【算法】Prediction2

张楚珩
清华大学 交叉信息院博士在读

这一篇讲第二章：Prediction with Expert Advice。

原文传送门

Cesa-Bianchi, Nicolo, and Gabor Lugosi.*Prediction, learning, and games*. Cambridge university press, 2006.

由于是图书，就不放链接了，直接 google 就能搜得到 PDF。

特色

针对前一讲提到的 expert problem 的设定，针对一些略微不同的问题设定，给出了相应的策略并且证明了不同策略下的 regret bound。这些策略都是『综合考虑不同专家给出的意见』，数学上来说，即采用 weighted average 的方式作出预测。

过程

1. 问题设定

和上一讲的设定一样，只不过这里更形式化一点。定义 $\mathcal{Y}$ 为 outcome space， $\mathcal{D}$ 为 decision space， $\mathcal{E}$ 为专家的集合。每一轮 t 按照如下顺序进行：

- 1）专家给出建议 $f_{E,t}\in\mathcal{D},E\in\mathcal{E}$ ；
- 2）玩家（forecaster）根据专家的建议做出预测 $\hat{p}_t\in\mathcal{D}$ ；
- 3）环境给出结果 $y_t\in\mathcal{Y}$ ；
- 4）玩家遭受损失， $l(\hat{p}_t,y_t)$ ，其中损失函数 $l:\mathcal{D}\times\mathcal{Y}\to\mathbb{R}$ 。

PREDICTION WITH EXPERT ADVICE

Parameters: decision space $\mathcal{D}$, outcome space $\mathcal{Y}$, loss function ℓ, set $\mathcal{E}$ of expert indices.

For each round $t = 1, 2, \ldots$

(1) the environment chooses the next outcome y_t and the expert advice $\{f_{E,t} \in \mathcal{D} : E \in \mathcal{E}\}$; the expert advice is revealed to the forecaster;
(2) the forecaster chooses the prediction $\widehat{p}_t \in \mathcal{D}$;
(3) the environment reveals the next outcome $y_t \in \mathcal{Y}$;
(4) the forecaster incurs loss $\ell(\widehat{p}_t, y_t)$ and each expert E incurs loss $\ell(f_{E,t}, y_t)$.

目标是最小化 regret，即

$$R_{E,n} = \sum_{t=1}^{n} \big(\ell(\widehat{p}_t, y_t) - \ell(f_{E,t}, y_t)\big) = \widehat{L}_n - L_{E,n},$$

注意到，它的比较基准是全部 follow 同一个专家 E 。这本书都考虑有限个专家，因此每个专家可以记做 $E \in \mathcal{E} \to i \in [N]$ 。

期望设计的策略能够使得 regret 的增长速度比玩的轮数 n 要慢得多，即

$$\max_{i=1,\dots,N} R_{i,n} = o(n) \quad \text{or, equivalently,} \quad \frac{1}{n}\left(\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n}\right) \overset{n\to\infty}{\longrightarrow} 0,$$

2. Weighted average prediction

一个最简单的想法还是对于每个专家维护一个权重，然后决定的时候使用加权平均来做预测。

$$\widehat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N} w_{j,t-1}},$$

玩家能获取的信息主要是玩家历史上产生的损失和各个专家在历史上产生的损失，一个比较自然的选择是让各个专家的权重取决于历史上玩家相对于该专家的 regret $R_{i,t-1} = \hat{L}_{t-1} - L_{i,t-1}$ 。（当然，也可以只取决于各个专家在历史上产生的损失，而不取决于玩家在历史上的损失；毕竟一个专家的靠谱程度应该和玩家之前有没有听从其建议无关，后面会有这样的方案，不过这里暂且假定取决于各个 regret。）

下面定义一类的权重函数，它给出了专家历史上 regret 到各个专家对应权重之间的关系；接下来将会看到它和 regret bound 之间的联系。

令 instantaneous regret vector $\mathbf{r}_t = (r_{1,t}, \cdots, r_{N,t}) \in \mathbb{R}^N$ ，使得 $\mathbf{R}_t = \sum_{k=1}^{n} \mathbf{r}_t$ 的各个分量为玩家相对于各个专家截止当前时刻的 regret。定义

$$\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^{N} \phi(u_i)\right) \qquad \text{(potential function)},$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is any nonnegative, increasing, and twice differentiable function, and $\psi : \mathbb{R} \to \mathbb{R}$ is any nonnegative, strictly increasing, concave, and twice differentiable auxiliary function.

可以看出势能函数是相对于 regret 的增函数，因此我们希望多轮之后势能函数尽可能小。

让玩家每次按照如下公式做出预测

$$\widehat{p}_t = \frac{\sum_{i=1}^{N} \nabla\Phi(\mathbf{R}_{t-1})_i \, f_{i,t}}{\sum_{j=1}^{N} \nabla\Phi(\mathbf{R}_{t-1})_j}$$

where $\nabla\Phi(\mathbf{R}_{t-1})_i = \partial\Phi(\mathbf{R}_{t-1})/\partial R_{i,t-1}$.

如果损失函数是 convex 的，可以发现（利用 Jensen 不等式），每一轮按照权重加权平均的 instantaneous regret 都是小于等于零的，即 Blackwell condition：

$$\sup_{y_t \in \mathcal{Y}} \mathbf{r}_t \cdot \nabla\Phi(\mathbf{R}_{t-1}) \leq 0 \qquad \text{(Blackwell condition).}$$

下图比较形象地说明了满足 Blackwell condition 下的更新情形。

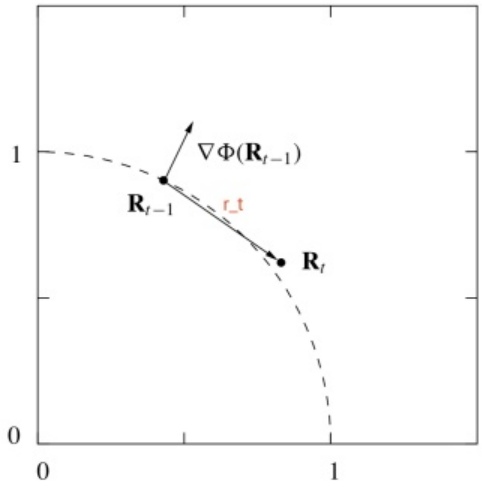

Figure 2.1. An illustration of the Blackwell condition with $N = 2$. The dashed line shows the points in regret space with potential equal to 1. The prediction at time t changed the potential from $\Phi(\mathbf{R}_{t-1}) = 1$ to $\Phi(\mathbf{R}_t) = \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)$. Though $\Phi(\mathbf{R}_t) > \Phi(\mathbf{R}_{t-1})$, the inner product between $\mathbf{r}_t$ and the gradient $\nabla\Phi(\mathbf{R}_{t-1})$ is negative, and thus the Blackwell condition holds.

这样的更新方式保证了 regret 的变化方向 $\mathbf{r}_t = \mathbf{R}_t - \mathbf{R}_{t-1}$ 和势能的梯度上升方向 $\nabla\Phi(\mathbf{R}_{t-1})$（也就是说各个专家的权重）夹角为钝角，这样虽然有可能 $\mathbf{R}_t$ 的势能还是比 $\mathbf{R}_{t-1}$ 大，但是不至于上升地太快。由于 regret 变化的***方向***是朝着势能变小的方向的，如果只看一阶近似， $\mathbf{R}_t$ 的势能应该比 $\mathbf{R}_{t-1}$ 更低。由此自然想到，如果我们 bound 更高阶的近似，就能够得到势能的上界。

***Theorem* 2.1.** *Assume that a forecaster satisfies the Blackwell condition for a potential* $\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^{N} \phi(u_i)\right)$. *Then, for all* $n = 1, 2, \ldots,$

$$\Phi(\mathbf{R}_n) \le \Phi(\mathbf{0}) + \frac{1}{2}\sum_{t=1}^{n} C(\mathbf{r}_t),$$

where

$$C(\mathbf{r}_t) = \sup_{\mathbf{u}\in\mathbb{R}^N} \psi'\left(\sum_{i=1}^{N} \phi(u_i)\right) \sum_{i=1}^{N} \phi''(u_i) r_{i,t}^2.$$

证明方法也比较简单，就是泰勒展开之后 bound 二阶导。

以上定理的意义在于 bound 了势能函数相当于就 bound 了这种策略下的 regret，注意到

$$\psi\left(\phi\left(\max_{i=1,\ldots,N} R_{i,n}\right)\right) = \psi\left(\max_{i=1,\ldots,N} \phi(R_{i,n})\right) \le \psi\left(\sum_{i=1}^{N} \phi(R_{i,n})\right) = \Phi(\mathbf{R}_n).$$

（最大的 regret 比 regret 的和要小，前一讲里面的证明也用到这个原理）

有

$$\max_{i=1,\dots,N} R_{i,n} \le \phi^{-1}\left(\psi^{-1}\left(\Phi(\mathbf{R}_n)\right)\right),$$

和前一讲里面问题设定的区别

前一讲中 decision space 和 outcome space 都是 binary 的，是离散的，不是 convex set，用 majority vote。 这里的 decision space 是一个 convex set，这样保证对于各个专家加权平均之后得到的结果还在 decision space 中，因此使用加权平均。很多情况下可以认为 decision space = outcome space。

3. Polynomially weighted average forecaster

前面定义了一种势能，多项式加权平均是满足前述定义势能函数的一个特殊的例子。多项式势能定义如下：

$$\Phi_p(\mathbf{u}) = \left(\sum_{i=1}^{N} (u_i)_+^p\right)^{2/p} = \|\mathbf{u}_+\|_p^2 \qquad \text{(polynomial potential)},$$

即，前面势能的定义中 $\psi(\cdot) = (\cdot)^{2/p}, \phi(\cdot) = (\cdot)_+^p$ 。这样每个 expert 对应的权重为

$$w_{i,t-1} = \nabla\Phi_p(\mathbf{R}_{t-1})_i = \frac{2(R_{i,t-1})_+^{p-1}}{\|(\mathbf{R}_{t-1})_+\|_p^{p-2}}$$

玩家每次的预测数值为

$$\widehat{p}_t = \frac{\sum_{i=1}^{N}\left(\sum_{s=1}^{t-1}\left(\ell(\widehat{p}_s, y_s) - \ell(f_{i,s}, y_s)\right)\right)_+^{p-1} f_{i,t}}{\sum_{j=1}^{N}\left(\sum_{s=1}^{t-1}\left(\ell(\widehat{p}_s, y_s) - \ell(f_{j,s}, y_s)\right)\right)_+^{p-1}}.$$

注意到势能定义里面的 $\psi(\cdot)$ 其实不影响玩家给出的预测数值，它的设定只是为了分析方便。因此，设定 $\Phi_p(\mathbf{u}) = \|\mathbf{u}_+\|_p$ 也是一样的。

可以套用前面的结论，可以得到这种策略下的 regret bound：

***Corollary* 2.1.** *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. Then, for any sequence $y_1, y_2, \ldots \in \mathcal{Y}$ of outcomes and for any $n \ge 1$, the regret of the polynomially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \sqrt{n(p-1)N^{2/p}}.$$

令 $\eta = 2\ln N$ ，可以得到一个最优的 bound：

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \sqrt{ne(2\ln N - 1)}$$

4. Exponentially weighted average forecaster

前面的介绍的多项式势能产生的策略不仅仅依赖于各个专家历史上产生的损失，还取决于玩家在历史上产生的损失。但是各个专家的权重其实可以不依赖于玩家产生的损失。当我们使用如下的这种指数势能函数的时候，玩家产生的权重部分就可以被抵消掉，因而形成一个和玩家历史无关的 forecaster。

定义指数势能为：

$$\Phi_\eta(\mathbf{u}) = \frac{1}{\eta}\ln\left(\sum_{i=1}^{N} e^{\eta u_i}\right) \qquad \text{(exponential potential)},$$

各个专家的权重为：

$$w_{i,t-1} = \nabla\Phi_\eta(\mathbf{R}_{t-1})_i = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^{N} e^{\eta R_{j,t-1}}},$$

或者写作：

$$w_{i,t} = \frac{w_{i,t-1}e^{-\eta\ell(f_{i,t},y_t)}}{\sum_{j=1}^{N} w_{j,t-1}e^{-\eta\ell(f_{j,t-1},y_t)}}.$$

相应的 forecaster 为，观察到玩家的历史损失被上下约掉了：

$$\widehat{p}_t = \frac{\sum_{i=1}^{N}\exp\left(\eta\left(\widehat{L}_{t-1} - L_{i,t-1}\right)\right) f_{i,t}}{\sum_{j=1}^{N}\exp\left(\eta\left(\widehat{L}_{t-1} - L_{j,t-1}\right)\right)} = \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1}} f_{i,t}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}.$$

相应地，还是套用前面的结论，可以得到这种 forecaster 对应的 regret bound：

***Corollary* 2.2.** *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. For any n and $\eta > 0$, and for all $y_1, \dots, y_n \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \frac{\ln N}{\eta} + \frac{n\eta}{2}.$$

选择 $\eta=\sqrt{2\ln N/n}$ ，可以得到最好情况下的 regret bound 为 $\sqrt{2n\ln N}$ 。

如果针对这种该方法进行分析（而不是直接套用前面的结论），可以得到一个更好的 bound。隐约记得这个 bound 导师的课讲过。

***Theorem* 2.2.** *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \le \frac{\ln N}{\eta} + \frac{n\eta}{8}.$$

In particular, with $\eta = \sqrt{8\ln N/n}$, the upper bound becomes $\sqrt{(n/2)\ln N}$.

可以看到，它比前一个 bound 改善了一个常数 2。这种策略比多项式势能对应的策略更好一些（常数 $2\sqrt{e}$ 倍）。

5. Uniform over time

注意到上述策略里面的超参数 $\eta=\sqrt{8\ln N/n}$ 中含有总共玩的轮数，亦即要求先知道总共玩多少轮。但很多时候我们希望不提前告诉玩的总轮数，并且不管玩多少轮，都有相应的 regret bound（即，这里讲的 uniform over time）。

一个简单的办法是使用 doubling trick，把时间划分为若干段，每一段的长度都是前一段的两倍，每一段使用和这一段长度对应的 η ，这样仍然可以套用前面的结论得到相应的 regret bound。该 regret bound 在条件假设上有所放松，因此得到的 bound 略差一些（差了常数倍 $\frac{\sqrt{2}}{\sqrt{2}-1}\approx 3.41$ ）。

(The doubling trick) Consider the following forecasting strategy ("doubling trick"): time is divided in periods $(2^m, \ldots, 2^{m+1}-1)$, where $m = 0, 1, 2, \ldots$. In period $(2^m, \ldots, 2^{m+1}-1)$ the strategy uses the exponentially weighted average forecaster initialized at time 2^m with parameter $\eta_m = \sqrt{8(\ln N)/2^m}$. Thus, the weighted average forecaster is reset at each time instance that is an integer power of 2 and is restarted with a new value of η. Using Theorem 2.2 prove that, for any sequence $y_1, y_2, \ldots \in \mathcal{Y}$ of outcomes and for any $n \ge 1$, the regret of this forecaster is at most

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \le \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{\frac{n}{2}\ln N}.$$

观察到 doubling trick 里面相当于把 η 中的总轮数换成了差不多为当前经历的时间步，另外一个看起来更优雅的方法是直接把参数中的总轮数换成当前的时间步，即 $\eta(t)=\sqrt{8\ln N/t}$ 。这样能够得到一个

更好的 bound：

Theorem 2.3. *Assume that the loss function ℓ is convex in its first argument and takes values in $[0, 1]$. For all $n \geq 1$ and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster with time-varying parameter $\eta_t = \sqrt{8(\ln N)/t}$ satisfies*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq 2\sqrt{\frac{n}{2}\ln N} + \sqrt{\frac{\ln N}{8}}.$$

6. An improvement for small losses

回忆前面的一个例子，如果告知存在一个不犯错误的专家，那么我们可以采取更为激进的策略（如果任何一个专家犯错，都直接把它剔除），同时能够获得一个更好 bound。这个regret bound 与玩家玩的轮数 n 无关，即不论玩多少轮，犯错次数都不超过某个数。

这个例子告诉我们，如果预知存在一个犯错较少的专家，那么能够采取一个更激进的策略，使得 regret bound 更紧。下面的定理告诉我们，假设已知有一个专家遭受的损失为 $L_n^* = \min_{i \in [N]} L_{i,n}$ ，那么玩家所受损失有如下上界：

Theorem 2.4. *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. Then for any $\eta > 0$ the regret of the exponentially weighted average forecaster satisfies*

$$\widehat{L}_n \leq \frac{\eta L_n^* + \ln N}{1 - e^{-\eta}}.$$

通过选择一个合适的 η 能够得到一个最优的上界：

Corollary 2.4. *Assume the exponentially weighted average forecaster is used with $\eta = \ln\left(1 + \sqrt{(2\ln N)/L_n^*}\right)$, where $L_n^* > 0$ is supposed to be known in advance. Then, under the conditions of Theorem 2.4,*

$$\widehat{L}_n - L_n^* \leq \sqrt{2L_n^* \ln N} + \ln N.$$

注意到，当 $L_n^* = o(\sqrt{n})$ 的时候，这个 bound 比之前的结果更好（Theorem 2.2），否则会更差。

7. Forecasters using the gradient of the losses

前面讲的 polynomially weighted average forecaster 和 exponentially weighted average forecaster 都是基于对于势能的导数的分析而得到的策略（Theorem 2.1）。这里讲另外一种 forecaster，它适用于损失函数可导并且 decision space 是有限维度的 convex linear space 的情形。

该 forecaster 可以被写作：

$$\widehat{p}_t = \frac{\sum_{i=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(\widehat{p}_s, y_s) \cdot f_{i,s}\right) f_{i,t}}{\sum_{j=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(\widehat{p}_s, y_s) \cdot f_{j,s}\right)}.$$

直观地来说，如果在历史上，某个专家给出的建议 $f_{i,s}$ 能够进一步帮助玩家（在历史上的那个时刻）减小损失，就给该专家分配更多的权重。类似地，如果损失函数的导数 $\|\nabla \ell\| \leq 1$ ，这样的 forecaster 也存在和 exponentially weighted average forecaster 类似的 regret bound：

***Corollary* 2.5.** *Assume that the decision space $\mathcal{D}$ is a convex subset of the euclidean unit ball $\{q \in \mathbb{R}^d : \|q\| \leq 1\}$, the loss function ℓ is convex in its first argument and that its gradient $\nabla \ell$ exists and satisfies $\|\nabla \ell\| \leq 1$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the gradient-based exponentially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}.$$

8. Scaled losses and signed games

前面假设了 loss 的范围在 $[0,1]$ 之间，假如 loss 的范围在 $[0,M]$ 之间，相应的 regret bound 会怎样呢？

对于之前的一个损失函数 ℓ ，考虑一个刚好被 scale M 倍的新损失函数，这样玩家受到的损失也会 scale M 倍，但是 outcome 序列可以刚好选择为使得 L_n^* 不变。观察到 $\hat{L}_n \leq \frac{\eta L_n^* + \ln N}{1 - e^{-\eta}}$ ，即前一部分不会被 scale，但是后面的一部分被 scale，则有 $\hat{L}_n \leq \frac{\eta L_n^* + M \ln N}{1 - e^{-\eta}}$ ，相应地有

$$\widehat{L}_n - L_n^* \leq \sqrt{2L_n^* M \ln N} + M \ln N.$$

（这只是一个口头的分析，不严谨，详细的证明需要把 $\ell / M \to \ell$ 带入到原本的证明中）

如果 loss 的范围在 $[-M,0]$ 之间，类似地有：

$$G_n^* - \widehat{G}_n \leq \sqrt{2G_n^* M \ln N} + M \ln N.$$

其中 $G = -L, G^* = -L^*$ 。

当 loss 的范围在 $[-M,M]$ 时，也可以用 scale 的方法分析，得到

$$H_n^* - \widehat{H}_n \leq \sqrt{4(H_n^* + Mn)(M \ln N)} + 2M \ln N = O\left(M\sqrt{n \ln N}\right).$$

可以大致认为 $H = -L, H^* = -L^*$ 。比较糟糕的是，这里多了一个和玩的轮数有关的项 $\sqrt{n}$ 。不过可以使用其他的策略（multilinear forecaster，具体定义参考书），使得

$$H_n^* - \widehat{H} \leq 2\sqrt{Q_n^* \ln N} + 4M \ln N.$$

其中

$$Q_n^* = h(f_{k,1}, y_1)^2 + \cdots + h(f_{k,n}, y_n)^2 \quad \text{where} \quad k \quad \text{is such that} \quad H_{k,n} = H_n^* = \max_{i=1,\ldots,N} H_{i,n}.$$

9. Simulatable experts and minimax regret

假定专家给出的建议只依赖于历史上的 outcome sequence，即 $f_{E,t}: \mathcal{Y}^{t-1} \to \mathcal{D}$ ，并且每个专家的在未来的预测结果（通过给定假想的未来的 outcome sequence）都能通过模拟得到。这个额外的可以被利用的条件称作 simulatable experts。这个额外的条件能够为策略带来更多的信息，从而形成更好的 regret bound。

另外一种特殊情况是 static experts，即 $f_{E,t}$ 是一个 constant function，即其数值可以和时间 t 有关，但是和历史上的 outcome 无关。

这件事情第一眼看上去有点微妙，不太好理解，看了一下书后面第八章的例子，大概是这么个情况。首先，regret 的目标是最小化相对于每个专家的损失差距，因此，如果各个专家之间意见产生的损失差距不大，那么我们选择他们意见的加权平均产生的 loss 跟最好的专家之间差距肯定不会太大。其次，如果专家的建议可以被模拟，那么我们可以按照 $n, n-1, \cdots, t$ 的顺序，把各种 outcome 产生的 regret 都倒推回来，并且尽量减小最坏 outcome sequence 产生的 regret。这个做法类似于强化学习里面的 planning，只不过强化学习里面是优化期望，这里是优化最坏情形。

注意到在这本书的分析里面着重强调的是 bound 最坏情况，即找一个策略，使得在环境给出最坏的 outcome sequence 以及专家给出最坏的建议的时候，还能保证相应的 regret 上界。这个问题可以被显式地写为如下目标：

$$V_n^{(N)} = \sup_{(f_{1,1},\ldots,f_{N,1})\in\mathcal{D}^N} \inf_{\hat{p}_1\in\mathcal{D}} \sup_{y_1\in\mathcal{Y}} \sup_{(f_{1,2},\ldots,f_{N,2})\in\mathcal{D}^N} \inf_{\hat{p}_2\in\mathcal{D}} \sup_{y_2\in\mathcal{Y}} \cdots \sup_{(f_{1,n},\ldots,f_{N,n})\in\mathcal{D}^N} \inf_{\hat{p}_n\in\mathcal{D}} \sup_{y_n\in\mathcal{Y}} \left(\sum_{t=1}^n \ell(\hat{p}_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=1}^n \ell(f_{i,t}, y_t) \right),$$

注意到蓝色框里面三个一组，说明了『策略』和『专家、环境』之间的对抗关系。

如果是 static expert（不过不能预知它们的序列，不然它就是 simulatable 的了），令这些 expert 的建议来自一个函数族 $\mathcal{F}$ ，该问题可以被表述为：

$$V_n^{(N)} = \inf_P \sup_{\{\mathcal{F}:|\mathcal{F}|=N\}} \sup_{y^n\in\mathcal{Y}^n} \max_{i=1,\ldots,N} \left(\widehat{L}_n(P, \mathcal{F}, y^n) - \sum_{t=1}^n \ell(f_{i,t}, y_t) \right),$$

关于它，前面我们得到一个最好的 regret bound： $V_n^{(N)} \le \sqrt{(n/2)\ln N}$ 。

如果 expert 是可以 simulatable 的，并且相应的函数族为 $\mathcal{F}$ ，那么它就不再成为一个『对抗』的因素了（不用分析关于它的最坏情形了），相应的 regret 就依赖于这个函数族，即：

$$V_n(\mathcal{F}) = \inf_P \sup_{y^n\in\mathcal{Y}^n} \left(\sum_{t=1}^n \ell\left(\widehat{p}_t(y^{t-1}), y_t\right) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell\left(f_t(y^{t-1}), y_t\right) \right).$$

当 $|\mathcal{F}| = n$ 时，有 $V_n(\mathcal{F}) \le V_n^{(N)}$ （来自第八章）。其原因是专家 $\mathcal{F}$ 可以被模拟之后，它就可以被策略所

利用，从而得到更好的 regret bound。

总结

对于最为基础的 expert problem 问题，这一章先提出了基于势能的 weighted average 方案，并且由此推出了 polynomially weighted average 和 exponentially weighted average。接着提出了另外一种（不基于势能）的把 loss 的梯度作为权重的 weighted average 方案。文章出了：1）超参数不含有总轮数 n 的修正方案；2）得知最优专家遭受较小损失 L_n^* 时的改进方案；3）每一局的损失函数不再是 $[0,1]$ 之间而是任意区间内的 bound。最后，文章做了 simulatable expert 和 discounted regret 的推广，其中后者不太感兴趣，没写出来。

思考

目前看到 expert problem 相比于 RL 问题，有两个比较重要的区别：

- Expert problem 假定 loss function 已知，这样即使你没有采取某个专家的建议，它的 loss 也会被知道；RL 里面只能知道采取某个行动（某个专家的建议）之后的收益，而未采取的行动对应的收益不知道。考虑到这一点，前面讲的 POLITEX 就需要使用价值函数的估计，来近似地得到未采取行动的收益/损失。
- Expert problem 的目标是与给定的 expert 作比较，因此策略的选择只考虑 expert 的建议们的『平均值』，即内部，而不要考虑这些建议『外部』的区域，即使可能这些 expert 都很差。隐约感觉 expert problem 不太有『探索』的问题，主要是如何『利用』的问题。而 RL 有探索的问题。如果像 POLITEX 一样用 expert problem 去套 RL 问题，一般就假设 expert 包含了所有可能的 action。

（最后这一条我感觉我没说清楚。。因为我还没想清楚。。）

发布于 2019-07-23

强化学习 (Reinforcement Learning)　博弈论

文章被以下专栏收录