Bayesian deep learning

一、贝叶斯建模

给定一个训练数据，$X = \{x_1, \ldots, x_N\}, Y = \{y_1, \ldots, y_N\}$，对于回归任务来说，如果是频率学派，可以通过最小化均方误差的方式得到拟合函数：$y = f^w(x)$。但是对于贝叶斯学派来说，则先假设模型的参数w服从某种先验分布，然后通过数据，来矫正参数的分布，得到参数的后验分布。最终可以通过参数的后验分布进行推断。

贝叶斯推断需要假设一个likelihood分布，举例来说，对于**分类**任务而言，我们可以假设softmax likelihood。

$$p\Big(y = d|\mathbf{x}, \boldsymbol{\omega}\Big) = \frac{\exp(f_d^{\boldsymbol{\omega}}(\mathbf{x}))}{\sum_{d'} \exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))}$$

对于**回归**，我们可以假设Gaussian likelihood。

$$p\Big(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}\Big) = \mathcal{N}(\mathbf{y}; \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}), \tau^{-1} I)$$

对于给定的数据集，X, Y，我们需要去寻找参数的后验分布（可以通过贝叶斯公式得到）。如下。

$$p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\omega}) p(\boldsymbol{\omega})}{p(\mathbf{Y}|\mathbf{X})}.$$

得到了后验分布之后，对于一个新的样本点，可以通过积分得到其预测 ：

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) \mathrm{d}\boldsymbol{\omega}.$$

这叫做贝叶斯**推断**。

在后验评估中有一个很重要的一项，叫做model evidence：

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\omega}) p(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}.$$

对于简单的贝叶斯线性模型来说，这个积分可以比较容易求得。但是对于相对复杂的模型，这个时候就不容易求积分，这个时候就需要通过近似方法来得到。

二、变分推断

参数的后验分布$p(\omega|X,Y)$通常不容易获得，同样也需要通过近似方法得到。变分推断通过一个结构简单的variational distribution $q_\theta(\omega)$来进行近似这个后验分布。而KL divergence是两个分布相似性的度量，通过最小化KL divergence可以求解这个$q_\theta(\omega)$。

$$\mathrm{KL}(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})) = \int q_\theta(\boldsymbol{\omega}) \log \frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})} \mathrm{d}\boldsymbol{\omega}.$$

用$q_\theta^*(\omega)$表示优化结果，则预测分布可以表示为：

$$p(\mathbf{y}^*|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) \approx \int p(\mathbf{y}^*|\mathbf{x}^*,\boldsymbol{\omega}) q_\theta^*(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} =: q_\theta^*(\mathbf{y}^*|\mathbf{x}^*).$$

但是这种方法不适用于大规模数据，因为计算需要遍历整个数据集$\int q_\theta(\omega) log p(Y|X,\omega) d\omega$。

三、什么是贝叶斯深度学习

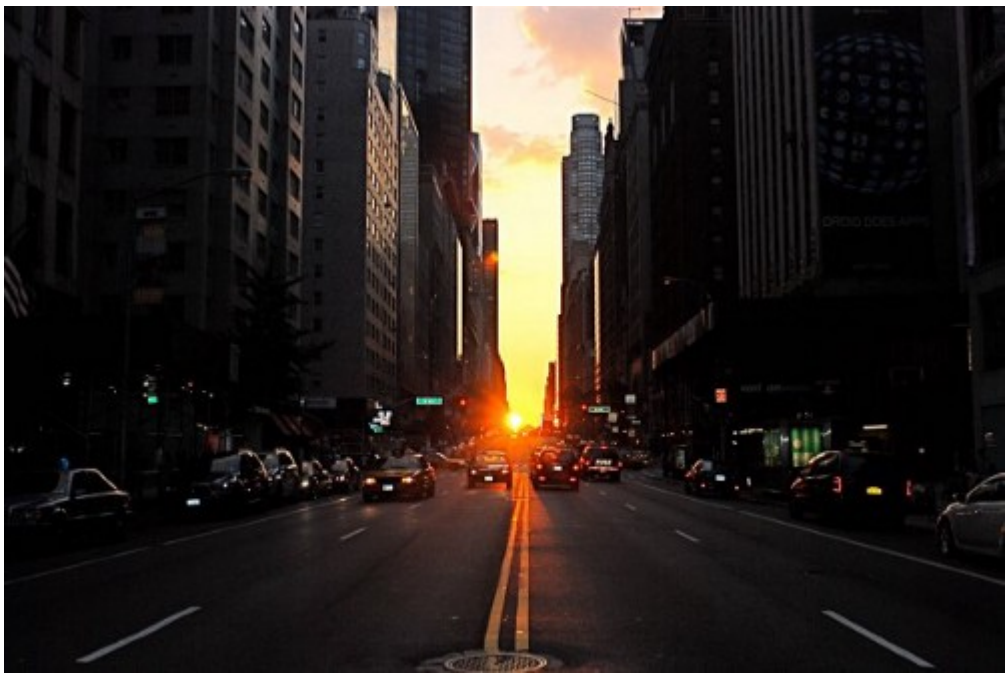
顾名思义，贝叶斯统计+深度学习就是贝叶斯深度学习。传统的贝叶斯神经网络的做法是给网络的参数加上一个先验。但是由于深度学习网络的参数通常很多，这种方法通常需要大量计算，因此在实践中通常不可行，因此近来的方法试图不改变网络参数数量的情况下引入不确定性。

四、不确定性

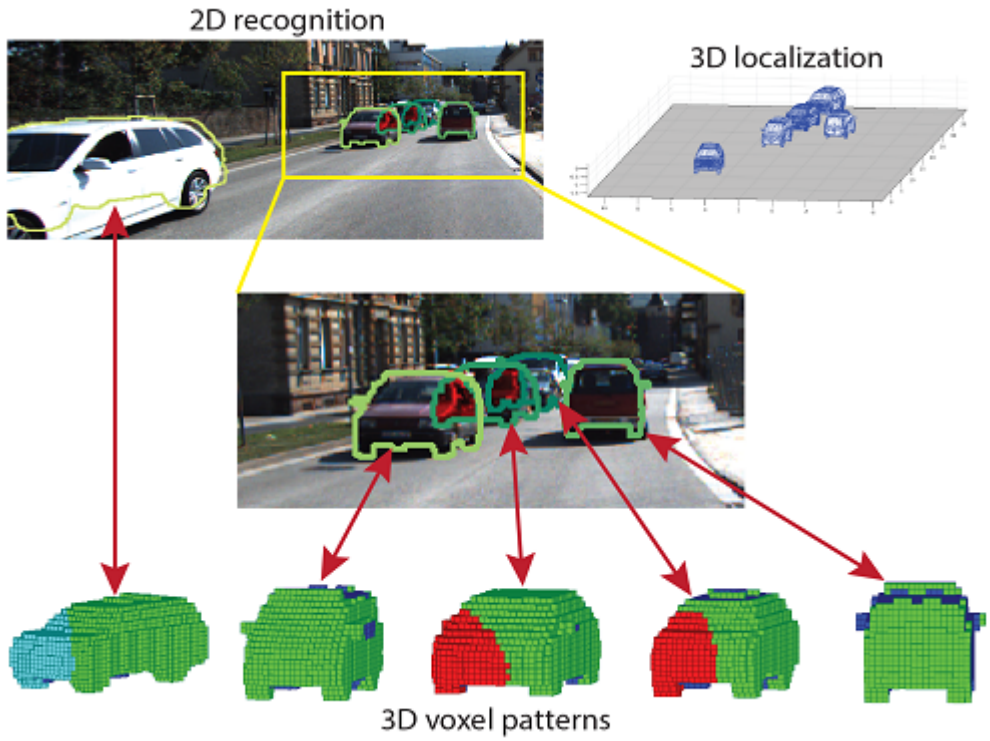
不确定性分为两种，Epistemic uncertainty和Aleatoric uncertainty。

1. Aleatoric uncertainty

Aleatoric uncertainty（任意不确定性）捕捉了观测（observation）中固有的噪声。比如sensor采集数据过程中的噪声。这种不确定性无法通过收集更多的数据来减少。任意不确定性的例子：occlusions, lack of visual features and under/over exposure.



过曝或者欠曝



遮挡

缺少视觉特征

事实上，有两种类型的不确定性，**异方差**和**同方差**，但一般值涉及异方差的不确定性。

2. Epistemic uncertainty

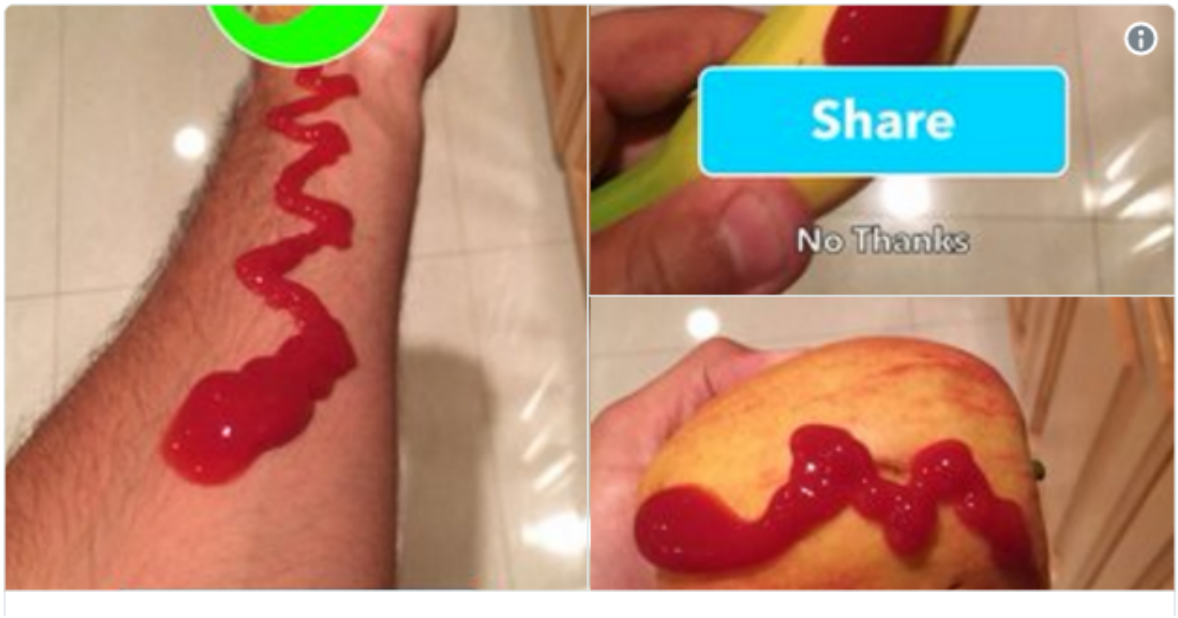
Epistemic uncertainty（认知不确定性）告诉我们模型不知道什么，这种不确定性可以用足够的数据来解释，通常被称为模型不确定性。

例子1.

观察认知不确定性的一个简单方法是在25%的数据集上训练一个模型，在整个数据集上训练第二个模型。仅在25%的数据集上训练的模型将比在整个数据集上训练的模型具有更高的认知不确定性，因为它看到的数据更少。

例子2

Not Hotdog APP。这个模型在正常情况下表现良好。但是如果你在腿上涂上番茄酱，然后拍照给这个程序判断，它很容易判断成是热狗。这是因为模型从来没有训练过这些数据，所以出现了误判。贝叶斯深度学习模型将预测在这些情况下的高认知不确定性。



五、不确定性建模

传统的方法基本上都是独立地对认知不确定性和任意不确定性进行评估。认知不确定性主要是通过对参数添加先验分布的方法进行建模，而任意不确定性则是对模型的输出施加一个分布进行建模。下面进行具体解释。

1.认知不确定性在深度学习中的建模

通过给网络参数施加一个先验分布的方法实际上就是贝叶斯神经网络(BNN)。贝叶斯神经网络的参数是不确定的，满足某种分布，这和参数确定的神经网络是不一样的。贝叶斯神经网络在预测阶段实际上是在所有可能参数上进行加权平均的一个结果。假设BNN的随机输出用$f^W(x)$表示，模型的likelihood为$p(y/f^W(x))$，给定一个数据集$X=\{x1,...,x_N\},Y=\{y_1,...,y_N\}$，贝叶斯推断需要计算参数的后验$p(W/X,Y)$。

对于回归任务来说，通常假定likelihood为高斯分布，$p(y/f^W(x))=N(f^W(x);\sigma^2)$，对于分类任务来说，通常将模型输出经过softmax：$p(y/f^W(x))=Softmax(f^W(x))$。

BNN形式化很容易，但是却很难计算，这是因为在$p(W/X,Y)=p(Y/X,W)p(W)=p(Y/X)$中，$p(Y/X)$很难进行评估。有很多方法对这个分布进行近似，这些方法都是通过$q_\theta^*(W)$去拟合$p(W/X,Y)$。

然而，在实践中，对于复杂的大规模模型，真正可以用的是**Dropout variational inference**。即在训练模型的时候采用dropout，在测试的时候也使用dropout（实际上是一种近似后验）。更为重要的是，dropout可以看做是一种**variational bayesian approximation**，其中近似的分布混合了两个高斯模型（具有小方差，其中一个高斯均值固定为0），具体证明过程略。

通过观察更多的数据来减少权重的认知不确定性。This uncertainty induces prediction uncertainty by marginalising over the (approximate) weights posterior distribution.

对于分类而言，这种近似可以通过monte carlo integration得到：

$$p(y=c|\mathbf{x},\mathbf{X},\mathbf{Y})\approx\frac{1}{T}\sum_{t=1}^{T}\text{Softmax}(\mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x}))$$

其中T表示T次采样，$\hat{W}_t\sim q_\theta^*(W)$, $q_\theta(W)$ 是 Dropout distribution。概率向量p的不确定性可以通过：

$$H(\mathbf{p}) = -\sum_{c=1}^{C} p_c \log p_c.$$

进行评估。

对于回归而言，预测结果表示为：

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x}).$$

相应的不确定性表示为：

$$\text{Var}(\mathbf{y}) \approx \sigma^2 + \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x})^T \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x}_t) - E(\mathbf{y})^T E(\mathbf{y})$$

预测方差中的第一项，σ^2，对应于数据中固有的噪声量。预测方差的第二部分测量模型对其预测有多大的不确定性。

2.Heteroscedastic Aleatoric Uncertainty建模

以上捕捉模型不确定性是通过近似$p(W/X;Y)$来得到。异方差任意不确定性的observation noise与每个输入x有关，对于数据中部分数据比其他数据有更大噪声的情况，异方差模型有非常好的效果。要与每个数据相关，那么可以通过下面这个式子建模：

$$\mathcal{L}_{\text{NN}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\sigma(\mathbf{x}_i)^2}||\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)||^2 + \frac{1}{2}\log \sigma(\mathbf{x}_i)^2$$

3.结合任意不确定性和认知不确定性

1. 认知不确定性：通过dropout variational distribution来近似BNN的后验分布。
2. 任意不确定性：通过模型输出方差值。

回归

对于视觉任务的回归来说：

$$\mathcal{L}_{BNN}(\theta) = \frac{1}{D}\sum_{i} \frac{1}{2}\hat{\sigma}_i^{-2}||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 + \frac{1}{2}\log \hat{\sigma}_i^2$$

其中，D是pixel输出y_i的总数，σ^2是每个pixel i的方差。

在实际中，为了数值的稳定性，改成如下的方式：

$$\mathcal{L}_{BNN}(\theta) = \frac{1}{D}\sum_{i} \frac{1}{2}\exp(-s_i)||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 + \frac{1}{2}s_i.$$

其中，$s_i := log\hat{\sigma}_i^2$。

对于pixel y，不确定衡量：

$$\text{Var}(\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t^2 \tag{9}$$

with $\{\hat{\mathbf{y}}_t, \hat{\sigma}_t^2\}_{t=1}^T$ a set of T sampled outputs: $\hat{\mathbf{y}}_t, \hat{\sigma}_t^2 = \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x})$ for randomly masked weights $\widehat{\mathbf{W}}_t \sim q(\mathbf{W})$.

异方差不确定性应用在分类中

对于分类任务而言，NN对每个像素都进行预测，每个像素输出一个向量f_i，这个向量再通过一个softmax操作形成一个概率向量p_i。我们改变这种方式，对这个向量施加一个高斯分布。

$$\hat{\mathbf{x}}_i|\mathbf{W} \sim \mathcal{N}(\mathbf{f}_i^{\mathbf{W}}, (\sigma_i^{\mathbf{W}})^2)$$
$$\hat{\mathbf{p}}_i = \text{Softmax}(\hat{\mathbf{x}}_i).$$

此处f_i^W, σ_i^W是网络的输出，向量f_i^W混杂了高斯噪声$(\sigma_i^W)^2$，这是一个对角矩阵，每个对角元素都对应一个logit value，加了噪声的向量再通过softmax，得到概率预测。

此时，log likelihood为：

$$\log E_{\mathcal{N}(\hat{\mathbf{x}}_i;\mathbf{f}_i^{\mathbf{W}},(\sigma_i^{\mathbf{W}})^2)}[\hat{\mathbf{p}}_{i,c}]$$

c是每个input i的label。理想情况下上式可以通过求积分得到，但实际上并不好求解，因此考虑使用Monte Carlo integration的方法来进行近似。实际上可以这么做，从logits中进行采样。因此，重新改写以上公式得到：

$$\hat{\mathbf{x}}_{i,t} = \mathbf{f}_i^{\mathbf{W}} + \sigma_i^{\mathbf{W}}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I)$$
$$\mathcal{L}_x = \sum_i \log \frac{1}{T}\sum_t \exp(\hat{x}_{i,t,c} - \log \sum_{c'} \exp \hat{x}_{i,t,c'})$$

附录

分类任务中不确定性的衡量

有三种方式可以衡量分类任务的不确定性：variation ratio，predictive entropy和mutual information。

1.variation ratio

在测试的时候，对于输入x，进行T次实验（每次实验都开启dropout），每次实验的输出为y^t，找出T次实验中所预测类别的众数$c^* = arg\,max_{c=1,...,C}$，以及其对应的数量$f_x = \sum 1[y^t = c]$。则variation ratio的计算为：

$variation\_ratio[x] := 1 - \frac{f_x}{T}$

对于二分类而言，当variation ratio达到0.5的时候，所表示的不确定最大。

2.predictive entropy

predictive entropy表示预测分布的平均信息量。计算方法如下：

$H[y|x, D_{train}] := -\sum_c p(y=c|x, D_{train}) log\, p(y=c|x, D_{train})$

对于一个测试样本x，predictive entropy在所有类别概率都相等的情况下达到最大值。

对于T次测试实验而言，可以用$\frac{1}{T}\sum_t p(y=c|x, \hat{\omega}_t)$来取代$p(y=c|x, D_{train})$，其中 $p(y=c|x, \hat{\omega}_t)$ 是输入x对应c类别输出的概率，参数$\hat{\omega}_t$是每次实验的模型参数。

3.mutual information

As an alternative to the predictive entropy, the mutual information between the prediction y and the posterior over the model parameters $\boldsymbol{\omega}$ offers a different measure of uncertainty:

$$\begin{aligned}\mathbb{I}[y, \boldsymbol{\omega}|\mathbf{x}, \mathcal{D}_{\text{train}}] &:= \mathbb{H}[y|\mathbf{x}, \mathcal{D}_{\text{train}}] - \mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D}_{\text{train}})}\Big[\mathbb{H}[y|\mathbf{x}, \boldsymbol{\omega}]\Big] \\ &= -\sum_c p(y=c|\mathbf{x}, \mathcal{D}_{\text{train}}) \log p(y=c|\mathbf{x}, \mathcal{D}_{\text{train}}) \\ &\quad + \mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D}_{\text{train}})}\Big[\sum_c p(y=c|\mathbf{x}, \boldsymbol{\omega}) \log p(y=c|\mathbf{x}, \boldsymbol{\omega})\Big]\end{aligned}$$

with c the possible classes y can take. This tractable view of the mutual information was suggested in [Houlsby et al., 2011] in the context of active learning. Test points $\mathbf{x}$ that maximise the mutual information are points on which the model is uncertain on average, yet there exist model parameters that erroneously produce predictions with high confidence.

The mutual information can be approximated in our setting in a similar way to the predictive entropy approximation:

$$\begin{aligned}\tilde{\mathbb{I}}[y, \boldsymbol{\omega}|\mathbf{x}, \mathcal{D}_{\text{train}}] := &-\sum_c \left(\frac{1}{T}\sum_t p(y=c|\mathbf{x}, \widehat{\boldsymbol{\omega}}_t)\right) \log \left(\frac{1}{T}\sum_t p(y=c|\mathbf{x}, \widehat{\boldsymbol{\omega}}_t)\right) \\ &+ \frac{1}{T}\sum_{c,t} p(y=c|\mathbf{x}, \widehat{\boldsymbol{\omega}}_t) \log p(y=c|\mathbf{x}, \widehat{\boldsymbol{\omega}}_t)\end{aligned}$$

关于三种度量的直觉：

对于二分类问题，假设现在有三种类型的实验结果（每种类型的实验结果是多次实验结果的集合）：

1. 预测的概率向量为：$\{(1,0), \ldots, (1,0)\}$
2. 预测的概率向量为：$\{(0.5,0.5), \ldots, (0.5,0.5)\})$
3. 预测的概率向量为：$\{(1,0), (0,1), (0,1), \ldots, (1,0)\}$

对于例子1预测结果具有很高的置信度，而2,3的置信度较低，这种置信度属于**predictive uncertainty**。

从另一方面考虑，例子1,2中模型对预测结果都非常有信心，只有例子3中模型对结果预测很没有信心，这种属于**model uncertainty**（自己的理解：可以理解为参数的后验分布非常宽，比如高斯分布的方差非常大。在这种情况下，每次进行dropout对模型参数采样的结果都非常不一致，这样导致每次预测的记结果都非常不稳定，这种就称为模型的不确定性）。

对于三个例子，三种不确定度值如下表：

	variation ratio	predictive entropy	mutual information
1	0	0	0
2	0.5	0.5	0
3	0.5	0.5	0.5