1、What are representation and representation learning meaning?

在深度学习领域内，**表示**是指通过模型的参数，采用何种形式、何种方式来表示模型的输入观测样本X。

**表示学习**指学习对观测样本X有效的表示。表示学习是学习一个特征的技术的集合：将原始数据转换成为能够被机器学习来有效开发的一种形式。

2、Why do we need to perform representation learning in many machine learning applications?

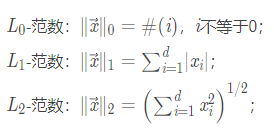
因为学习到的表示往往比手动设计的表示表现得更好，并且它们只需最少的人工干预，就能让AI系统迅速适应新的任务。表示学习算法只需几分钟就可以为简单的任务发现一个很好的特征集，对于复杂任务也只需要几小时到几个月。而手动为一个复杂的任务设计特征则需要耗费大量的人工时间和精力；甚至需要花费整个社群研究人员几十年的时间。

3、What’s the relationship between AI, machine learning and deep learning?

**机器学习**是实现人工智能的一种方法。机器学习表明，人工智能系统需要有获取自身知识的能力。机器学习使用概率理论从原始数据中学习。

**深度学习**是一种机器学习，通过建立更深的网络获得更好的结果。

4、What are L0 norm, L1 norm, and L2 norm?



5、What’s the definition of the Frobenius norm of matrices?

Frobenius 范数，简称F-范数，是一种矩阵范数，记为||·||F。矩阵A的Frobenius范数定义为矩阵A各项元素的绝对值平方的总和开根，即

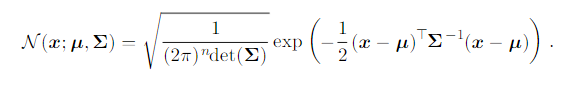
6、What’s the chain rule of conditional probabilities?

任何多维随机变量的联合概率分布，都可以分解成只有一个变量的条件概率相乘的形式：

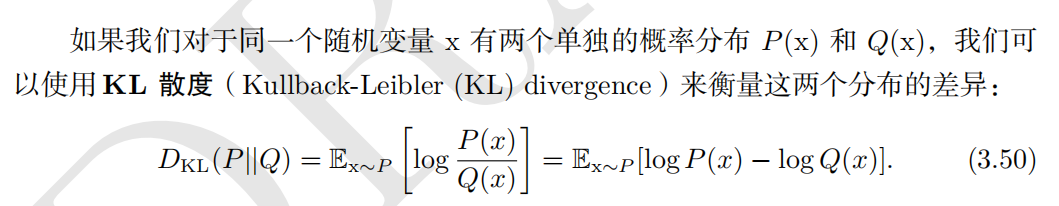


这个规则被称为概率的链式法则或者乘法法则。

7、Please write out the probabilistic density function of multivariate Gaussian.

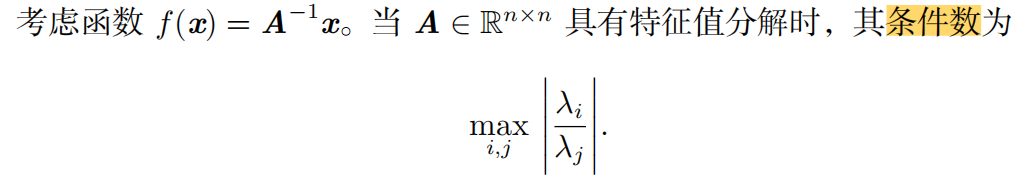


8、How to compute the Kullback-Leibler (KL) divergence of two distributions?



9、What is the condition number?

**条件数**是指函数相对于输入的微小变化而变化的快慢程度。



10、What are the local minimum and saddle point?

一个局部极小点意味着这个点的 f(x) 小于所有邻近点，因此不可能通过移动无穷小的步长来减小f(x)。有些临界点既不是最小点也不是最大点。这些点被称为**鞍点**。

11、How to evaluate the abilities of a machine learning algorithms?

通常性能度量 P 是特定于系统执行的任务 T 而言的。对于诸如分类、缺失输入分类和转录任务，我们通常度量模型的准确率，是指该模型输出正确结果的样本比率。也可以通过 错误率得到相同的信息，错误率是指该模型输出错误结果的样本比率。对于密度估计这类任务而言，最常用的是输出模型在一些样本上概率对数的平均值。也可以使用测试集数据来评估系统性能。

12、What are overfitting and underfitting?

**过拟合**是指训练误差和和测试误差之间的差距太大；

**欠拟合**是指模型不能在训练集上获得足够低的误差。

13、What is the curse of dimensionality?

当数据的维数很高时，很多机器学习问题变得相当困难，这种现象被称为维数灾难。

低维（特征少）转向高维的过程中，样本会变的稀疏（可以有两种理解方式：1.样本数目不变，样本彼此之间距离增大。2.样本密度不变，所需的样本数目指数倍增长）。

高维度带来的影响：

**1.变得可分。**由于变得稀疏，之前低维不可分的，在合适的高维度下可以找到一个可分的超平面。

**2.过拟合风险。**过高维度会带来过拟合的风险（会学习到数据集中的特例或异常，对现实测试数据效果较差）。增加维度的线性模型等效于低维空间里较复杂的非线性分类器。

**3.需要更多训练数据。**我们需要更多的训练数据进行参数估计。

**4.过高维度会让分类变难。**

**5.高维度中用距离来衡量样本相似性的方法已经渐渐失效。**

14、For a regression problem with real outputs, what kind of output units should we choose? Why？

线性单元常用于回归问题。给定特征 h，线性输出单元层产生一个向量因为线性模型不会饱和，所以它们易于采用基于梯度的优化算法，甚至可以使用其他多种优化算法。

15、Why do we generally use ReLU as the activation function other than sigmoid/tanh in CNNs?

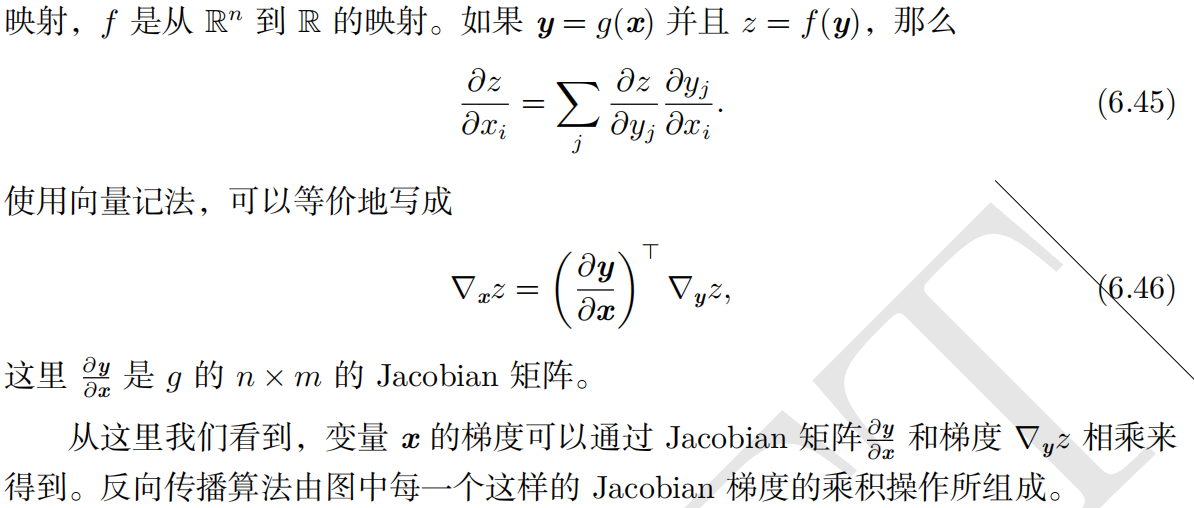
**第一**，采用sigmoid等函数，算激活函数时（指数运算），计算量大，反向传播求误差梯度时，求导涉及除法和指数运算，计算量相对大，而采用Relu激活函数，整个过程的计算量节省很多。

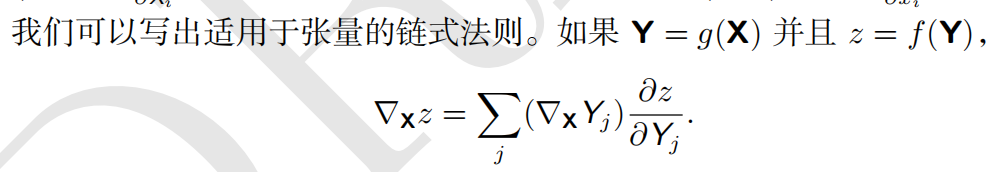
**第二**，对于深层网络，sigmoid函数反向传播时，很容易就会出现梯度消失的情况（在sigmoid接近饱和区时，变换太缓慢，导数趋于0，这种情况会造成信息丢失），这种现象称为饱和，从而无法完成深层网络的训练。而ReLU就不会有饱和倾向，不会有特别小的梯度出现。

**第三**，Relu会使一部分神经元的输出为0，这样就造成了网络的稀疏性，并且减少了参数的相互依存关系，缓解了过拟合问题的发生。

16、How to use the back propagation algorithm to optimize deep neural networks?

反向传播允许来自代价函数的信息通过网络向后流动，以便计算梯度。反向传播算法大体上来说是链式求导法则在神经网络训练中的应用。





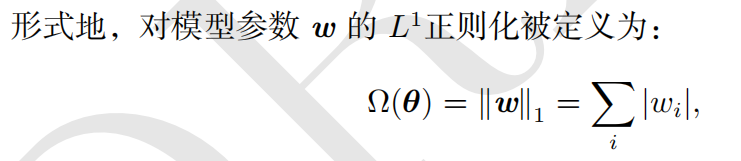
17、Why do we apply regularization on deep learning models?

深层网络架构的学习要求大量数据，对计算能力的要求很高。神经元和参数之间的大量连接需要通过梯度下降及其变体以迭代的方式不断调整。此外，有些架构可能因为强大的表征力而产生测试数据过拟合等现象。正则化通过避免训练完美拟合数据样本的系数而有助于算法的泛化。

1. Please introduce at least 3 norm based regularization methods.

**（1）参数范数惩罚**：基本公式

L2参数正则化：通过向目标函数添加一个正则项

L1参数正则化：

**（2）数据集增强**

**（3）注入噪声**

1. Why do we consider multi-task learning as a regularization method?

多任务学习是通过合并几个任务中的样例（可以视为对参数施加的软约束）来提高泛化的一种方式。正如额外的训练样本能够将模型参数推向具有更好泛化能力的值一样，当模型的一部分被多个额外的任务共享时，这部分将被约束为良好的值（如果共享合理），通常会带来更好的泛化能力。因此可作为一种正则化的手段。

1. How to perform dropout regularization?

Dropout训练的集成包括所有从基础网络除去非输出单元后形成的子网络。具体来说，在训练中使用Dropout时，我们会使用基于小批量的学习算法和较小的步长，如梯度下降等。我们每次在小批量中加载一个样本，然后随机抽样应用于网络中所有输入和隐藏单元的不同二值掩码。对于每个单元，掩码是独立采样的。掩码值为 1 的采样概率（导致包含一个单元）是训练开始前一个固定的超参数。它不是模型当前参数值或输入样本的函数。通常在每一个小批量训练的神经网络中，一个输入单元被包括的概率为 0*.*8，一个隐藏单元被包括的概率为 0*.*5。然后，我们运行和之前一样的前向传播、反向传播以及学习更新。

更正式地说，假设一个掩码向量 ***µ*** 指定被包括的单元，*J*(***θ****,* ***µ***) 是由参数 ***θ*** 和掩码 ***µ*** 定义的模型代价。那么Dropout训练的目标是最小化 E***µ****J*(***θ****,* ***µ***)。期望包含多达指数级的项，但我们可以通过抽样 ***µ*** 获得梯度的无偏估计。