你的模型可靠吗：模型的预测置信度和校准

问题引入

> 试想，你约一个朋友打篮球，他说：“现在有点事儿，处理完事情大约1小时后球场见。”结果1小时后你到了球场，完全看不到他人影。直到超预定时间1小时后，他才出现在球场，他尴尬着解释说：没想到花了2小时才处理完事情。这时你心想：这人说好1小时后到结果2小时才到，真不靠谱啊。
>
> 不过总算是打起球了。他说：“我最近苦练三分球，十投九中！”然后他拿起球投出去10个，进了2个。你心想：这人水平不行吹的倒厉害，真不靠谱啊

上面给出了做事不靠谱的两个例子。在第一个例子中，朋友认为自己1小时就能处理完，结果实际上2小时才处理完，这说明他对处理事情所要花的时间预测不准。而实际上，模型的预测不可能绝对准确，我们也不要求模型能够对每一个样本都能预测的很准，但是我们希望模型在面对拿不准的样本时，能够给出低置信度的预测。比如在这个例子中，如果朋友能够这么回复：“我处理事情有七成可能只需要1小时，但有三成可能需要2小时。”那么我们就能够提前做好他2小时后才来的准备，也不至于后面尴尬的同时让人感觉不靠谱。

在第二个例子中，朋友认为自己能够有90%的命中率，实际上只有20%的命中率。这就是另一种不靠谱：他实际的投篮水平远低于他所认为的水平，这说明他对自己投篮水平的预测过于自信了。最靠谱的情况是：他说自己只能10投2中，而事实也是如此。

在日常中，这种不靠谱的朋友可能不算什么大问题。但是在一些安全关键型应用场景中，如果模型也这么不靠谱，就可能会酿成大祸。比如，在医疗图像诊断和自动驾驶领域，如果模型做出了不靠谱的预测，就将危及人们的生命安全。

上述的两个例子，实际上是从两个层面给出了我们对模型可靠性的要求：

- 模型要能对预测给出置信度
- 模型给出的置信度要符合实际的准确率

为此，我们需要解决两个问题：

- 如何使模型能够预测置信度？——不确定性估计
- 如何使置信度符合实际的准确率？——模型校准

模型的预测置信度

我们可以从两个角度理解置信度：

- 从数值上考虑：置信度是满足归一化条件的[0,1]之间的值。
- 从不确定性的角度考虑：预测是随机变量，置信度是一种概率，其与预测服从的分布有关。

基于这两种不同的理解，我将预测置信度的方法分为两大类：

- 直接映射法：将模型的输入和输出映射到满足归一化条件的[0,1]之间的值。例如，分类问题中的softmax（逻辑回归）。
- 构造分布法：通过一些手段构造模型预测的分布，从分布中计算概率。例如，集成学习、贝叶斯神经网络（通过多次采样估计分布）、证据神经网络（通过预测分布的控制参数）等。

模型校准

当我们的模型可以预测置信度时，我们希望模型给出的置信度和准确率相匹配，也就是希望模型的“自信程度”和“客观能力”相匹配，这样置信度才是一种可靠的度量。使模型的置信度和准确率相匹配就叫做**校准**。

记置信度为p，准确率为$accuracy(p)$，我们可将模型校准公式化地定义为：

$$p \rightarrow accuracy(p), \forall p \in [0,1]$$

上式中，校准的目标是置信度在范围$[0,1]$内都与准确率匹配。

模型校准有以下三点好处：

- 模型校准后，我们就可以通过置信度评估准确率。
- 置信度有更明确的物理意义，提升了模型的可解释性。
- 当模型的预测处于决策流程的中间环节，我们可以利用预测置信度进一步地做出决策。

下面我们从分类问题和回归问题两方面，介绍模型如何预测置信度和进行校准。

一种分类模型的校准方法——温度缩放

以上介绍了分类问题中的校准目标和校准性能指标，下面介绍一种简单高效的分类模型校准方法：温度缩放。注意，温度缩放仅适用于逻辑回归模型，即利用softmax得到置信度的分类模型。

温度缩放具体的做法为：利用与训练集不相交的一个校准集，以最小化交叉熵为目标学习最优的温度缩放参数T：

$$\operatorname{softmax}(z_i, T) = \frac{e^{z_i/T}}{\sum_{j=1}^{K} e^{z_j/T}}$$

> 交叉熵是一种度量概率模型质量的指标。理论上，只有估计的分布与实际的分布一致，交叉熵才会最小。通过最小化交叉熵，就可以使预测的分布越来越接近实际，因此能够提高模型的校准性能。

一种常用的校准方法——保序回归

温度缩放仅适用于分类问题，而保序回归分类和回归问题均适用。

我们需要一个与训练集不相交的校准集，其中包含的是观测点和标签。然后我们需要构建再校准数据集（Recalibration Dataset）。再校准数据集中，一条数据包括准确率和置信度：

$$D(p) = \{(accuracy(p), p)\}$$

为了使得置信度在整个[0,1]上都有样本，我们可以等间隔地取一些值$0 \leqslant a_1 < a_2 < \cdots < a_m \leqslant 1$，比如$\{0, 0.01, 0.02, \cdots, 0.99, 1.00\}$，构成整个再校准数据集$D$：

$$D = \bigcup D(p) = \{(accuracy(p), p)\}_{p \in \{0, a_1, a_2, \cdots, a_m, 1\}}$$

在分类问题中，再校准数据集的大小与分箱数目一致。

有了再校准数据集，我们就可以训练一个新模型，将准确率作为标签，使置信度与准确度尽量吻合，从而达到校准的目的。保序回归模型将每个箱子映射到一个新的值q_m，在最小化校准误差的同时保证q_m是递增的。优化问题可以表述为：

$$\begin{aligned} \min_{q_1, \cdots, q_m, a_1, \cdots, a_m} & \sum_{j=1}^{m} \sum_{i=1}^{n} 1(a_j \leqslant p_i < a_{j+1})(q_m - accuracy(p_i))^2 \\ s.t. \quad & 0 = a_1 \leqslant a_2 \leqslant \cdots \leqslant a_{M+1} = 1 \\ & q_1 \leqslant q_2 \leqslant \cdots \leqslant q_m \end{aligned}$$

保序回归模型实际上就是一个单调递增的分段函数。

分类模型的置信度预测和校准

分类问题的置信度

在分类问题中，对于**单个样本**而言，置信度代表着模型对于该样本的“自信程度”。例如，模型认为某个样本的类别为A，并给出置信度为80%。

对于某个测试集而言，我们亦可以统计测试样本集合中的每一个样本的置信度，取平均来作为测试集的置信度。

分类问题中，最常用的置信度预测方法是利用**softmax（逻辑回归）**来得到置信度预测值。对于一个K分类问题，我们记模型的输出为$z = (z_1, z_2, \cdots, z_K)$，利用softmax得到置信度预测：

$$\operatorname{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

分类问题的准确率

分类问题中的一个模型指标是**准确率**（accuracy），准确率为在某个**测试样本集**上，模型预测正确的样本占总样本数的百分比。准确率代表模型的“客观能力”。需要指出的是，对于单个样本讨论准确率没有意义，因为要么预测正确，要么预测错误。

分类模型的校准

理想中，我们希望对于80%置信度的样本，模型的准确率也是80%。我们无法做到单个样本的置信度和准确率匹配，因为单个样本的准确率要么是0，要么是100%。因此，我们只能希望在有一定数量的样本集上模型是校准的。

> 例如，某个测试集中各样本的置信度分别为{0.33，0.33，0.33，0.67，0.67，0.67，1.00，1.00}，当3个置信度为0.33的样本里有1个样本预测正确时，在置信度为0.33的样本构成的集合上，模型的准确率也是0.33，和置信度相匹配。同理，我们希望在3个置信度为0.67的样本里有2个预测正确，准确率与置信度均为67%；置信度为1.00的样本全部预测正确，准确率和置信度均为100%。

上面给出的例子能直观地说明模型的置信度与准确率相匹配的情况，但是并不符合实际，因为实际中几乎不会出现置信度完全一致的样本。举一个看起来更实际的例子，测试集中各样本的置信度为{0.302，0.331，0.358，0.645，0.673，0.698，0.996，0.998}。这样情况下，已经没有确切地为0.33和0.67置信度的样本。对此，我们可以取一些区间，统计置信度落入区间的样本，这种方法我们称为**分箱**（binning）。

> 例如，我们取区间[0.30,0.36)作为一个“箱子”（bin），可见，置信度落在这个区间的样本有3个，他们的置信度分别为{0.302，0.331，0.358}，置信度均值约为0.33，那么如果模型是校准的，则这3个样本中有1个预测正确，准确率也为0.33。同理，我们可以取[0.64,0.70)为第二个箱子，[0.94,1.00)为第三个箱子。

最常见的分箱方法是均匀地划分$[0,1]$区间来分箱，比如当我想划分10个箱子时，就取$[0,0.1)$为第一个箱子，$[0.1,0.2)$为第二个箱子，$[0.9,1]$为第10个箱子。另一种分箱方法是样本数均匀划分，使每个箱子里的样本数相同。

通过将置信度分为m个分箱，分类模型的校准可以公式化地表示为：

$$p\left(B_i\right) \rightarrow accuracy\left(B_i\right), \forall i \in \{1,2,\cdots,m\}$$

分类问题中的校准性能指标

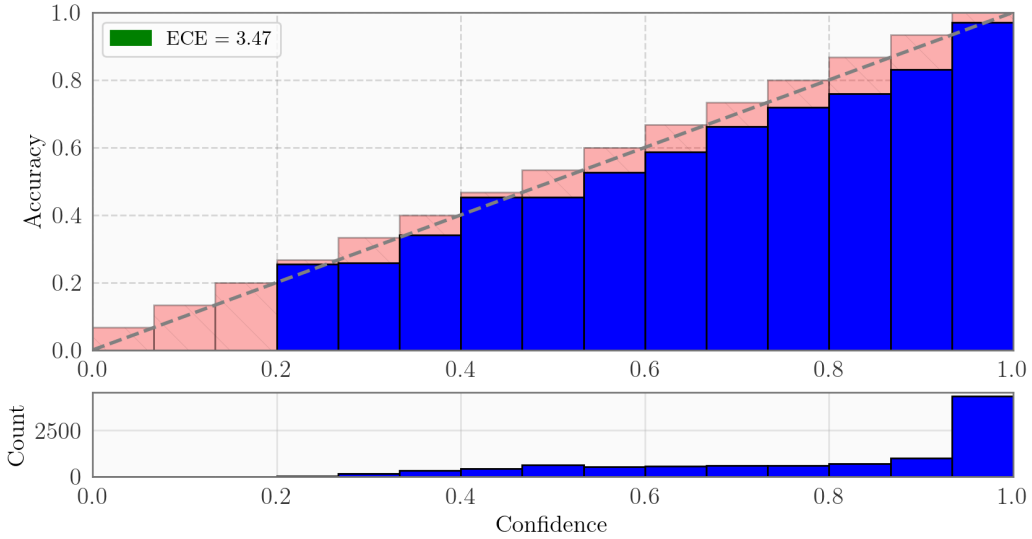
最常见的校准性能指标为期望校准误差（Expected Calibration Error，ECE），通过将置信度分为m个分箱，ECE可以通过下式估计：

$$\widehat{\mathrm{ECE}}=\sum_{i=1}^{m} \frac{\left|B_i\right|}{n}\left|\operatorname{accuracy}\left(B_i\right)-\operatorname{confidence}\left(B_i\right)\right|$$

其中$accuracy(B_i)$表示第i个分箱中样本的准确率，$confidence(B_i)$表示第i个分箱中样本的平均置信度，$|B_i|$表示第i个分箱中的样本数量，n为总样本数。由于准确率和置信度常用%作为单位，因此ECE也常以%为单位。

可靠性图

可靠性图可以直观地展示分类模型的校准性能。例如下图：



该图有两个部分，上面的为可靠性图，下面的为样本数柱状图。

可靠性图的纵坐标为准确率，横坐标为置信度，图中将$[0,1]$的置信度等宽分成了15个分箱。理想的可靠性曲线为灰色的斜线，每个箱子对应的理想准确率用红色标识出。

下方的柱状图给出了每个分箱中的样本数，可见高置信度的样本最多，随着置信度降低，样本数量也随之减少。

该可靠性图是十分类问题中画出来的。十分类问题中，所预测的类别对应的置信度一定不低于10%，因此低于10%置信度的箱子中一定不存在样本。

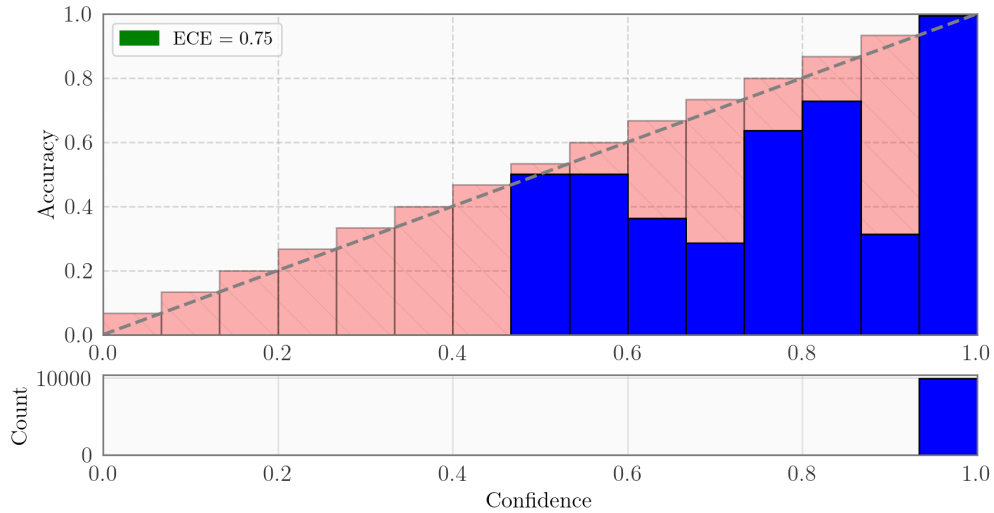
分类问题仿真

MNIST数据集仿真

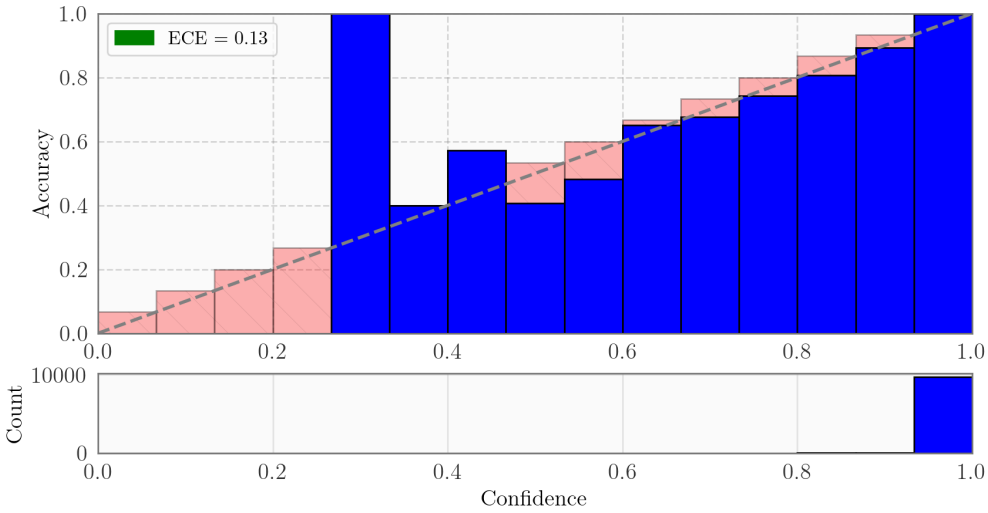
利用LeNet5网络，测试集上准确率达到**99.01%**。

温度缩放校准前后的可靠性图如下：

- 校准前：



- 校准后：



我们发现，模型的准确率很高，接近100%时，本身的校准性能已经不错。利用温度缩放校准也可以进一步提升校准性能。

> 可以试想，当模型的准确率为100%时，就不存在校准的问题了，因为对所有的样本模型都能精准预测。当模型准确率接近100%，其校准性能也会不错，对后校准的要求比较小。

Cifar10数据集仿真

对Cifar10数据集进行仿真，对照文献中的结果。均采用LeNet5网络，利用保序回归和温度缩放进行校准，校准集大小为5000，采用ECE作为指标（越小越好）。

校准性能验证

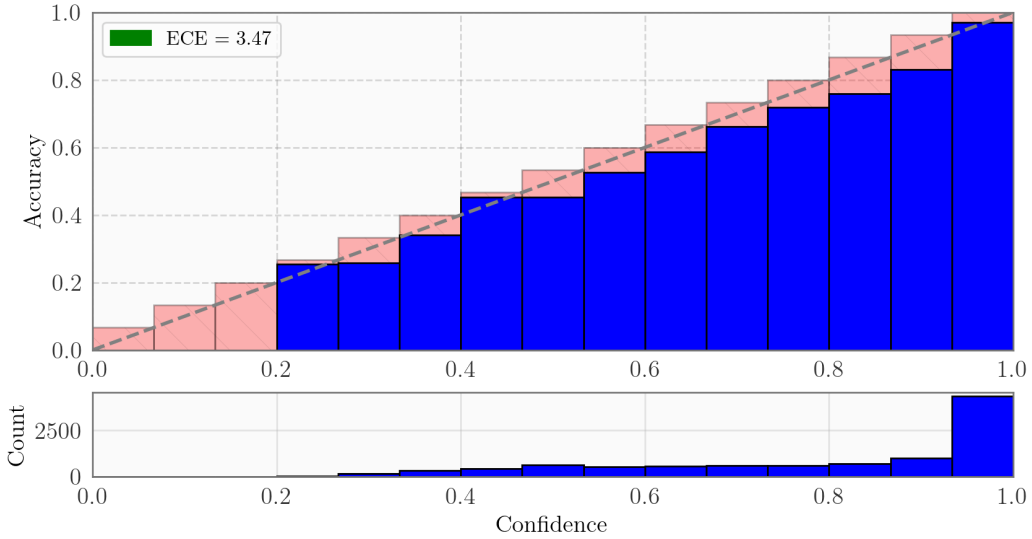
	校准前	保序回归	温度缩放
文章[1]	3.02	1.85	0.93
我	3.47	1.59	0.81

> [1]"On Calibration of Modern Neural Networks"

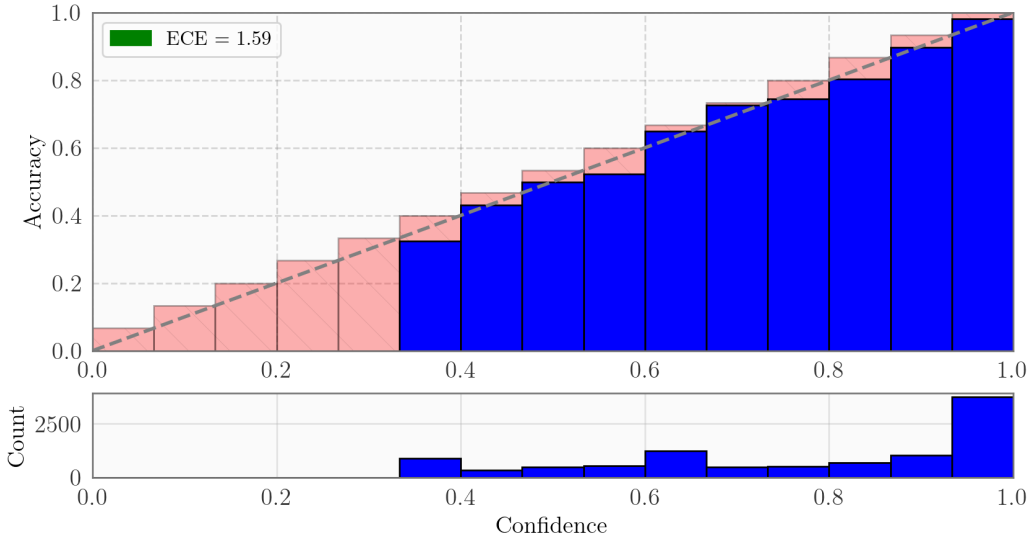
仿真得到的校准结果与论文结果大致吻合。

可靠性图

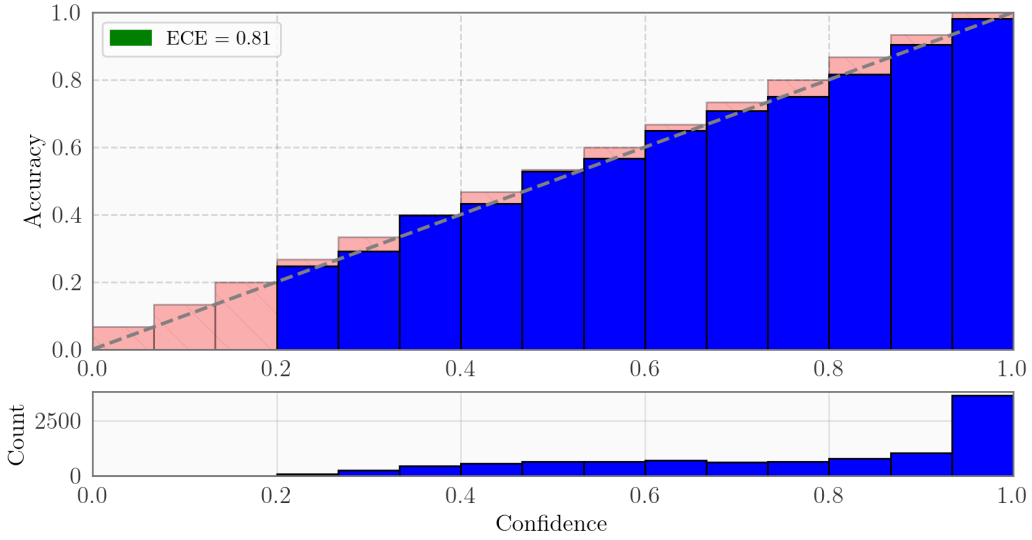
- 校准前：



- 保序回归：



- 温度缩放：

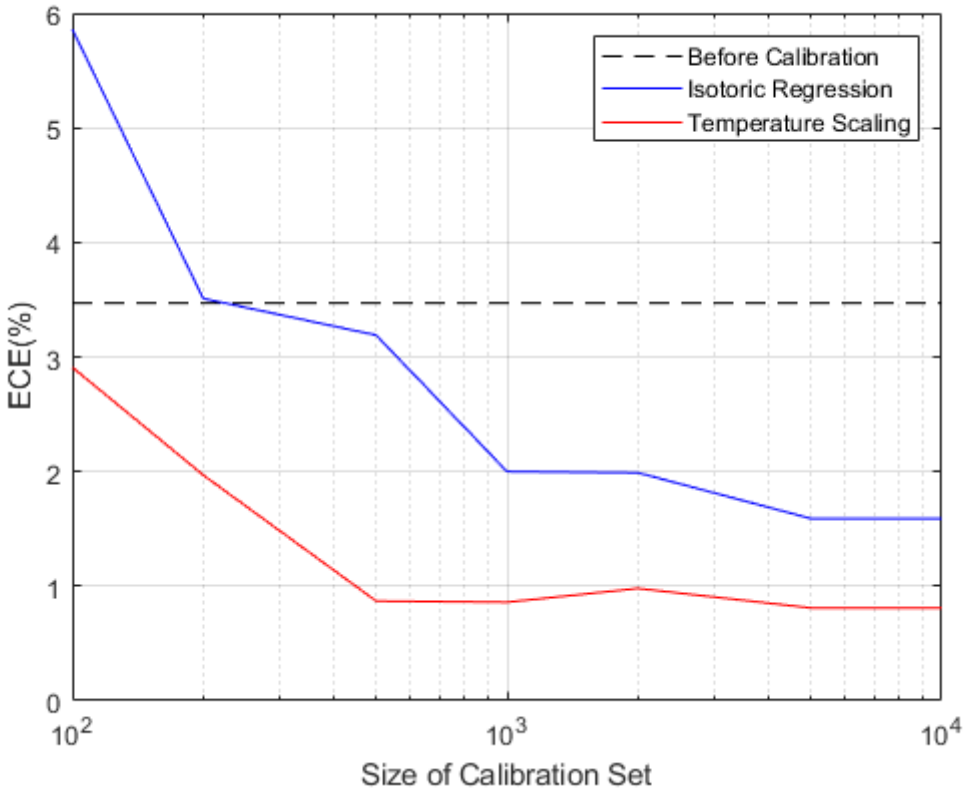


从仿真结果可见，校准前的模型过自信，通过保序回归和温度缩放可以缓解过自信，提升校准性能。

校准集大小的影响

校准的质量受到校准集的大小影响。下面是ECE随着校准集大小变化的趋势：

校准集大小	0	100	200	500	1000	2000	5000	10000
保序回归	3.47	5.86	3.51	3.19	2.00	1.99	1.59	1.59
温度缩放	3.47	2.91	1.97	0.87	0.86	0.98	0.81	0.81



加入不确定性度量方法

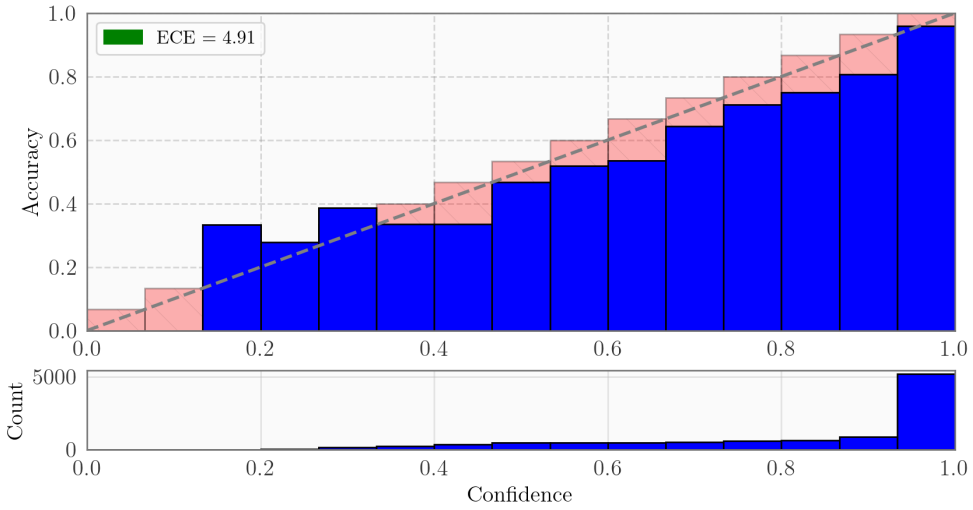
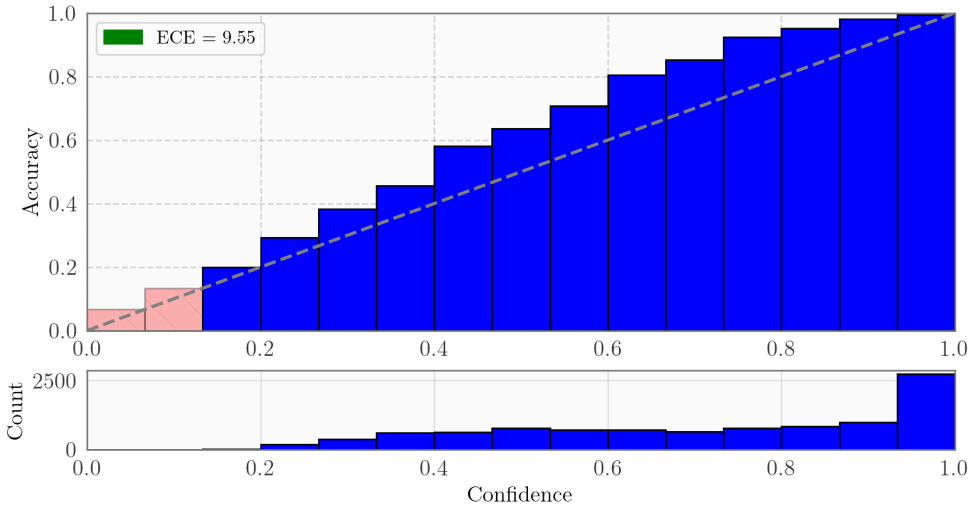
分类问题可依靠softmax得到置信概率，并不一定需要不确定性度量的机器学习算法。但是我们依然使用贝叶斯神经网络、MCDropout或集成方法来训练模型。

下面是BNN和集成方法，校准集大小为5000的校准性能结果：

ECE（%）	LeNet5	贝叶斯LeNet5	集成LeNet5
校准前	3.47	4.91	9.55
温度缩放	0.81	1.33	1.01
保序回归	1.59	1.79	1.09

仿真结果表明，利用BNN或集成方法不能提升校准性能。但在准确率上有所提升：

LeNet5	贝叶斯LeNet5	集成LeNet5
77.00%	78.87%	81.77%

> 需要补充的一点是，BNN模型依然表现为过自信，而集成方法训练得到的模型表现为欠自信。因此，虽然集成模型的ECE较大，但更容易校准，校准后的性能比BNN更好。
>
> BNN：
>
> 
>
>
> 集成：
>
> 
>

总结

1. 利用softmax的DNN得到的常表现为过自信，利用温度缩放或保序回归进行校准，可以缓解过自信，提升校准性能。
2. 在分类问题中，温度缩放的校准性能常比保序回归更好。
3. 校准集的大小会影响校准性能。在校准集较小时，温度缩放依然能够一定程度提升校准性能，而保序回归则有可能恶化校准性能。
4. 在分类问题中可采用softmax直接获得置信度，BNN和集成等不确定性度量方法不是必须的，并且应用这些方法也不能有效提升校准性能。
5. 虽然BNN等不确定性度量方法不能提升有效提升模型校准性能，但是可能提升模型准确率。

回归模型的置信度预测和校准

回归问题的置信度

下面我们从不确定性的角度考虑回归问题。由于数据和模型存在不确定性，我们可以设想：预测值实际上是一个随机变量，服从某个分布，而模型给出的预测只是在该分布上的一次采样。

预测值作为随机变量，对于某个区间，我们可以给出一次采样落在该区间的概率，这就是概率统计中**置信区间**和**置信概率**。例如，当已知均值和标准差时，正态分布常用的双侧置信区间与置信概率如下表：

置信区间	$[\mu-\sigma,\mu+\sigma]$	$[\mu-1.64\sigma,\mu+1.64\sigma]$	$[\mu-1.96\sigma,\mu+1.96\sigma]$	$[\mu-3\sigma,\mu+3\sigma]$
置信概率	68.27%	90%	95%	99.73%

在没有可用的先验信息时，我们一般假设预测值服从正态分布。这样，我们只要知道了预测值的均值和方差，就能计算不同置信区间下的置信概率。

> 注意，每个置信区间对应一个置信概率，但一个置信概率可以对应很多种置信区间。比如，除了上表展示的双侧置信区间，我们还可以使用单侧置信区间。
>
> 如何选择置信区间是与任务相关的。比如，想要预测某个零件的尺寸，尺寸应该既不能过大也不能过小，此时就用双侧置信区间较好；如果想要预测的是用户掉线率，我们希望预测的掉线率不低于实际掉线率（保守估计），此时用单侧置信区间较好。

度量不确定性的机器学习方法

为了能够估计预测值的分布，有以下几种度量不确定性的机器学习方法：

- 多个网络、多次推理：集成学习
- 单个网络、多次推理：贝叶斯神经网络、MCDropout
- 单个网络、多次推理：先验神经网络、证据神经网络

集成学习的思路是：训练多个网络，通过多个网络的预测估计预测值的分布。

贝叶斯神经网络和MCDropout的思路是：将模型权重看成随机变量，通过对权重的多次采样，获得多次预测值来估计预测值分布。

先验神经网络和证据神经网络的思路是：直接假设预测值服从的分布形式，通过预测分布的控制超参数来估计预测值分布。

回归问题的准确率

假设我们现在已经得到了一个可以给出预测值分布的模型，我们如何判断模型给出的分布是否合适呢？

> 如果一个模型，在其给出的置信度为95%的置信区间内（比如$[\mu-1.96\sigma,\mu+1.96\sigma]$），却只观测到了80%的样本落在这个区间，那么我们可以说这个模型过于自信。
>
> 这个例子中，模型在$[\mu-1.96\sigma,\mu+1.96\sigma]$这个区间上的准确率为80%，而置信度是95%，准确率与置信度不吻合，我们就可以认为模型不是校准的。

通过上面这个例子可以直观地展示回归问题中的**准确率**与**置信度**。下面公式化地给出定义：

记T个样本的校准集（Calibration Set）$S=\{(x_t,y_t)\}_{t=1}^{T}$，模型对于样本$x_t$预测其标签的累计概率分布(CDF)为$F_t$。对于置信概率为$p$的某个置信区间$I(F_t,p)$，可将准确率accuracy定义为：

$$accuracy(p)=\frac{\sum_{t=1}^{T}1\left(y_t\in I\left(F_t,p\right)\right)}{T}$$

其中$1\left(\cdot\right)$为指示函数。如果我们利用单侧置信区间$(-\infty,F_t^{-1}(p)]$，就可将准确率accuracy定义为：

$$accuracy(p)=\frac{\sum_{t=1}^{T}1\left(y_t\leqslant F_t^{-1}\left(p\right)\right)}{T}$$

同理，我们也可以利用双侧置信区间来定义准确率。

我们校准的目标可以写为：

$$p \rightarrow accuracy(p), \forall p \in [0,1]$$

注意：准确率由置信区间内的样本数决定，是需要测试得到的。而置信度和置信区间是对应关系，可以直接换算。

> 我本来想用“命中率”这个词来表示回归问题中的“准确率”，二者的英文都是accuracy，我觉得这个词更形象。
>
> 用投篮打个比方：对于直径50cm的篮筐，我可以说有信心10球投进5个，置信度为50%，而实际测试中我10个只进了3个，命中率为30%；然后换了个直径80cm的篮筐，我说有信心10球进8个，而实际测试中我只进了6个。这说明我对我的投篮能力过于自信了，自己的信心和能力不匹配，需要校准。校准的任务就是，对于任意大小的篮筐，我的自信和能力能够匹配的上。
>
> 回归问题中，置信区间就好比“篮筐”，样本就好比“篮球”，模型预测就好比投篮这个动作，如果投出去的“球”在置信区间这个“篮框内”，那么就“命中个数”就+1，最终“命中率”就是“命中总数”除以“篮球个数”。回归模型的校准，就是对于任意大小的置信区间，模型的“自信”和“命中率”都能够匹配。

回归问题中的校准性能评估指标

我们常采用均方校准误差(Mean Square Calibration Error, MSCE)来评估校准性能：

$$MSCE = \frac{1}{m}\sum_{j=1}^{m}(p_j - accuracy(p_j))^2$$

注意：MSCE是评估校准性能的指标，需要利用到预测值分布来计算不同置信区间下的置信度与准确率。而MSE通常是指评估模型预测精度的指标，其误差的计算是根据预测（均）值减去标签值，不需要利用预测值分布的二阶以上信息。

回归问题仿真

非时间序列回归问题仿真

对UCI公开数据集中的wine数据集和crime数据集进行仿真，对照文献中的结果。均采用MCDropout，网络结构一致，采用单侧置信区间，利用保序回归进行后校准，采用均方校准误差作为指标（越小越好）。

数据集	文章[1]（校准前）	我（校准前）	文章[1]（校准后）	我（校准后）
wine	0.096	0.068	0.028	0.036
crime	0.070	0.057	0.015	0.033

> [1] "Accurate Uncertainties for Deep Learning Using Calibrated Regression"

仿真结果显示，我训练的模型在校准前的校准性能好于文章，校准后的校准性能差于文章。但经过校准，我的模型校准性能也有明显的提升。

分析可能是以下原因造成结果不完全一致：

- 数据预处理可能不一致。在crime数据集中，有一些属性含有缺省值，我的处理方法是将含有缺省值的属性删除。我采用的是z-score归一化，且没有进行特征选择或数据降维。在这些具体的处理上文章都并未提及。
- 超参数选择不一致。文章并未提及训练超参数的选择，而我也没有进行过多的调参。
- 训练集和测试集的划分不一致。数据集划分的随机性会导致结果不同。

单侧置信区间vs双侧置信区间

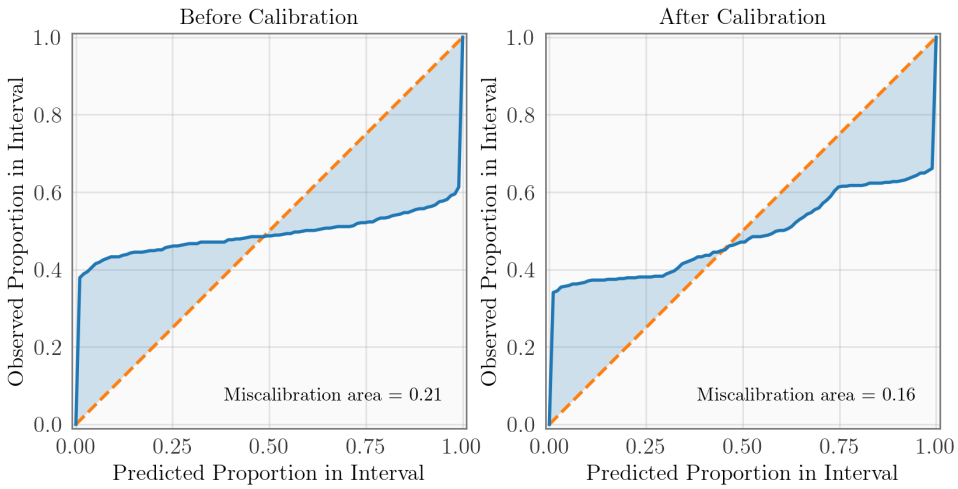
我们同样可以利用双侧置信区间来评估准确率，并构造再校准数据集进行校准。以下是在双侧置信区间下，校准前后的均方校准误差MSCE。

数据集	双侧置信区间，校准前	双侧置信区间，校准后
wine	0.185	0.066
crime	0.227	0.124

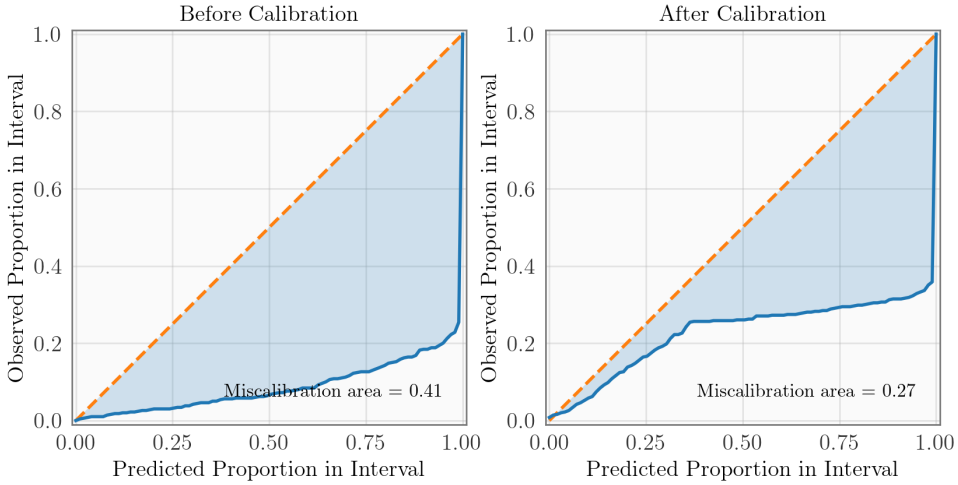
不论采用单侧置信区间还是双侧置信区间，我们发现保序回归都可以有效降低了校准误差。一个有趣的现象是，利用双侧置信区间来评估准确率后计算得到的均方校准误差要比单侧置信区间大。

下面以crime数据集为例，对比单侧和双侧置信区间的可靠性曲线：

> 单侧置信区间可靠性曲线：



> 双侧置信区间可靠性曲线：



- **采用单侧置信区间时，校准前的可靠性曲线几乎中心对称**

我们发现采用单侧置信区间时，校准前的可靠性曲线是几乎是中心对称的。这是因为：

$$accuracy\left(p\right)+accuracy\left(1-p\right)=\frac{1}{T}\sum_{t=1}^{T}1\left(y_t\leqslant F_t^{-1}\left(p\right)\right)+1\left(y_t\leqslant F_t^{-1}\left(1-p\right)\right)$$

只要标签的分布是近似对称的，那么

$$\begin{aligned} accuracy\left(p\right)+accuracy\left(1-p\right) &= \frac{1}{T}\sum_{t=1}^{T}1\left(y_t \leqslant F_t^{-1}\left(p\right)\right)+1\left(y_t \leqslant F_t^{-1}\left(1-p\right)\right) \\ &\approx \frac{1}{T}\sum_{t=1}^{T}1\left(y_t \leqslant F_t^{-1}\left(p\right)\right)+1\left(y_t \geqslant F_t^{-1}\left(p\right)\right) \\ &= 1 \end{aligned}$$

而双侧置信区间的可靠性图没有这种近似对称的性质，这直观说明了采用双侧置信区间得到的校准误差数值为何比单侧置信区间大。

我们同时发现，在保序回归校准后，单侧置信区间的可靠性图的近似对称被破坏。这是因为保序回归将置信度映射到一个新的值，不保证对称性质不变。

> 单侧置信区间的可靠性曲线近似中心对称的性质需要标签的分布近似对称才成立。我猜想，更一般的结论是单侧置信区间的可靠性曲线会与理想斜线至少有一个交点。

- **过自信对后校准产生的影响**

观察到双侧置信区间的可靠性曲线总是在理想曲线下方，可以说明模型是过于自信的。并且我们发现，后校准似乎仅改善了置信度在$[0, 0.35]$范围内的置信度。这是由于模型过于自信，在模型给出几乎100%的置信度时，实际的准确率仅有约35%。这就导致再校准数据集中，标签的范围只有$[0, 0.35]$，因此保序回归模型也只能校准低于0.35的置信度。

若模型是欠自信的，则不会出现上面这种情况。如下面这个例子：

> 一个欠自信的模型，在给出的置信度为80%时，对应的准确率已经接近100%。那么在再校准数据集中，置信度范围为$[0, 0.8]$，而标签的范围为$[0, 1]$，保序回归模型可以有效校准任意的置信度。

下面我们将看到对于欠自信的模型，保序回归能够发挥更好的作用。

利用随机不确定性度量提升校准性能

我考虑通过度量随机不确定性来改进校准性能，其只需要添加一个输出作为对随即不确定性的度量即可（增加的参数量几乎可忽略不计）。

> | 数据集 | 文章[1]（校准前） | 我（校准前） | 文章[1]（校准后） | 我（校准后） |
> | --- | --- | --- | --- | --- |
> | wine | 0.096 | 0.021 | 0.028 | 0.006 |
> | crime | 0.070 | 0.006 | 0.015 | 0.012 |
>
> [1] "Accurate Uncertainties for Deep Learning Using Calibrated Regression"

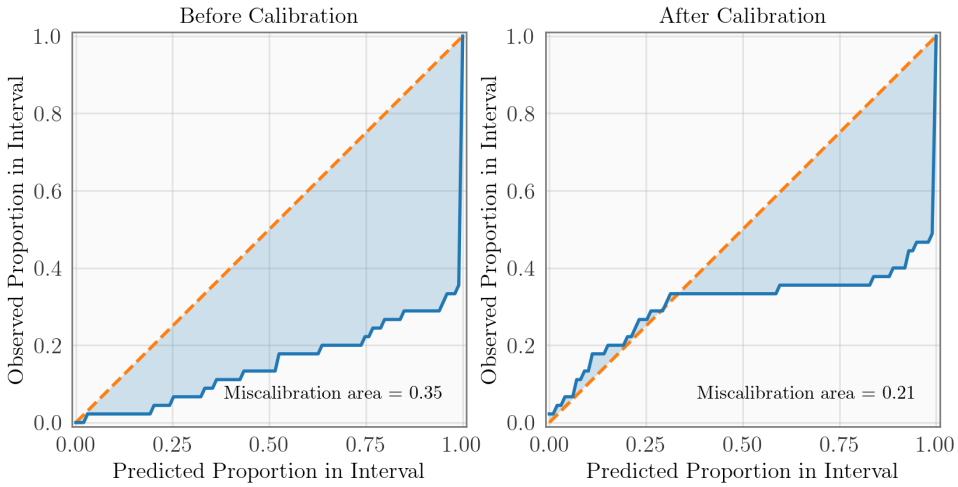
可见通过度量随机不确定性，即使不进行后校准，校准性能已超过文章校准后的性能。在crime数据集上，校准前的校准性能已经极好，再进行校准反而降低了校准性能。

在预测精确性方面，通过RMSE来评估度量随机不确定性对预测精确性的影响，仿真显示度量随机不确定性没有负面影响：

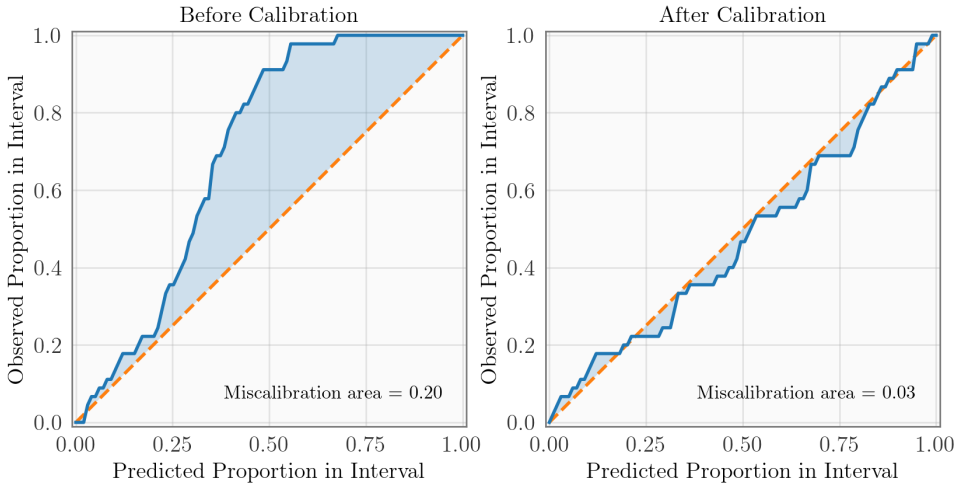
> | 数据集 | 不度量随机不确定性 | 度量随机不确定性 |
> | --- | --- | --- |
> | wine | 0.163 | 0.122 |
> | crime | 0.576 | 0.578 |

下面给出wine数据集仿真中，度量和不度量随机不确定性的可靠性曲线，均采用双边置信区间：

> **不度量随机不确定性**：



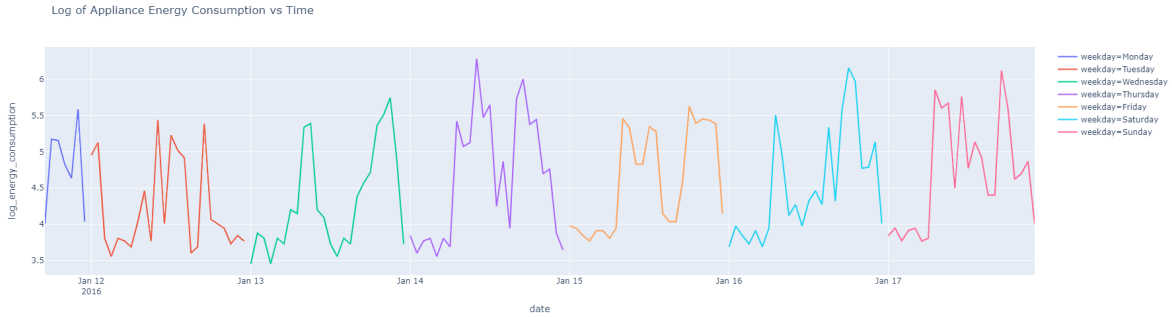
度量随机不确定性:



仿真结果展示了度量随机不确定性的效果：通过度量随机不确定性，解决了模型过于自信的问题，保序回归在模型欠自信时可以更好地校准模型。

时间序列回归问题仿真

下面对UCI上的一个能源预测数据集进行仿真。数据集为实际测量某地每个小时消耗的能源，时序数据示意图：

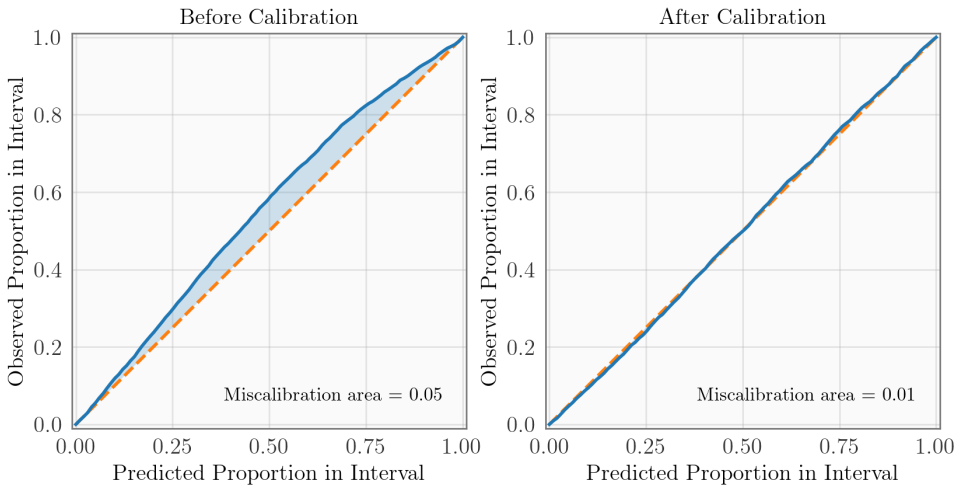


我们考虑用过去10小时数据，预测未来10小时数据。

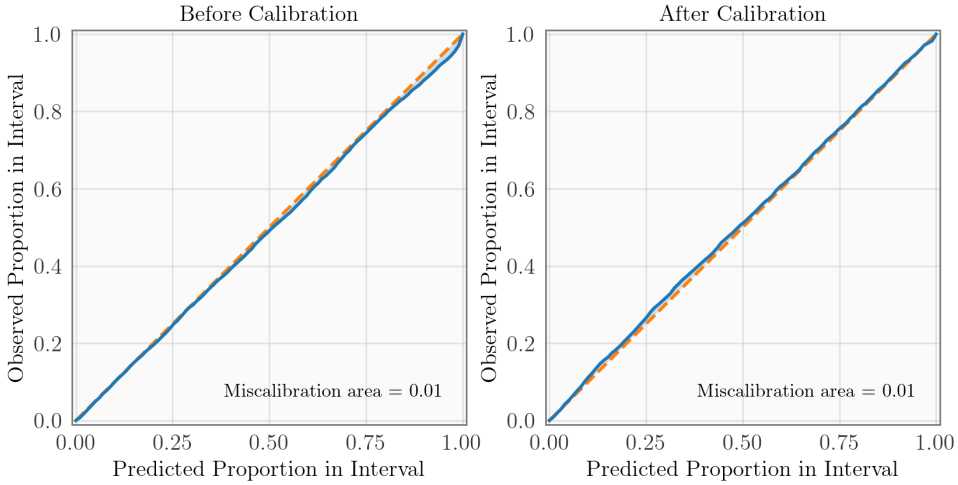
可靠性图

下面我们给出MCDropout、集成和证据神经网络，保序回归校准前后的可靠性曲线：

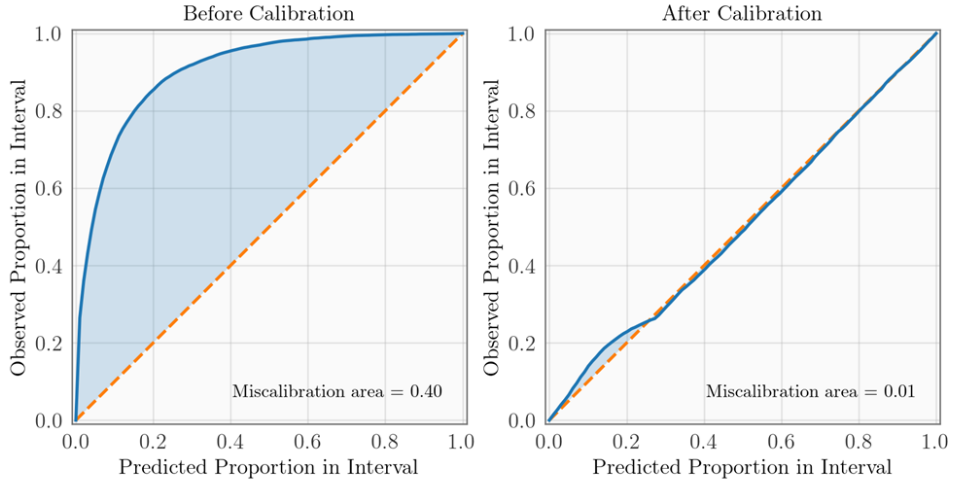
- MCDropout



- 集成



- 证据神经网络

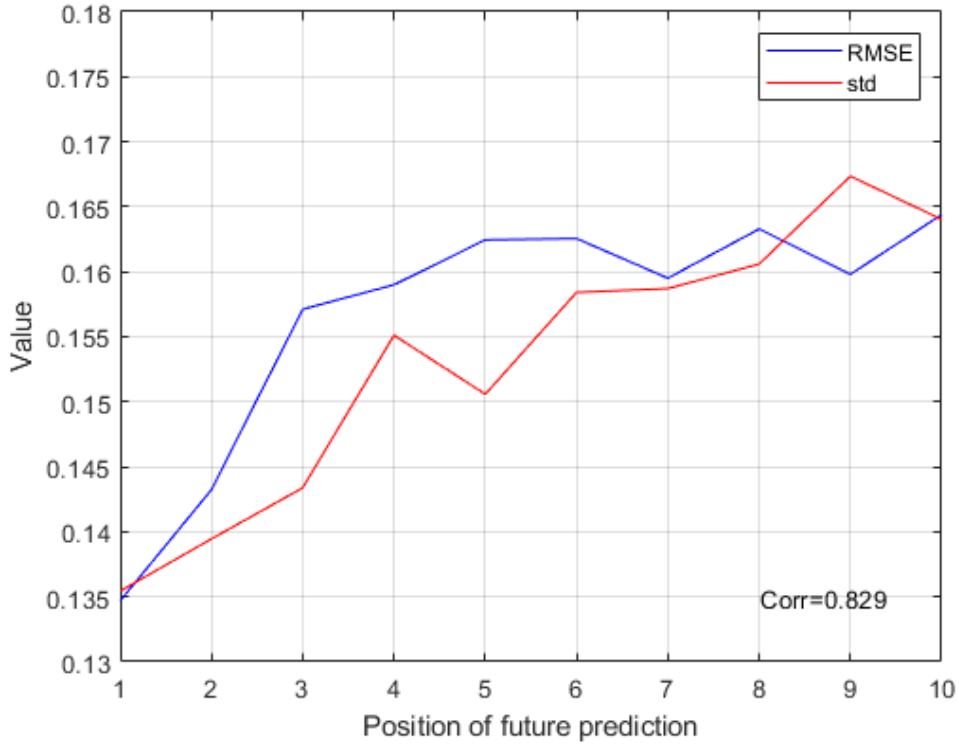


在校准前，集成方法得到的校准性能最好，证据神经网络欠自信。利用保序回归校准后三个模型校准性能都很好。

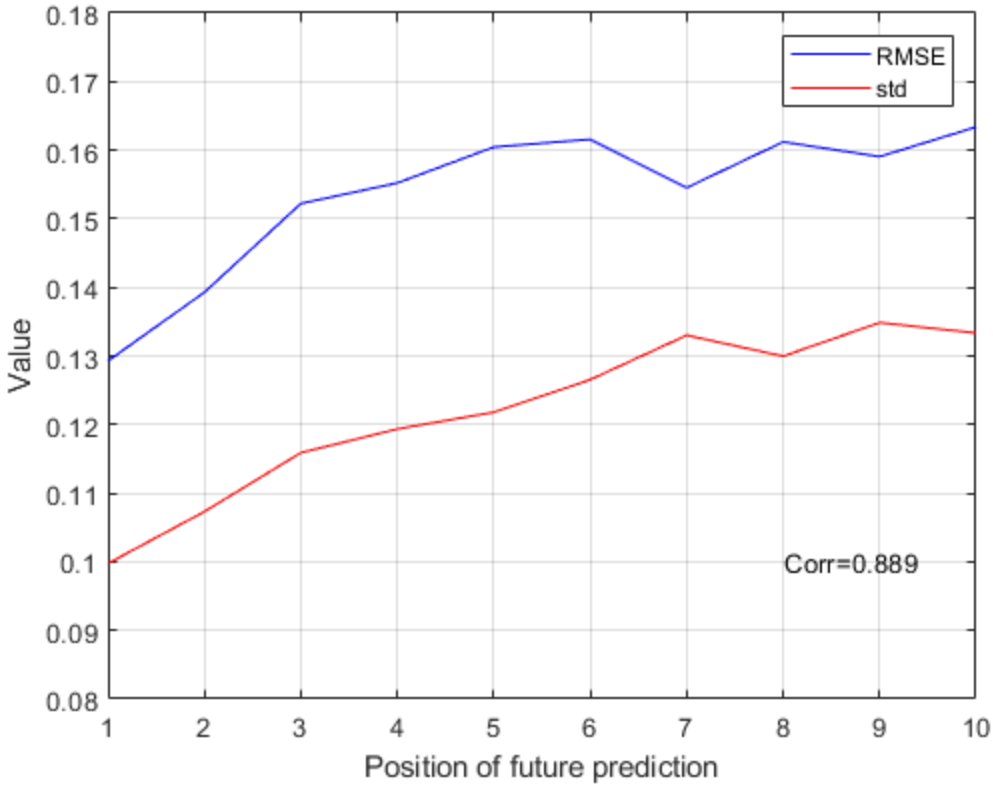
时序预测的精度与不确定性

在时序预测任务中，可想而知，对于越是未来的信息预测精度越差。那么对于越为未来的预测，模型给出的不确定性是否也越大呢？下面几张图中，横坐标表示为未来的第几个小时，曲线给出了关于预测精度（RMSE，蓝线）和不确定性（std，红线）结果。

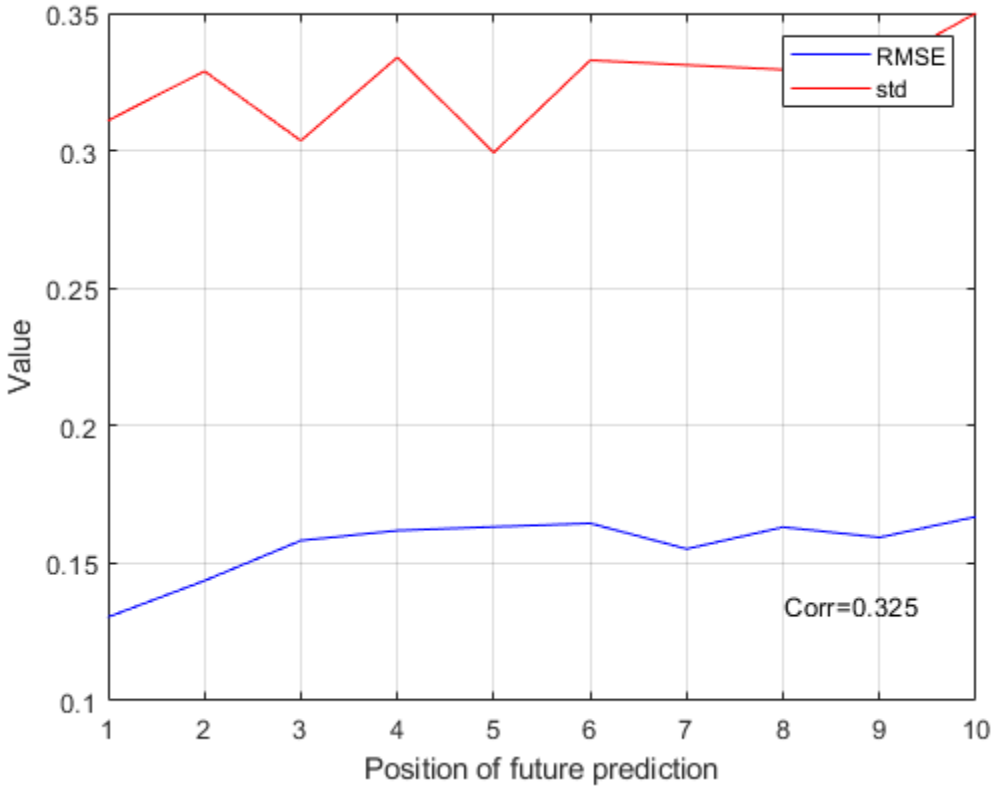
- MCDropout，校准前



- 集成，校准前



- 证据神经网络，校准前



仿真结果显示，模型的预测精度关于小时数下降，验证了预测的越远、精度越差的直觉。并且，对于MCDropout和集成方法而言，预测不确定性和预测精度的相关性高；但证据神经网络的预测不确定性则与预测精度相关性较差。

> 从校准前的MCDropout、集成和证据神经网络可靠性曲线可知，MCDropout和集成的校准性能较好，而证据神经网络较差。这可能是MCDropout和集成的预测不确定性（std）与预测精度（RMSE）相关性高，而证据神经网络相关性不高的直接原因。
>
> 如果对于校准较差的模型，想要建立预测标准差与RMSE之间的强相关性，一种直接的想法是对模型进行校准。但是，保序回归校准的对象是置信度，而不是标准差，无法直接得到校准之后的预测标准差。因此，在回归问题中，如何校准预测标准差可能是个值得研究的问题。

总结

- 为了后校准回归模型，我们需要构建再校准数据集。选择单侧或双侧置信区间可以根据实际任务决定。
- 神经网络通常表现为过于自信，即给出的置信度高于准确率。在过于自信时，后校准方法通常无法很好地校准较高的置信度。
- 度量随机不确定性是一种简单有效的避免模型过于自信的方法。
- 保序回归能有效校准欠自信的回归模型。
- 良好校准的模型给出的预测标准差，与预测精度RMSE高度相关。因此，良好校准的模型给出的预测不确定性可以间接反映预测精度。