第五讲
信息论基础
Lecture 5
Elements of Information Theory

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

  The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

引论
Introduction

- 当将神经网络视为一个参数化的概率模型$f(y|X;w)$（依赖于w），即观察到某个图片输入时，输出为特定类别的概率时，图片分类问题可视为一个贝叶斯决策问题：

  When treat a neural network as a probabilistic model $f(y|X;w)$, which is parameterized by w, aka, the probability of a specific class given the image data, the image classification problem can be superficially interpreted as a Bayesian decision problem:

$$D(y|X=x)=\frac{p(X=x|y=0)P(y=0)}{p(X=x|y=1)P(y=1)}$$

- 但是条件概率$p(X|y=0)$和$p(X|y=1)$，并不好估计

  But it is not easy to estimate the conditional probability $p(X|y=0)$ and $p(X|y=1)$

信息
Information

- 从学科的角度给信息一个非常清晰的定义有些困难，一个可能的说法是，信息是衡量不确定性的量。考虑事件A，其发生具有随机性，则其信息量定义为$I(A) = -\log_2 P(A)$，即其发生的概率的对数的取反值。

  The definition of information from the disciplinary perspective is difficult, one possible way is to depict it as the quantity for measuring uncertainty. For example, considering a random event A, the information it contains is defined as $I(A) = -\log_2 P(A)$, aka, the negative of the logarithm of its probability.

- 以掷硬币为例，记A为出现正面的概率。若硬币无暇，则出现正面的概率为0.5，此时$I_{perfect}(A) = \log_2 2 = 1$；若硬币有瑕，导致出现正面概率为0.25，则$I_{defect}(A) = 2\log_2 2 = 2$，即此种情况下出现正面更会让人惊讶。

  Take tossing a coin for example, let A denote the case of head up. For a prefect coin, we know $P(A) = 0.5$, which induces an information quantity $I_{perfect}(A) = \log_2 2 = 1$. A defect coin might reduce the chance of head up from 0.5 to 0.25, now we have $I_{defect}(A) = 2\log_2 2 = 2$, which means when it happens, it surprise us more.

信息
Information

- 假设一个随机试验S有N种结果$\{x_1, x_2, \cdots, x_N\}$，每种结果的信息量的加权平均值定义为$S$的熵，即$H(S) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i)$

  Assume the outcomes of an experiment are enumerated as $\{x_1, x_2, \cdots, x_N\}$, the weighted sum of the information of each outcome is called the entropy of S, defined as $H(S) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i)$.

- 以掷硬币为例，若硬币无暇，出现正面与反面的概率均为0.5，此时

  Take tossing a coin for example, let A denote the case of head up. For a prefect coin, we know $P(A) = 0.5$, which induces an information quantity

$$H_{perfect}(S) = 0.5 * \log_2 2 + 0.5 * \log_2 2 = 1$$

信息
Information

- 若硬币有瑕，导致出现正面概率为0.25，反面概率为0.75，则

  A defect coin might reduce the chance of head up from 0.5 to 0.25, now we have

$$H_{defect}(S) = 0.25 = 0.25 * \log_2 2 + 0.75(2 * \log_2 2 - \log_2 3) = 1 + 0.75(1 - \log_2 3)$$

- 由于$1 - \log_2 3 < 0$，熵减小，即表明此时情况更容易猜结果。

  Since $\log 2 - \log 3 < 0$, decrease in entropy means more easier to guess the outcome.

Kullback–Leibler散度（Divergence）

- 我们可能会用一个简化的模型去估计一个随机系统的输出的概率分布函数。有时候，我们需要知道这个模型的优劣。因此，我们需要知道估计的概率分布函数$q(x)$与真实的概率分布函数$p(x)$差多少，我们定义如下式子来度量这种差异，称为相对熵或KL散度：

  We might use a simplified model to estimate the probability of outputs of a stochastic system. Sometimes, we need to assess the performance of the model. Hence, we need to benchmark the difference between the estimated distribution compared against the real distribution $p(x)$. The following formula is defined to cater to the need, which is called Kullback-Leibler divergence or relative entropy:

$$D_{KL}(P\|Q) = E_{x\sim P}\left[\log\frac{p(x)}{q(x)}\right]$$

Kullback–Leibler散度（Divergence）

- 我们有以下关于KL散度的一些性质：

  We have the following properties about KL divergence:

  - KL散度是非对称的，即对于P与Q而言，$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$

    KL divergence is non-symmetric, namely, for P and Q, usually $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$

  - KL散度是非负的，即$D_{KL}(P\|Q) \geq 0$，等于0仅当$P = Q$时成立

    LK divergence is non-negative, namely, $D_{KL}(P\|Q) \geq 0$, equality holds if and only if $P = Q$

  - 若存在x使得$p(x) > 0$，$q(x) = 0$，则$D_{KL}(P\|Q) = \infty$，即KL散度可能计算不稳定

    If there exists x satisfying $p(x) > 0$ and $q(x) = 0$, then $D_{KL}(P\|Q) = \infty$, which means KL divergence might be unstable in computation

Kullback–Leibler散度（Divergence）

- 我们来分析一下KL散度可能计算不稳定的原因。若X在$X = x_0$点为小概率事件（假设考虑离散变量的情况），则在实际中x_0点可能采样不到，则$Q(x_0) = 0$但$P(x_0) > 0$。因此

  We now analyze the reason for computational instability of KL divergence. If $X = x_0$ is of small probability (suppose discrete variable is considered), of great chance that x_0 will be missed during sampling, leading to $Q(x_0) = 0$ whilst $P(x_0) > 0$. Therefore

  $$P(x_0) \cdot \log \frac{P(x_0)}{Q(x_0)} = \infty$$

  实际上，由于X在$X = x_0$点为小概率事件，则在x_0的估算，不应对计算估计偏差时，贡献太多。则考虑形式：$P(x_0) \cdot \log Q(x_0)$，此时$P(x_0) \cdot \log Q(x_0) = P(x_0) \to 0$，正符合我们的需求。

  Actually, since $X = x_0$ is a small probability event, so bias for $Q(x_0)$ shall contribute minute impact on the probability estimation. Consider the form $P(x_0) \cdot \log Q(x_0)$, now $P(x_0) \cdot \log Q(x_0) = P(x_0) \to 0$, and this satisfying what we desired.

交叉熵
Cross-Entropy

- 定义：给定随机变量X，其真实分布P与估计分布Q之间的差异由如下形式的交叉熵度量：

  Definition: given random variable X, the divergence between the real distribution P and estimated distribution Q is measured by the cross entropy defined as below:

$$\mathrm{CE}(P,Q) = -E_{x\sim P}\left[\log\left(q(x)\right)\right]$$

- 可以看出，结合上面定义的自信息与KL散度，容易得知其三者的关系如下：

  It is easy to verify by definition of self-information and KL divergence, that cross entropy is linked with the previous two terms via equation below:

$$\mathrm{CE}(P,Q) = H(P) + D_{KL}(P\|Q)$$

交叉熵
Cross-Entropy

- 对于交叉熵，我们有如下重要性质：

  We have the following important property for cross entropy:

$$H(X) = -E_{x\sim P}\left[\log\left(p(x)\right)\right] \leq -E_{x\sim P}\left[\log\left(q(x)\right)\right] = \mathrm{CE}(P, Q)$$

- 这个式子通俗地说，如果随机事件的结果和我们预想的不一样，我们都会惊讶，越不一样，越惊讶。

  The inequality or formula above intuitively states that if the outcomes of an event are out of our expectation, we will feel surprised. The more diverse from our expectation, the more astonishment we expose.

交叉熵
Cross-Entropy

- 上述性质是如此之重要，我们对其证明给一个概略地描述：

  We have the following important property for cross entropy:

- 琴生不等式

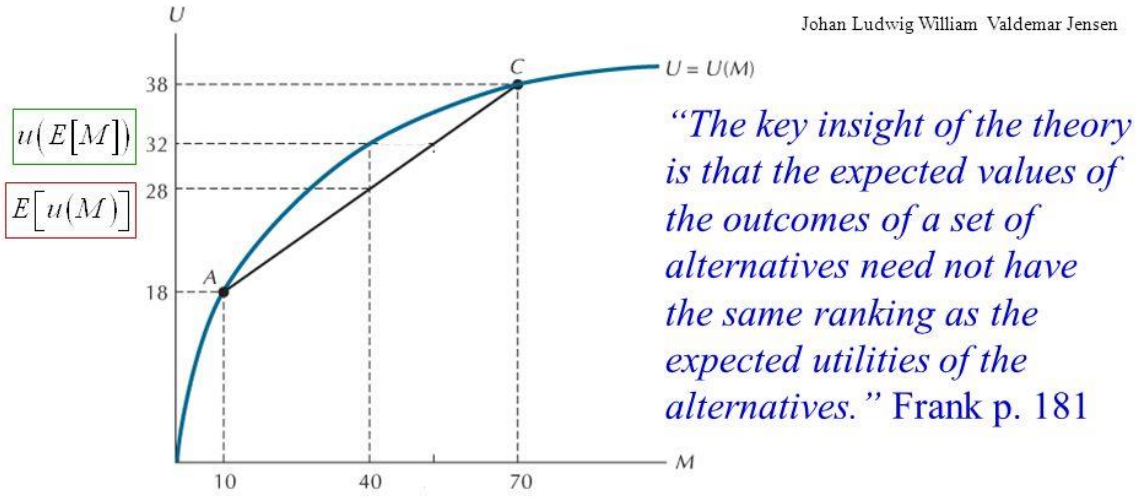
  Jensen’s Inequality

Jensen’s Inequality

$$E\left[u(M)\right] < u\left(E\left[M\right]\right)$$

For any concave function, the expected value of the function is less than the function of the expected value.

Johan Ludwig William Valdemar Jensen



“The key insight of the theory is that the expected values of the outcomes of a set of alternatives need not have the same ranking as the expected utilities of the alternatives.” Frank p. 181

交叉熵
Cross-Entropy

- 对于交叉熵，我们有如下重要性质：

  We have the following important property for cross entropy:

$$
\begin{aligned}
H(X) - \mathrm{CE}(P,Q) &= -E_{x\sim P}\left[\log\left(p(x)\right)\right] - \left(-E_{x\sim P}\left[\log\left(q(x)\right)\right]\right) \\
&= \sum_{i=1}^{n} P(x_i)\log\left(\frac{1}{P(x_i)}\right) - \sum_{i=1}^{n} P(x_i)\log\left(\frac{1}{Q(x_i)}\right) \\
&= \sum_{i=1}^{n} P(x_i)\log\left(\frac{Q(x_i)}{P(x_i)}\right) \leq \log\left(\sum_{i=1}^{n} P(x_i)\frac{Q(x_i)}{P(x_i)}\right) \\
&= \log\left(\sum_{i=1}^{n} Q(x_i)\right) = \log(1) = 0
\end{aligned}
$$

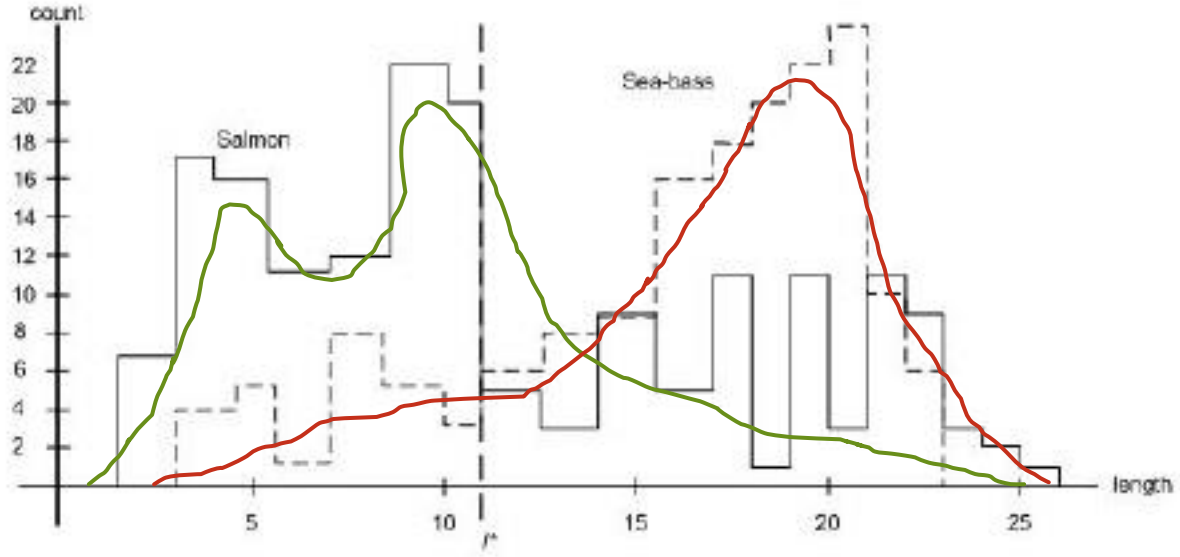
- 即$H(X) - \mathrm{CE}(P,Q) \leq 0$，所以$H(X) \leq \mathrm{CE}(P,Q)$

  That's $H(X) - \mathrm{CE}(P,Q) \leq 0$, therefore, $H(X) \leq \mathrm{CE}(P,Q)$

回忆：贝叶斯决策理论
Recall: Bayesian Decision Theory

- 假设在长期的实践中，观察到鱼体长度与种类的关系。经平滑，得到以下概率分布函数。则又该如何判断呢？

  Suppose during the practice, the relation between the class of fish and its corresponding length has been drawn. The probability distributions of the length of fishes are shown below after smoothing. Then how to decide?



回忆：贝叶斯决策理论
Recall: Bayesian Decision Theory

- 若$D(X|L=l)>1$，则判断为鲈鱼，否则判断为三文鱼。

  If $D(X|L=l)>1$, Then the output is sea bass, otherwise it is salmon.

- 注意到我们对$P(L|X=0)$或$P(L|X=1)$，均是得到平滑之后的概率分布函数，即$p(L|X=0)$或$p(L|X=1)$，但对$D(X|L=l)$的定义，在有的项是概率分布，有的项是概率分布函数的时候，依然有效，即依然可以依据以下定义的$D(X|L=l)$进行判断：

  It is mentioned here that for $P(L|X=0)$ or $P(L|X=1)$, what we obtained are the distribution functions after smoothing, aka, $p(L|X=0)$ and $p(L|X=1)$. However, in practice, the definition of $D(X|L=l)$ still holds by mixing probability and probability distribution together, just as above. We can still rely on the value of $D(X|L=l)$ to make the decision:

$$D(X|L=l)=\frac{p(L=l|X=0)P(X=0)}{p(L=l|X=1)P(X=1)}$$

回顾
Retrospection

- 可将神经网络视为一个参数化的概率模型$f(y|X;w)$（依赖于w），即观察到某个图片输入时，输出为特定类别的概率时。若视为一个贝叶斯决策问题时，条件概率$p(X|y=0)$和$p(X|y=1)$，并不好估计，那算了，干脆直接比较似然。对于逻辑回归，即二分类问题，即比较$f(y=0|X=x;w)$与$f(y=1|X=x;w)$。对于没有见过的x怎么办？就由见过的x内插或外插，即光滑过去就好。若光滑的效果不好怎么办？那能怎么办？就说模型的泛化性能不好。

  To treat a neural network as a probabilistic model $f(y|X;w)$, which is parameterized by w, it can be interpreted as the probability of a specific class given the image data. However, the superficial interpretation as a Bayesian decision problem leads to the intractability to estimate $p(X|y=0)$ and $p(X|y=1)$. So let it be and a direct thought is to just compare the likelihood. For logistic regression or binary classification problem, it is to compare $f(y=0|X=x;w)$ with $f(y=1|X=x;w)$. Then how to tackle the unseen x, that's easy, by interpolation or extrapolation (smoothing) via the already seen x. Then how we deal with the situation of poor outcomes by smoothing? There is nothing we can do about it, just say the model generalizes bad.

交叉熵
Cross-Entropy

- 既然要比较似然，例如$f(y=0|X=x;w)$与$f(y=1|X=x;w)$，那么至少对已知的x，估计要准确。怎么估计准确呢？我们已经对神经网络进行改造，其输出在观察到x时，特定类别的概率。那么，估计准确意味着这个概率要趋近真实概率。那么真实概率是什么呢？不知道。

  Since it is to compare the likelihood, for example, $f(y=0|X=x;w)$ and $f(y=1|X=x;w)$, so at least for already-seen x, the estimate should be accurate as possible. Then how? We already modify the network to output probability of specific class upon observation of x. Acuity means the estimation is convergent to the real likelihood at maximum. The question is what's the real distribution? Sorry, it is unknown.

交叉熵
Cross-Entropy

- 特征脸：一张是脸的图片之所以会被认成脸，可以认为特定位置的像素的值导致的结果，这些位置像素值可假设是满足一定概率分布的。设$x_{i,j}$表示位置为(i,j)的随机变量，则人脸的这些位置的像素值的联合概率分布可记为$P(\{x_{i,j}\})$。在简化情况下，我们可以认为对这个分布采样，就会得到一张人脸。

  Eigenface: a face of a human being is just being identified as it is, is due to the pixels in specific positions. And these pixel values are hypothesized to satisfy a certain probability distribution. Let $x_{i,j}$ denote the rando of pixel at (i,j), then we have a joint distribution of the pixel values $P(\{x_{i,j}\})$. And by sampling such a distribution, we can get a face.

交叉熵
Cross-Entropy

- 我们从自然界随便拍到一张猫的照片或狗的照片供神经网络分析，我们称为一次采样（假设像上面讲述那样，有一个狗的或猫的distribution供我们采）。第i次采样T_i的数据与标签，记为$T_i = (X_i, \bar{y}_i)$。 我们可以假设T_i为随机变量。因为做之前并不清楚对何种物体进行采样及采样得到的数据。同时我们可以称$\{T_i\}$为独立同分布随机变量，因为第i次采样与第j次采样没有任何关系。

  We call the image, of which such as dogs or cats, to be fed into neural network for analysis, is a sampling instance (suppose there exist distributions of dogs and cats that we can sample from). The data and label generated from the i^{th} sampling are denoted by $T_i = (X_i, \bar{y}_i)$. T_i is supposedly to be random variable, since before sampling, we have no information of it. And $\{T_i\}$ are independent identical variables since there is no relation or entanglement between the i^{th} and j^{th} sampling.

- 假设我们只采样一次，去估计似然，怎么办？

  We can we do for estimating the likelihood by just one sampling?

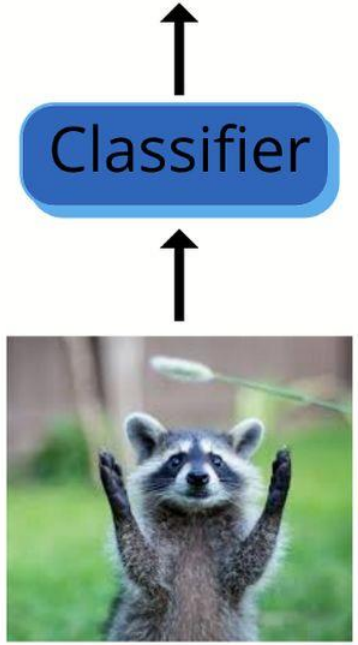
交叉熵
Cross-Entropy

- 假设我们只有一个样本，$T_1 = (X_1, \bar{y}_1)$，处理的是K分类问题。设$\bar{y}_1 = k \in \{1, 2, \cdots, K\}$，$\vec{y}_1 = f(y|X; w)$。注意$\bar{y}_1$是一个标量，表示属于特定类的真实标签，$\vec{y}_1$是一个向量，表示属于各个类的概率。由于只有一个样本，但似然该估还得估。那什么是真实分布呢？对其而言，那只能属于第k类的概率是100%，属于其它类的概率为0。

  Suppose we just have one example $T_1 = (X_1, \bar{y}_1)$ to process a K classification problem. Denote $\bar{y}_1 = k \in \{1, 2, \cdots, K\}$ and $\vec{y}_1 = f(y|X; w)$. Note $\bar{y}_1$ is a scalar, indicating the ground-truth label, while $\vec{y}_1$ is a vector, representing the probabilities of belong to each classes. Although we just have one sample, it is still obligatory to estimate the likelihood. Now what's the real distribution? Without other choices, the chance of belonging to class k is 100%, and zero probabilities for all other classes.

交叉熵
Cross-Entropy

	CAT	DOG	FOX	RED PANDA	RACCOON	COW	BEAR
True Distribution:	0%	0%	0%	0%	100%	0%	0%
Predicted Distribution:	5%	2%	45%	10%	30%	1%	2%



交叉熵
Cross-Entropy

- 若对$\bar{y}_1$进行独热编码，即

  If $\bar{y}_1$ is encoded in one-hot format, namely,

$$\bar{y}_1 = \begin{matrix} 1 & 2 & \cdots & k & \cdots & K \\ 0 & 0 & \cdots & 1 & \cdots & 0 \end{matrix}$$

- 可以看出，其与真实概率分布一致，所以，我们有

  It is manifest that it is coincident with the real distribution, so we have

$$L = \mathrm{CE}(\bar{y}_1, \vec{y}_1) = -\bar{y}_{1,1} \log \vec{y}_{1,1} - \bar{y}_{1,2} \log \vec{y}_{1,2} - \cdots - \bar{y}_{1,K} \log \vec{y}_{1,K} = -\sum_{k=1}^{K} \bar{y}_{1,k} \log \vec{y}_{1,k}$$

- 若有N个样本，则损失函数变为

  If there are N samples, then the loss function is updated to the formular below

$$L = \sum_{i=1}^{N} \mathrm{CE}(\bar{y}_i, \vec{y}_i) = \sum_{i=1}^{N} \sum_{k=1}^{K} \bar{y}_{i,k} \log \vec{y}_{i,k}$$

习题
Problems

1. 解释为什么eigenface可以不精确地用于表示组成脸的像素的概率分布？

   Explain why eigenface can grossly represents the pixel value distributions of people's face?

2. 交叉熵有何性质？

   Enumerate some properties of cross-entropy.

3. 写出二分类问题的损失函数.

   Write down the loss function for binary classification.

习题
Problems

4. 证明对于分类问题，如下图所示的交叉熵损失函数L，其对于输入x_i的导数$\frac{\partial L}{\partial x_i}$为预测值$y_i$与真实值$\tilde{y}_i$之差，即$\frac{\partial L}{\partial x_i} = y_i - \tilde{y}_i$。

   Prove for the classification problem, the derivative of input x_i with regard to the cross-entropy loss function L, where y_i indicates the predicted value and $\tilde{y}_i$ the ground-truth label, can be as simplified as $\frac{\partial L}{\partial x_i} = y_i - \tilde{y}_i$.

$$L = -\sum_{k=1}^{n} \tilde{y}_k \ln y_k$$