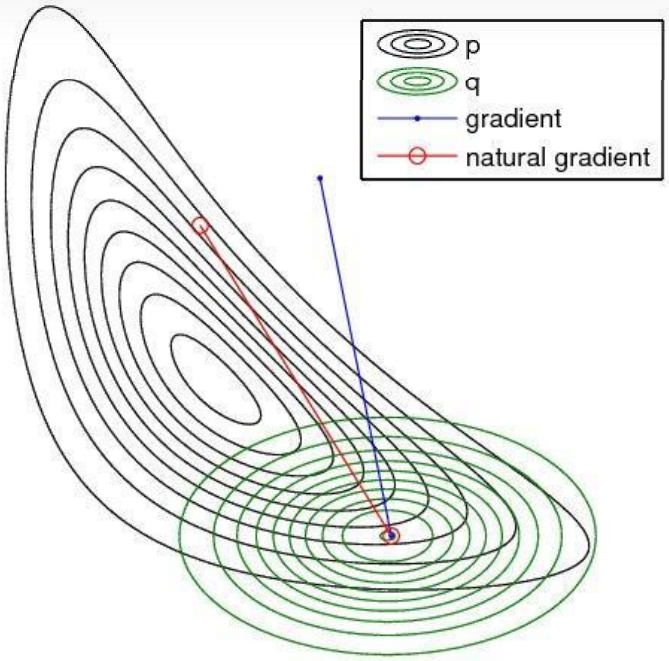


费舍尔信息矩阵及自然梯度法

Monica Li

关注她

236 人赞同了该文章

原作者：Agustinus Kristiadi
本翻译增加了一些补（Tu）充（Cao）

费舍尔信息矩阵（Fisher Information Matrix, FIM）

假设我们有一个参数为向量 θ 的模型，它对分布 $p(x|\theta)$ 建模。在频率派统计学中，我们学习θ的方法是最大化$p(x|\theta)$与参数θ的似然（likelihood）。 为了评估我们对 θ 估计的好坏，我们定义了一个评分函数(score function)：

$$s(\theta) = \nabla_\theta \log p(x|\theta),$$

评分函数即为对数似然函数的梯度。以下关于评分函数的结果是我们讨论的重要基础。

声明：我们模型的评分期望值为零。

证明. 以下对 θ 求梯度

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(x|\theta)}[s(\theta)] &= \mathop{\mathbb{E}}_{p(x|\theta)}[\nabla \log p(x|\theta)] \\
&= \int \nabla \log p(x|\theta) p(x|\theta) \mathrm{d}x \\
&= \int \frac{\nabla p(x|\theta)}{p(x|\theta)} p(x|\theta) \mathrm{d}x \\
&= \int \nabla p(x|\theta) \mathrm{d}x \\
&= \nabla \int p(x|\theta) \mathrm{d}x \\
&= \nabla 1 \\
&= 0
\end{aligned}$$

据模型评分的协方差，

那么，我们可以把它看作是一种信息。上面评分函数的协方差就是费舍尔信息的定义。由于我们假设θ是一个向量，所以费舍尔信息是以矩阵形式存在的，称为费舍尔信息矩阵（FIM）：

$$\mathbf{F} = \mathop{\mathbb{E}}_{p(x|\theta)}[\nabla \log p(x|\theta) \nabla \log p(x|\theta)^{\mathrm{T}}].$$

然而，通常我们的似然函数是复杂的，很难计算期望值。我们可以用经验分布 $\tilde{q}(x)$ 来近似 $\mathbf{F}$ 中的期望值，它由我们的训练数据 $X = \{x_1, x_2, \cdots, x_N\}$ 给出。在这种形式下，$\mathbf{F}$被称为经验费舍尔：

$$\mathbf{F} = \frac{1}{N} \sum_{i=1}^{N} p(x_i|\theta) \nabla \log p(x_i|\theta) \nabla \log p(x_i|\theta)^{\mathrm{T}}.$$

费舍尔信息矩阵与黑森（Hessian）矩阵的联系

$\mathbf{F}$ 有一个不是很明显的属性，它可以理解为，模型对数似然的黑森矩阵期望的负值。

声明：对数似然的负期望黑森，等于费舍尔信息矩阵$\mathbf{F}$。

证明. 对数似然的黑森由其梯度的雅可比矩阵给出：

$$\begin{aligned}
\mathbf{H}_{\log p(x|\theta)} &= \mathrm{J}\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right) \\
&= \frac{\mathbf{H}_{p(x|\theta)} p(x|\theta) - \nabla p(x|\theta) p(x|\theta)^{\mathrm{T}}}{p(x|\theta) p(x|\theta)} \\
&= \frac{\mathbf{H}_{p(x|\theta)} p(x|\theta)}{p(x|\theta) p(x|\theta)} - \frac{\nabla p(x|\theta) p(x|\theta)^{\mathrm{T}}}{p(x|\theta) p(x|\theta)} \\
&= \frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}},
\end{aligned}$$

其中，第二行使用了函数除法导数规则。对上述式子求期望，有：

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(x|\theta)}\left[\mathbf{H}_{\log p(x|\theta)}\right] &= \mathop{\mathbb{E}}_{p(x|\theta)}\left[\frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}}\right] \\
&= \mathop{\mathbb{E}}_{p(x|\theta)}\left[\frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)}\right] - \mathop{\mathbb{E}}_{p(x|\theta)}\left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}}\right] \\
&= \int \frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} p(x|\theta) \mathrm{d}x - \mathop{\mathbb{E}}_{p(x|\theta)}\left[\nabla \log p(x|\theta) \nabla \log p(x|\theta)^{\mathrm{T}}\right] \\
&= \mathbf{H}_{\int p(x|\theta) \mathrm{d}x} - \mathbf{F} \\
&= \mathbf{H}_1 - \mathbf{F} \\
&= -\mathbf{F}.
\end{aligned}$$

故此，$\mathbf{F} = -\mathop{\mathbb{E}}_{p(x|\theta)}\left[\mathbf{H}_{\log p(x|\theta)}\right]$

费舍尔信息矩阵被定义为评分函数的协方差，它是一个曲率矩阵，可以理解为对数似然函数的黑森负期望。因此，$\mathbf{F}$的直接应用，是在二阶优化方法中替换 $\mathbf{H}$ 。

此外，$\mathbf{F}$与KL散度之间有着振奋人心的联系，就产生了自然梯度法。

自然梯度法（Natural Gradient Descent）

分布空间及KL散度

根据上文，我们有一个概率模型，用它的似然 $p(x|\theta)$ 来表示，我们想最大化这个似然函数，找到最有可能的参数。等价的表述是最小化损失函数 $\mathcal{L}(\theta)$ ——即负对数似然（negative log

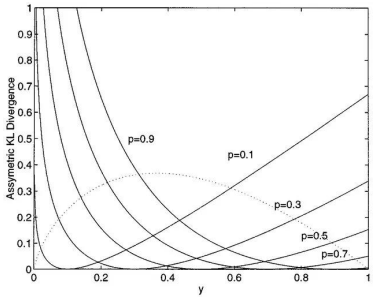
$$\frac{-\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\|} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \arg\min_{d \text{ s.t. } \|d\| \leq \epsilon} \mathcal{L}(\theta + d).$$

上述表达式的意思是，在参数空间中最陡峭的下降方向上，选取一个向量 d ，使得新参数 $\theta + d$ 在当前参数 θ 的 ϵ 邻域范围内，且使得损失函数最小化。注意，这里采用欧式范数来表达邻域的情况。因此，梯度下降法的优化建立在参数空间的欧氏距离上。

然而，如果我们的目标是最小化损失函数（最大化似然），那么我们自然可以在所有可能由参数 θ 实现的似然函数所构成的空间中进行移动。由于似然函数本身是一个概率分布，我们称这个空间为分布空间。因此，在这个分布空间，而不是参数空间中，采取最陡峭的下降方向，实际上对我们的优化更有意义。

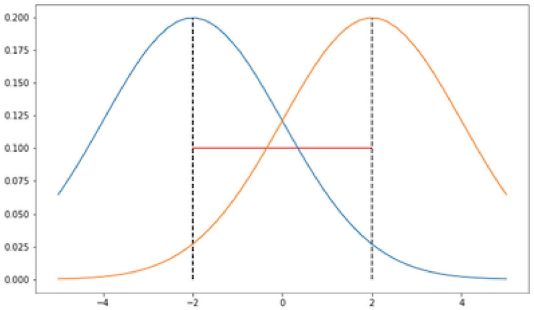
那么在这个空间中我们需要使用哪种度量方式来衡量不同似然函数之间的距离呢？一个常见选择是KL散度（KL-divergence）：$D_{\mathrm{KL}}(p(x)\|q(x)) = \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$

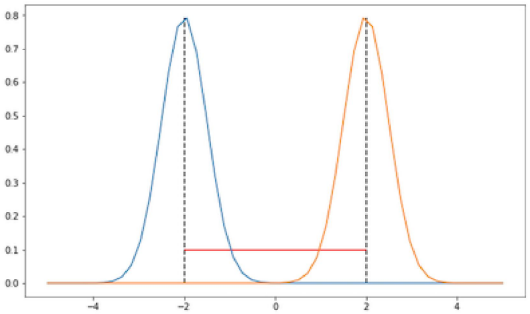
KL散度衡量了两个分布的"紧密程度"。尽管KL散度是非对称的，因此不算是一个严格的距离函数，我们还是可以使用它。这是因为当d趋近于零时，KL散度是渐进对称的。所以，在一个局部邻域内，KL散度是近似对称的 。



KL散度在概率小时上升更快，因此在p值较小时可以提供更好的概率估计

下图显示了在参数空间中使用欧氏距离时可能会遇到的问题。假设以下分布均为高斯分布，且均值在-2和2，一幅图中分布的方差为2，另一幅图中分布的方差为0.5：





在两图中，根据欧氏距离（红色线段），两幅图中高斯分布的距离是相同的，也就是4。然而，如果考虑到高斯分布的形状，在第一幅图和第二幅图中，两个分布的“紧密度”明显是不同的，所以在分布空间中，它们的距离应该不同。在第一幅图中，两个分布的KL散度应该更低，因为它们曲

处理问题，就无法将由

我们还需要回答一个问题：FIM和KL散度之间到底有什么联系？答案是，KL散度是分布空间的度量，而FIM定义了分布空间的局部曲率。

> 注：黑森矩阵本身就描述了函数的局部曲率

声明：两个分布 $p(x|\theta)$ 和 $p(x|\theta')$ 之间，相对于 θ' 的KL散度，在 $\theta'=\theta$ 时的黑森矩阵，即为费舍尔信息矩阵 $\mathbf{F}$ 。

证明. KL散度可以写成熵（entropy）和交叉熵（cross-entropy）两项，即：

$$\mathbf{KL}\left[p(x|\theta)\,||\,p(x|\theta')\right]=\mathop{\mathbb{E}}_{p(x|\theta)}[\log p(x|\theta)]-\mathop{\mathbb{E}}_{p(x|\theta)}[\log p(x|\theta')]$$

对 θ' 求一阶导：

$$\begin{aligned}\nabla_{\theta'}\mathbf{KL}\left[p(x|\theta)||p(x|\theta')\right]&=\nabla_{\theta'}\mathop{\mathbb{E}}_{p(x|\theta)}[\log p(x|\theta)]-\nabla_{\theta'}\mathop{\mathbb{E}}_{p(x|\theta)}[\log p(x|\theta')]\\&=-\mathop{\mathbb{E}}_{p(x|\theta)}[\nabla_{\theta'}\log p(x|\theta')]\\&=-\int p(x|\theta)\nabla_{\theta'}\log p(x|\theta')\mathrm{d}x.\end{aligned}$$

> 第一步第一项里面没有θ'，所以为0

对 θ' 求二阶导：

$$\nabla^2_{\theta'}\mathbf{KL}\left[p(x|\theta)||p(x|\theta')\right]=-\int p(x|\theta)\nabla^2_{\theta'}\log p(x|\theta')\mathrm{d}x.$$

因此，相对于 θ' 的KL散度，在 $\theta'=\theta$ 时的黑森矩阵为：

$$\begin{aligned}\mathbf{H}_{\mathbf{KL}[p(x|\theta)\|p(x|\theta')|}&=-\int p(x|\theta)\nabla^2_{\theta'}\log p(x|\theta')|_{\theta'=\theta}\mathrm{d}x\\&=-\int p(x|\theta)\mathbf{H}_{\log p(x|\theta)}\mathrm{d}x\\&=-\mathop{\mathbb{E}}_{p(x|\theta)}\left[\mathbf{H}_{\log p(x|\theta)}\right]\\&=\mathbf{F}.\end{aligned}$$

最后一步请参见上方第二部分“FIM与黑森矩阵的联系”的结论。

分布空间的最速下降法

现在我们准备使用FIM来改进梯度下降法。
首先，我们需要推导出KL散度在 θ 处的泰勒展开。

声明：当 $d\to 0$ 时，KL散度的二阶泰勒展开为 $\mathbf{KL}\left[p(x|\theta)\|p(x|\theta+d)\right]\approx\frac{1}{2}d^{\mathrm{T}}\mathbf{F}d.$

证明. 根据定义，KL散度的二阶泰勒数列展开为：

$$\begin{aligned}\mathbf{KL}\left[p(x|\theta)\|p(x|\theta+d)\right]&\approx\mathbf{KL}\left[p(x|\theta)\|p(x|\theta)\right]+\left(\nabla_{\theta'}\mathbf{KL}\left[p(x|\theta)\|p(x|\theta')\right]|_{\theta=\theta'}\right)^{\mathrm{T}}d+\frac{1}{2}d^{\mathrm{T}}\mathbf{F}d\\&=\mathbf{KL}\left[p(x|\theta)\|p(x|\theta)\right]-\mathop{\mathbb{E}}_{p(x|\theta)}\left[\nabla_{\theta}\log p(x|\theta)\right]^{\mathrm{T}}d+\frac{1}{2}d^{\mathrm{T}}\mathbf{F}d\end{aligned}$$

第一项中的两个分布都是 $p(x|\theta)$ ，所以KL散度为0
第二项中，对数似然函数的梯度的期望（也就是上面证明过程中KL散度的梯度），已在本文第一个声明中被证明为0

> 因此必须展开到二阶

是通常情况下在参数空间中以欧氏距离为度量。对d的要求可以写成：

$$d^* = \underset{d\ s.t.\ \mathrm{KL}[\mathrm{p}(\mathrm{x}|\theta)\|\mathrm{p}(\mathrm{x}|\theta+\mathrm{d})]=\epsilon}{\arg\min} \mathcal{L}(\theta+d),$$

其中 ϵ 是某个常数。将KL散度固定为某个常数的目的，是为了确保不管曲率如何，参数都以恒定的速度在分布空间中移动，这使得算法对模型的更新更加稳健，即算法不关心模型本身引发的参数改变，它只关心参数改变引起的分布变化。

如果将 d^* 写成拉格朗日算子的形式，并用二阶泰勒序列展开逼近约束条件中的KL散度，用一阶泰勒展开逼近 $\mathcal{L}(\theta+d)$ ：

$$\begin{aligned} d^* &= \arg\min_d \mathcal{L}(\theta+d) + \lambda(\mathrm{KL}\left[p(x|\theta)\|p(x|\theta+d)\right]-\epsilon) \\ &\approx \arg\min_d \mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^{\mathrm{T}} d + \frac{1}{2}\lambda d^{\mathrm{T}}\mathbf{F}d - \lambda\epsilon \end{aligned}$$

为求得最小值，我们求它对 d 导数为0时方程的解：

$$\begin{aligned} &\frac{\partial}{\partial d}\left(\mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^T d + \tfrac{1}{2}\lambda d^T \mathbf{F} d - \lambda\epsilon\right) = 0 \\ &\nabla_\theta \mathcal{L}(\theta) + \lambda \mathbf{F} d = 0 \\ &\lambda \mathbf{F} d = -\nabla_\theta \mathcal{L}(\theta) \\ &d = -\tfrac{1}{\lambda}\mathbf{F}^{-1}\nabla_\theta \mathcal{L}(\theta) \end{aligned}$$

只要系数 $\frac{1}{\lambda}$ 是常数（可以将它吸收到学习率中），就可以得到最优下降方向，即考虑以 $\mathbf{F}^{-1}$ 定义局部曲率的分布空间中，梯度的相反方向。在此，我们得到了自然梯度（Natural Gradient）和自然梯度法（Natural Gradient Descent）的定义：

定义：自然梯度被定义为 $\tilde{\nabla}_\theta \mathcal{L}(\theta) = \mathbf{F}^{-1}\nabla_\theta \mathcal{L}(\theta).$

学习率：我们将KL散度固定为常数的限制单独写出来

$$\mathrm{KL}\left[\mathrm{p}(\mathrm{x}|\theta)\|\mathrm{p}(\mathrm{x}|\theta+\mathrm{d})\right] \approx \tfrac{1}{2}(\theta-\theta')^{\mathrm{T}}\mathbf{F}(\theta-\theta') = \epsilon$$

由于在实际训练时，梯度 g 是通过使用在参数 θ 的情况下采样模拟的分布而计算出来的，我们将其记为 $\hat{g}$ ，上面的等式可以转写为：

$$\tfrac{1}{2}(\alpha\hat{g})^{\mathrm{T}}\mathbf{F}(\alpha\hat{g}) = \epsilon$$

解得 $\alpha = \sqrt{\frac{2\epsilon}{\hat{g}^{\mathrm{T}}\mathbf{F}\hat{g}}}$

算法：自然梯度法

```
Repeat:
    对模型进行正向传播，计算损失L(θ)
    计算梯度g = ∇θL(θ)
    计算Fisher信息矩阵F，或其相对于训练数据的经验版本
    计算自然梯度
    更新参数：θ=θ- α自然梯度，其中α是学习率。
Until 收敛
```

即 $\theta_{k+1} = \theta_k + \sqrt{\frac{2\epsilon}{\hat{g_k}^{\mathrm{T}}\mathbf{F}\hat{g_k}}}\mathbf{F}^{-1}\hat{g}_k, \text{for } k = 0, 1, 2, \cdots$

编辑于 2021-08-09 04:43

「先给TRPO打个底」