sigma_0}^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(\frac{-{\mu_0}^2}{2{\sigma_0}^2}\right) \exp\left[\boldsymbol{\eta}^T \begin{bmatrix}\frac{\mu_0}{\sigma_0^2} \\ -\frac{1}{2\sigma_0^2}\end{bmatrix}\right] \end{aligned} \tag{16}$$

We know from equation 16 that we need μ outside $\exp$ to construct $g(\boldsymbol{\eta})$.

If we create two helper parameters α, β to help us split the coefficients of μ and μ^2, we'll have

$$\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(\frac{-{\mu_0}^2}{2{\sigma_0}^2}\right) \exp\left[\boldsymbol{\eta}^T \begin{bmatrix}\frac{\mu_0}{\sigma_0^2} \\ -\frac{1}{2\sigma_0^2}\end{bmatrix}\right] = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[\frac{\alpha}{{\sigma_0}^2}\mu^2 + \frac{\beta\mu_0}{{\sigma_0}^2}\mu\right] \left\{-\frac{(1+2\alpha)\mu^2 - 2\mu_0(1-\beta)\mu + {\mu_0}^2}{2{\sigma_0}^2}\right\} \tag{17}$$

We can assume that ν here is the common denominator of $\frac{\alpha}{{\sigma_0}^2}$ and $\frac{\beta\mu_0}{{\sigma_0}^2}$, which is $\frac{1}{{\sigma_0}^2}$. Under this assumption

$$g(\boldsymbol{\eta}) = \exp(\alpha\boldsymbol{\eta}_2 + \beta\mu_0\boldsymbol{\eta}_1).$$

Unfortunately, there's so far no way for us to determine the values of α and β.

When it comes to the posterior, we first have to tidy up the original expression and get

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\right\} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\left(x^{(i)}\right)^2 - 2\mu\sum_{i=1}^{N}x^{(i)} + N\mu^2\right]\right\}.$$

2 Scoring Rules

Scoring rules is the simplest form of prediction mechanism. For every agent with some information, a scoring rule evaluates how close the information is from the actual outcome, and pays the agent in return for his/her information.

2.1 Motivations

Taking an event $\mathbf{X}$ with outcome space $\mathcal{X}$, we want to know something about each agent's belief $p(\mathbf{x})$ on the actual outcome $\mathbf{x} \in \mathcal{X}$, compare it with the actual outcome, and see how accurate the agent's prediction is. However, it is impossible to ask agent for his/her entire $p(\mathbf{x})$ probability distribution. What should we do?

We've already known that sufficient statistic is a very suitable parameter to characterize a probability distribution. Therefore, we can use it to represent agent's belief. We just ask each agent for the sufficient statistic $\mathbf{u}(\mathbf{x})$ as a representation of his/her belief $p(\mathbf{x})$. Then we are able to measure how accurate the agent can predict with a scoring rule based on the actual outcome.

Assume $\boldsymbol{\mu} = \mathbb{E}_p[\mathbf{u}(\mathbf{x})]$ is the *expected* sufficient statistic over some $p(\mathbf{x})$ that takes $\mathbf{u}(\mathbf{x})$ as its sufficient statistic and is taken as an estimate of the agent's belief. Assume $\hat{\boldsymbol{\mu}}$ is the agent's report as an estimate of $\boldsymbol{\mu}$. Then, with the actual outcome denoted $\mathbf{x}$, the scoring rule takes the form:

$$S(\hat{\boldsymbol{\mu}}, \mathbf{x}).$$

We can see that the scoring rule is a measure of how close the agent's belief is from the actual outcome.

2.2 Incentive Compatibility

Incentive compatibility is a property of prediction mechanism with which the best strategy for an agent to earn the most profit is to *honestly* report all the information as soon as he/she has it. As a prediction mechanism, a proper scoring rule should leverage **incentive compatibility** in order to get real information from agents.

Assume that we already set the sufficient statistic to be $\mathbf{u}(\mathbf{x})$. Then for each $p \in \mathcal{P}$ that takes this $\mathbf{u}(\mathbf{x})$ as its sufficient statistic, we can calculate its *expected* sufficient statistic $\boldsymbol{\mu} = \mathbb{E}_p[\mathbf{u}(\mathbf{x})]$. A scoring rule is thus called **proper** if it satisfies the following, for all such p, for all $\hat{\boldsymbol{\mu}} \neq \boldsymbol{\mu}$:

$$\mathbb{E}_p[S(\boldsymbol{\mu}, \mathbf{x})] \geqslant \mathbb{E}_p[S(\hat{\boldsymbol{\mu}}, \mathbf{x})].$$

Any scoring rule with this property actually encourages agents to report a probability distribution as close to the actual probability distribution of $\mathbf{X}$ as possible, which is in accordance with the essence of incentive compatibility.

2.3 Logarithmic Scoring Rule

Suppose that we have set a form of $\mathbf{u}(\mathbf{x})$. For some p that takes $\mathbf{u}(\mathbf{x})$ as its sufficient statistic, a classic logarithmic scoring rule takes the form:

$$S(\boldsymbol{\mu}, \mathbf{x}) = \ln p(\mathbf{x}; \boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = \mathbb{E}_p[\mathbf{u}(\mathbf{x})]$ is the expected sufficient statistic over $p(\mathbf{x}; \boldsymbol{\mu})$. **For the following sections, we will be only talking about logarithmic scoring rules.**

2.4 Maximum Entropy Optimization

Now that we have the general expression of logarithmic scoring rules, how can we determine which $p(\mathbf{x};\boldsymbol{\mu})$ to choose as an estimate of agents' beliefs, given a specific $\mathbf{u}(\mathbf{x})$?

For some particular form of $\mathbf{u}(\mathbf{x})$, denote the set $\mathcal{P}$ which contains all p that takes $\mathbf{u}(\mathbf{x})$ as its sufficient statistic. Then, we formulate the following optimization problem:

$$\begin{aligned}\min_{p\in\mathcal{P}} \quad & F(p) = \int_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}h(\mathbf{x}) \\ \text{s.t.} \quad & \mathbb{E}_p[\mathbf{u}(\mathbf{x})] = \boldsymbol{\mu}\end{aligned}$$

where $\boldsymbol{\mu}$ is the expected sufficient statistic over p, and $h(\mathbf{x})$ is the base measure.

It is clear to see that $F(p)$ is actually the negative of entropy of $p(\mathbf{x})$. Therefore, minimizing $F(p)$ is thus maximizing the entropy of p. This actually mean that of all $p \in \mathcal{P}$, we tend to choose a p that shows the most uncertainty of $\mathbf{X}$ as an estimate of agents' beliefs. This estimate is not necessarily exactly the same as agents' beliefs, as the general goal is just to find something that we can use to construct a metric in order to measure how close the agents' beliefs are from the actual outcome.

It has been proved that $S(\boldsymbol{\mu}, \mathbf{x}) = \ln p(\mathbf{x};\boldsymbol{\mu})$ is **proper** if and only if p is the exponential family. It has also been proved that the solutions to the optimization problem are exponential family distributions which takes the form

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\eta}) &= h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}\cdot\mathbf{u}(\mathbf{x})\} \\ &= h(\mathbf{x}) \exp\{\boldsymbol{\eta}\cdot\mathbf{u}(\mathbf{x}) - T(\boldsymbol{\eta})\}\end{aligned}$$

where $h(\mathbf{x})$ is the base measure, $\boldsymbol{\eta}$ is the natural parameter of p, $T(\boldsymbol{\eta}) = -\ln g(\boldsymbol{\eta})$ is the log-partition function of p.

Either *expected* sufficient statistic $\boldsymbol{\mu}$ or natural parameter $\boldsymbol{\eta}$ can parametrize a probability distribution p, and actually they are closely related. If we denote the convex conjugate of $T(\boldsymbol{\eta})$ as $G(\boldsymbol{\mu})$, then we have $\boldsymbol{\mu} = \nabla T(\boldsymbol{\eta}), \boldsymbol{\eta} = \nabla G(\boldsymbol{\mu})$.

2.4.1 Examples

1. If we take the outcome space $\mathcal{X} = [0, +\infty)$ and $u(x) = x$, the maximum entropy optimization produces an exponential distribution. Hence, the log scoring rule becomes

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

   If we take the partial derivative with respect to μ

   $$\frac{\partial S}{\partial \mu} = \frac{x}{\mu^2} - \frac{1}{\mu} = \frac{x-\mu}{\mu^2},$$

   we can see that the scoring rule reaches maximum when $\mu = x$. This actually means that agents should report information as close to the actual outcome as possible in order to maximize profit, which corresponds to the essence of incentive compatibility.

2. If we take the outcome space $\mathcal{X} = \mathbb{R}$ and $\mathbf{u}(x) = (x, x^2)$, the maximum entropy optimization produces a Gaussian distribution. Hence, the log scoring rule becomes

   $$S((\mu, \sigma^2), x) = -\frac{(x-\mu)^2}{\sigma^2} - \ln\sigma^2.$$

   As this scoring rule is a basic parabola with respect to μ, it is also clear that it reaches the maximum when $\mu = x$. Similarly, this scoring rule is also incentive compatible.

3 Cost Function based Prediction Market with Bayesian Traders

Based on what we have in previous sections, we could simulate the prediction market by a python program.

For the setup, there would be a group of Bayesian agents (the number of agents as hyper-parameter would be an input for the program), a market maker and an executable program driving the market. They are explained in Sec 3.1, Sec 3.2 and Sec 3.3. The dataset the market is based on is subject to Bernoulli distribution.

3.1 Bayesian Agents

Each Bayesian agents would be initialized with several data points, based on which he would form his prior $p(\boldsymbol{\eta}; \boldsymbol{\chi}, \nu)$. The dataset is subject to Bernoulli distribution, in which ν would be the number of data points Bayesian agents have access to initially, $\nu\boldsymbol{\chi}$ would be the sum of these data points.

Once the agent intends to enter the market, he would have access to several more data points before his entrance. He would update his belief by $b_1 = \begin{bmatrix} n\nu + m\hat{\mu} \\ n + m \end{bmatrix}$, that is, he would update the first entry of the vector as the sum of data he is now having access to, the second entry as the number of these data. He would also derive the amount of outstanding shares θ from the current market price, in our case, $\theta = -1/p$ (p is the current market price).

Based on

$$\nabla C(\theta + \delta) = \frac{n\nu + m\hat{\mu}}{n + m},$$

the agent would be able to decide δ, the amount he would like to trade by

$$\delta = -\frac{1}{\frac{n\nu + m\hat{\mu}}{n+m}} - \theta.$$

3.2 Market Maker

The market maker is initialized with the initial outstanding security amount θ and the initial market price, which are hyper-parameter predefined.

The market maker holds the sufficient statistic $\phi(x) = x$, cost function for the market $C(\theta) = -\log(-\theta)$ and price function $p = -1/\theta$. He would keep track of number of trades, outstanding security amount and current market price when the market runs.

3.3 Executable Program

The driving program takes agent number and max iteration number as hyper-parameters. The complete dataset is divided into two parts as random chosen data in the first part would be assigned to agents to form their priors, data in the second part would be fed to agents when they intend to enter the market.

Each agent comes into the market iteratively with the total number of iteration not exceeding max iteration number in program arguments. The interval between each agent's arrival is subject to Poisson distribution. As discussed in Sec 3.1 and Sec 3.2, each agent would update their belief and then trade with market maker. The final market price is considered as the aggregated belief among agents.