

谷歌开发者机器学习词汇表：纵览机器学习基本词汇与概念

2017-10-05 机器之心

选自Google Developers

机器之心编译

机器之心曾开放过人工智能术语集，该术语库项目目前收集了人工智能领域 700 多个专业术语，但仍需要与各位读者共同完善与修正。本文编译自谷歌开发者机器学习术语表项目，介绍了该项目所有的术语与基本解释。之后，我们也将表内术语更新到了机器之心 GitHub 项目中。

机器之心人工智能术语项目：<https://github.com/jiqizhixin/Artificial-Intelligence-Terminology>

A

准确率 (accuracy)

分类模型预测准确的比例。在多类别分类中，准确率定义如下：

$$Accuracy = \frac{Correct\ Predictions}{Total\ Number\ Of\ Examples}$$

在二分类中，准确率定义为：

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ Of\ Examples}$$

激活函数 (Activation function)

一种函数（例如 ReLU 或 Sigmoid），将前一层所有神经元激活值的加权和输入到一个非线性函数中，然后向下一层传递该函数的输出值（典型的非线性）。

AdaGrad

一种复杂的梯度下降算法，重新调节每个参数的梯度，高效地给每个参数一个单独的学习率。详见论文：<http://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf>。

AUC (曲线下面积)

一种考虑到所有可能的分类阈值的评估标准。ROC 曲线下面积代表分类器随机预测真正类 (True Positives) 要比假正类 (False Positives) 概率大的确信度。

B

反向传播 (Backpropagation)

神经网络中完成梯度下降的重要算法。首先，在前向传播的过程中计算每个节点的输出值。然后，在反向传播的过程中计算与每个参数对应的误差的偏导数。

基线 (Baseline)

被用为对比模型表现参考点的简单模型。基线帮助模型开发者量化模型在特定问题上的预期表现。

批量

模型训练中一个迭代 (指一次梯度更新) 使用的样本集。

批量大小 (batch size)

一个批量中样本的数量。例如，SGD 的批量大小为 1，而 mini-batch 的批量大小通常在 10-1000 之间。批量大小通常在训练与推理的过程中确定，然而 TensorFlow 不允许动态批量大小。

偏置 (bias)

与原点的截距或偏移量。偏置 (也称偏置项) 被称为机器学习模型中的 b 或者 w_0 。例如，偏置项是以下公式中的 b ： $y' = b + w_1x_1 + w_2x_2 + \dots + w_nx_n$ 。

注意不要和预测偏差混淆。

二元分类器 (binary classification)

一类分类任务，输出两个互斥（不相交）类别中的一个。例如，一个评估邮件信息并输出「垃圾邮件」或「非垃圾邮件」的机器学习模型就是一个二元分类器。

binning/bucketing

根据值的范围将一个连续特征转换成多个称为 buckets 或者 bins 二元特征，称为 buckets 或者 bins。例如，将温度表示为单一的浮点特征，可以将温度范围切割为几个离散的 bins。假如给定的温度的敏感度为十分之一度，那么分布在 0.0 度和 15.0 度之间的温度可以放入一个 bin 中，15.1 度到 30.0 度放入第二个 bin，30.1 度到 45.0 度放入第三个 bin。

C

标定层 (calibration layer)

一种调整后期预测的结构，通常用于解释预测偏差。调整后的预期和概率必须匹配一个观察标签集的分布。

候选采样 (candidate sampling)

一种优化训练时间的，使用 Softmax 等算法计算所有正标签的概率，同时只计算一些随机取样的负标签的概率。例如，有一个样本标记为「小猎兔狗」和「狗」，候选取样将计算预测概率，和与「小猎兔狗」和「狗」类别输出（以及剩余的类别的随机子集，比如「猫」、「棒棒糖」、「栅栏」）相关的损失项。这个想法的思路是，负类别可以通过频率更低的负强化（negative reinforcement）进行学习，而正类别经常能得到适当的正强化，实际观察确实如此。候选取样的动力是计算有效性从所有负类别的非计算预测的得益。

检查点 (checkpoint)

在特定的时刻标记模型的变量的状态的数据。检查点允许输出模型的权重，也允许通过多个阶段训练模型。检查点还允许跳过错误继续进行（例如，抢占作业）。注意其自身的图式并不包含于检查点内。

类别 (class)

所有同类属性的目标值作为一个标签。例如，在一个检测垃圾邮件的二元分类模型中，这两个类别分别是垃圾邮件和非垃圾邮件。而一个多类别分类模型将区分狗的种类，其中的类别可以是贵宾狗、小猎兔狗、哈巴狗等等。

类别不平衡数据集 (class-imbalanced data set)

这是一个二元分类问题，其中两个类别的标签的分布频率有很大的差异。比如，一个疾病数据集中若 0.01% 的样本有正标签，而 99.99% 的样本有负标签，那么这就是一个类别不平衡数据集。但对于一个足球比赛预测器数据集，若其中 51% 的样本标记一队胜利，而 49% 的样本标记其它队伍胜利，那么这就不是一个类别不平衡数据集。

分类模型 (classification)

机器学习模型的一种，将数据分离为两个或多个离散类别。例如，一个自然语言处理分类模型可以将一句话归类为法语、西班牙语或意大利语。分类模型与回归模型 (regression model) 成对比。

分类阈值 (classification threshold)

应用于模型的预测分数以分离正类别和负类别的一种标量值标准。当需要将 logistic 回归的结果映射到二元分类模型中时就需要使用分类阈值。例如，考虑一个确定给定邮件为垃圾邮件的概率的 logistic 回归模型，如果分类阈值是 0.9，那么 logistic 回归值在 0.9 以上的被归为垃圾邮件，而在 0.9 以下的被归为非垃圾邮件。

混淆矩阵 (confusion matrix)

总结分类模型的预测结果的表现水平 (即，标签和模型分类的匹配程度) 的 NxN 表格。混淆矩阵的一个轴列出模型预测的标签，另一个轴列出实际的标签。N 表示类别的数量。在一个二元分类模型中，N=2。例如，以下为一个二元分类问题的简单的混淆矩阵：

	Tumor (predicted)	Non-Tumor (predicted)
Tumor (actual)	18	1
Non-Tumor (actual)	6	452

上述混淆矩阵展示了在 19 个确实为肿瘤的样本中，有 18 个被模型正确的归类 (18 个真正)，有 1 个被错误的归类为非肿瘤 (1 个假负类)。类似的，在 458 个确实为非肿瘤的样本中，有 452 个被模型正确的归类 (452 个真负类)，有 6 个被错误的归类 (6 个假正类)。

多类别分类的混淆矩阵可以帮助发现错误出现的模式。例如，一个混淆矩阵揭示了一个识别手写数字体的模型倾向于将 4 识别为 9，或者将 7 识别为 1。混淆矩阵包含了足够多的信息可以计算很多的模型表现度量，比如精度 (precision) 和召回 (recall) 率。

连续特征 (continuous feature)

拥有无限个取值点的浮点特征。和离散特征 (discrete feature) 相反。

收敛 (convergence)

训练过程达到的某种状态，其中训练损失和验证损失在经过了确定的迭代次数后，在每一次迭代中，改变很小或完全不变。换句话说就是，当对当前数据继续训练而无法再提升模型的表现水平的时候，就称模型已经收敛。在深度学习中，损失值下降之前，有时候经过多次迭代仍保持常量或者接近常量，会造成模型已经收敛的错觉。

凸函数 (concex function)

一种形状大致呈字母 U 形或碗形的函数。然而，在退化情形中，凸函数的形状就像一条线。例如，以下几个函数都是凸函数：

- L2 损失函数
- Log 损失函数
- L1 正则化函数
- L2 正则化函数

凸函数是很常用的损失函数。因为当一个函数有最小值的时候（通常就是这样），梯度下降的各种变化都能保证找到接近函数最小值的点。类似的，随机梯度下降的各种变化有很大的概率（虽然无法保证）找到接近函数最小值的点。

两个凸函数相加（比如，L2 损失函数+L1 正则化函数）后仍然是凸函数。

深度模型通常是非凸的。出乎意料的是，以凸优化的形式设计的算法通常都能在深度网络上工作的很好，虽然很少能找到最小值。

成本 (cost)

loss 的同义词。

交叉熵 (cross-entropy)

多类别分类问题中对 Log 损失函数的推广。交叉熵量化两个概率分布之间的区别。参见困惑度 (perplexity) 。

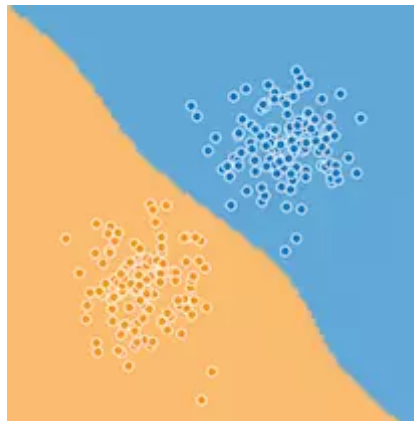
D

数据集 (data set)

样本的集合。

决策边界 (decision boundary)

在一个二元分类或多类别分类问题中模型学习的类别之间的分离器。例如，下图就展示了一个二元分类问题，决策边界即橙点类和蓝点类的边界。



深度模型 (deep model)

一种包含多个隐藏层的神经网络。深度模型依赖于其可训练的非线性性质。和宽度模型对照 (wide model) 。

密集特征 (dense feature)

大多数取值为非零的一种特征，通常用取浮点值的张量 (tensor) 表示。和稀疏特征 (sparse feature) 相反。

派生特征 (derived feature)

合成特征 (synthetic feature) 的同义词。

离散特征 (discrete feature)

只有有限个可能取值的一种特征。例如，一个取值只包括动物、蔬菜或矿物的特征就是离散 (或类别) 特征。和连续特征 (continuous feature) 对照。

dropout 正则化 (dropout regularization)

训练神经网络时一种有用的正则化方法。dropout 正则化的过程是在单次梯度计算中删去一层网络中随机选取的固定数量的单元。删去的单元越多，正则化越强。

动态模型 (dynamic model)

以连续更新的方式在线训练的模型。即数据连续不断的输入模型。

E

早期停止法 (early stopping)

一种正则化方法，在训练损失完成下降之前停止模型训练过程。当验证数据集 (validation data set) 的损失开始上升的时候，即泛化表现变差的时候，就该使用早期停止法了。

嵌入 (embeddings)

一类表示为连续值特征的明确的特征。嵌入通常指将高维向量转换到低维空间中。例如，将一个英语句子中的单词以以下任何一种方式表示：

- 拥有百万数量级（高维）的元素的稀疏向量，其中所有的元素都是整数。向量的每一个单元表示一个单独的英语单词，单元中的数字表示该单词在一个句子中出现的次数。由于一个句子中的单词通常不会超过 50 个，向量中几乎所有的单元都是 0。少量的非零的单元将取一个小的整数值（通常为 1）表示句子中一个单词的出现次数。
- 拥有数百个（低维）元素的密集向量，其中每一个元素取 0 到 1 之间的浮点数。

在 TensorFlow 中，嵌入是通过反向传播损失训练的，正如神经网络的其它参量一样。

经验风险最小化 (empirical risk minimization , ERM)

选择能最小化训练数据的损失的模型函数的过程。和结构风险最小化 (structural risk minimization) 对照。

集成 (ensemble)

多个模型预测的综合考虑。可以通过以下一种或几种方法创建一个集成方法：

- 设置不同的初始化；
- 设置不同的超参量；
- 设置不同的总体结构。

深度和广度模型是一种集成。

评估器 (Estimator)

tf.Estimator 类的一个例子，封装 logic 以建立一个 TensorFlow 图并运行一个 TensorFlow session。你可以通过以下方式创建自己的评估器：<https://www.tensorflow.org/extend/estimators>

样本 (example)

一个数据集的一行内容。一个样本包含了一个或多个特征，也可能是一个标签。参见标注样本 (labeled example) 和无标注样本 (unlabeled example)。

F

假负类 (false negative , FN)

被模型错误的预测为负类的样本。例如，模型推断一封邮件为非垃圾邮件（负类），但实际上这封邮件是垃圾邮件。

假正类 (false positive , FP)

被模型错误的预测为正类的样本。例如，模型推断一封邮件为垃圾邮件（正类），但实际上这封邮件是非垃圾邮件。

假正类率 (false positive rate , FP rate)

ROC 曲线 (ROC curve) 中的 x 轴。FP 率的定义是：假正率=假正类数/(假正类数+真负类数)

特征 (feature)

输入变量，用于做出预测。

特征列 (feature columns/FeatureColumn)

具有相关性的特征的集合，比如用户可能居住的所有可能的国家的集合。一个样本的一个特征列中可能会有一个或者多个特征。

TensorFlow 中的特征列还可以压缩元数据比如下列情况：

- 特征的数据类型；
- 一个特征是固定长度的或应该转换为嵌入。
- 一个特征列可以仅包含一个特征。「特征列」是谷歌专用的术语。在 VW 系统 (Yahoo/Microsoft) 中特征列的意义是「命名空间」 (namespace)，或者场 (field)。

特征交叉 (feature cross)

将特征进行交叉 (乘积或者笛卡尔乘积) 运算后得到的合成特征。特征交叉有助于表示非线性关系。

特征工程 (feature engineering)

在训练模型的时候，决定哪些特征是有用的，然后将记录文件和其它来源的原始数据转换成上述特征的过程。在 TensorFlow 中特征工程通常意味着将原始记录文件输入 tf.Example 协议缓存中。参见 tf.Transform。特征工程有时候也称为特征提取。

特征集 (feature set)

机器学习模型训练的时候使用的特征群。比如，邮政编码，面积要求和物业状况可以组成一个简单的特征集，使模型能预测房价。

特征定义 (feature spec)

描述所需的信息从 tf.Example 协议缓存中提取特征数据。因为 tf.Example 协议缓存只是数据的容器，必须明确以下信息：

- 需要提取的数据 (即特征的关键信息)
- 数据类型 (比如，浮点数还是整数)

- 数据长度（固定的或者变化的）

Estimator API 提供了从一群特征列中生成一个特征定义的工具。

完全 softmax (full softmax)

参见 softmax。和候选采样对照。

G

泛化 (generalization)

指模型利用新的没见过数据而不是用于训练的数据作出正确的预测的能力。

广义线性模型 (generalized linear model)

最小二乘回归模型的推广/泛化，基于高斯噪声，相对于其它类型的模型（基于其它类型的噪声，比如泊松噪声，或类别噪声）。广义线性模型的例子包括：

- logistic 回归
- 多分类回归
- 最小二乘回归

广义线性模型的参数可以通过凸优化得到，它具有以下性质：

- 最理想的最小二乘回归模型的平均预测结果等于训练数据的平均标签。
- 最理想的 logistic 回归模型的平均概率的预测结果等于训练数据的平均标签。

广义线性模型的能力局限于其特征的性质。和深度模型不同，一个广义线性模型无法「学习新的特征」。

梯度 (gradient)

所有变量的偏导数的向量。在机器学习中，梯度是模型函数的偏导数向量。梯度指向最陡峭的上升路线。

梯度截断 (gradient clipping)

在应用梯度之前先修饰数值，梯度截断有助于确保数值稳定性，防止梯度爆炸出现。

梯度下降 (gradient descent)

通过计算模型的相关参量和损失函数的梯度最小化损失函数，值取决于训练数据。梯度下降迭代地调整参量，逐渐靠近权重和偏置的最佳组合，从而最小化损失函数。

图 (graph)

在 TensorFlow 中的一种计算过程展示。图中的节点表示操作。节点的连线是有指向性的，表示传递一个操作（一个张量）的结果（作为一个操作数）给另一个操作。使用 TensorBoard 能可视化计算图。

H

启发式 (heuristic)

一个问题的实际的和非最优的解，但能从学习经验中获得足够多的进步。

隐藏层 (hidden layer)

神经网络中位于输入层（即特征）和输出层（即预测）之间的合成层。一个神经网络包含一个或多个隐藏层。

折页损失函数 (Hinge loss)

损失函数的一个类型，用于分类模型以寻找距离每个样本的距离最大的决策边界，即最大化样本和边界之间的边缘。KSVMs 使用 hinge 损失函数（或相关的函数，比如平方 hinge 函数）。在二元分类中，hinge 损失函数按以下方式定义：

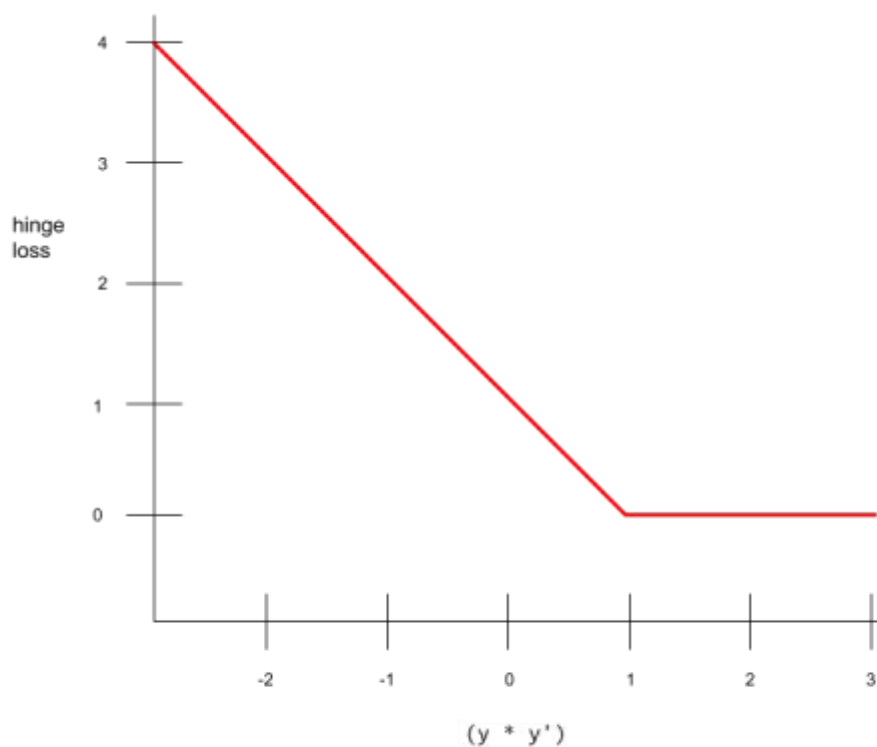
$$\text{loss} = \max(0, 1 - (y' * y))$$

其中 y' 是分类器模型的列输出：

$$y' = b + w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

y 是真实的标签，-1 或 +1。

因此，hinge 损失将是下图所示的样子：



测试数据 (holdout data)

有意不用于训练的样本。验证数据集 (validation data set) 和测试数据集 (test data set) 是测试数据 (holdout data) 的两个例子。测试数据帮助评估模型泛化到除了训练数据之外的数据的能力。测试集的损失比训练集的损失提供了对未知数据集的损失更好的估计。

超参数 (hyperparameter)

连续训练模型的过程中可以拧动的「旋钮」。例如，相对于模型自动更新的参数，学习率 (learning rate) 是一个超参数。和参量对照。

I

独立同分布 (independently and identically distributed , i.i.d)

从不会改变的分布中获取的数据，且获取的每个值不依赖于之前获取的值。i.i.d. 是机器学习的理想情况——一种有用但在现实世界中几乎找不到的数学构建。例如，网页访客的分布可能是短暂时间窗口上的 i.i.d；即分布不

会在该时间窗口发生改变，每个人的访问都与其他人的访问独立。但是，如果你扩展了时间窗口，则会出现网页访客的季节性差异。

推断 (inference)

在机器学习中，通常指将训练模型应用到无标注样本来进行预测的过程。在统计学中，推断指在观察到的数据的基础上拟合分布参数的过程。

输入层 (input layer)

神经网络的第一层（接收输入数据）。

评分者间一致性 (inter-rater agreement)

用来衡量一项任务中人类评分者意见一致的指标。如果意见不一致，则任务说明可能需要改进。有时也叫标注者间信度 (inter-annotator agreement) 或评分者间信度 (inter-rater reliability) 。

K

Kernel 支持向量机 (Kernel Support Vector Machines/KSVM)

一种分类算法，旨在通过将输入数据向量映射到更高维度的空间使正类和负类之间的边际最大化。例如，考虑一个输入数据集包含一百个特征的分类问题。为了使正类和负类之间的间隔最大化，KSVM 从内部将特征映射到百万维度的空间。KSVM 使用的损失函数叫作 hinge 损失。

L

L1 损失函数 (L1 loss)

损失函数基于模型对标签的预测值和真实值的差的绝对值而定义。L1 损失函数比起 L2 损失函数对异常值的敏感度更小。

L1 正则化 (L1 regularization)

一种正则化，按照权重绝对值总和的比例进行惩罚。在依赖稀疏特征的模型中，L1 正则化帮助促使（几乎）不相关的特征的权重趋近于 0，从而从模型中移除这些特征。

L2 损失 (L2 loss)

参见平方损失。

L2 正则化 (L2 regularization)

一种正则化，按照权重平方的总和的比例进行惩罚。L2 正则化帮助促使异常值权重更接近 0 而不趋近于 0。（可与 L1 正则化对照阅读。）L2 正则化通常改善线性模型的泛化效果。

标签 (label)

在监督式学习中，样本的「答案」或「结果」。标注数据集中的每个样本包含一或多个特征和一个标签。比如，在房屋数据集中，特征可能包括卧室数量、卫生间数量、房龄，而标签可能就是房子的价格。在垃圾邮件检测数据集中，特征可能包括主题、发出者何邮件本身，而标签可能是「垃圾邮件」或「非垃圾邮件」。

标注样本 (labeled example)

包含特征和标签的样本。在监督式训练中，模型从标注样本中进行学习。

lambda

正则化率的同义词。（该术语有多种含义。这里，我们主要关注正则化中的定义。）

层 (layer)

神经网络中的神经元序列，可以处理输入特征序列或神经元的输出。

它也是 TensorFlow 的一种抽象化概念。层是将张量和配置选项作为输入、输出其他张量的 Python 函数。一旦必要的张量出现，用户就可以通过模型函数将结果转换成估计器。

学习率 (learning rate)

通过梯度下降训练模型时使用的一个标量。每次迭代中，梯度下降算法使学习率乘以梯度，乘积叫作 gradient step。

学习率是一个重要的超参数。

最小二乘回归 (least squares regression)

通过 L2 损失最小化进行训练的线性回归模型。

线性回归 (linear regression)

对输入特征的线性连接输出连续值的一种回归模型。

logistic 回归 (logistic regression)

将 sigmoid 函数应用于线性预测，在分类问题中为每个可能的离散标签值生成概率的模型。尽管 logistic 回归常用于二元分类问题，但它也用于多类别分类问题（这种情况下，logistic 回归叫作「多类别 logistic 回归」或「多项式 回归」。

对数损失函数 (Log Loss)

二元 logistic 回归模型中使用的损失函数。

损失

度量模型预测与标签距离的指标，它是度量一个模型有多糟糕的指标。为了确定损失值，模型必须定义损失函数。例如，线性回归模型通常使用均方差作为损失函数，而 logistic 回归模型使用对数损失函数。

M

机器学习 (machine learning)

利用输入数据构建（训练）预测模型的项目或系统。该系统使用学习的模型对与训练数据相同分布的新数据进行有用的预测。机器学习还指与这些项目或系统相关的研究领域。

均方误差 (Mean Squared Error/MSE)

每个样本的平均平方损失。MSE 可以通过平方损失除以样本数量来计算。TensorFlow Playground 展示「训练损失」和「测试损失」的值是 MSE。

小批量 (mini-batch)

在训练或推断的一个迭代中运行的整批样本的一个小的随机选择的子集。小批量的大小通常在 10 到 1000 之间。在小批量数据上计算损失比在全部训练数据上计算损失要高效的多。

小批量随机梯度下降 (mini-batch stochastic gradient descent)

使用小批量的梯度下降算法。也就是，小批量随机梯度下降基于训练数据的子集对 梯度进行评估。Vanilla SGD 使用 size 为 1 的小批量。

模型 (model)

机器学习系统从训练数据中所学内容的表示。该术语有多个含义，包括以下两个相关含义：

- TensorFlow 图，显示如何计算预测的结构。
- TensorFlow 图的特定权重和偏差，由训练决定。

模型训练 (model training)

确定最佳模型的过程。

动量 (Momentum)

一种复杂的梯度下降算法，其中的学习步不只依赖于当前步的导数，还依赖于先于它的步。动量包括随着时间计算梯度的指数加权移动平均数，类似于物理学中的动量。动量有时可以阻止学习陷于局部最小值。

多类别 (multi-class)

在多于两类的类别中进行分类的分类问题。例如，有约 128 种枫树，那么分类枫树品种的模型就是多类别的。反之，把电子邮件分成两个类别（垃圾邮件和非垃圾邮件）的模型是二元分类器模型。

N

NaN trap

训练过程中，如果模型中的一个数字变成了 NaN，则模型中的很多或所有其他数字最终都变成 NaN。NaN 是「Not a Number」的缩写。

负类 (negative class)

在二元分类中，一个类别是正类，另外一个负类。正类就是我们要找的目标，负类是另外一种可能性。例如，医疗测试中的负类可能是「非肿瘤」，电子邮件分类器中的负类可能是「非垃圾邮件」。

神经网络 (neural network)

该模型从大脑中获取灵感，由多个层组成（其中至少有一个是隐藏层），每个层包含简单的连接单元或神经元，其后是非线性。

神经元 (neuron)

神经网络中的节点，通常输入多个值，生成一个输出值。神经元通过将激活函数（非线性转换）应用到输入值的加权和来计算输出值。

归一化 (normalization)

将值的实际区间转化为标准区间的过程，标准区间通常是-1 到+1 或 0 到 1。例如，假设某个特征的自然区间是 800 到 6000。通过减法和分割，你可以把那些值标准化到区间-1 到+1。参见缩放。

numpy

Python 中提供高效数组运算的开源数学库。pandas 基于 numpy 构建。

O

目标 (objective)

算法尝试优化的目标函数。

离线推断 (offline inference)

生成一组预测并存储，然后按需检索那些预测。可与在线推断对照阅读。

one-hot 编码 (one-hot encoding)

一个稀疏向量，其中：

- 一个元素设置为 1。
- 所有其他的元素设置为 0。

独热编码常用于表示有有限可能值集合的字符串或标识符。例如，假设一个记录了 15000 个不同品种的植物数据集，每一个用独特的字符串标识符来表示。作为特征工程的一部分，你可能将那些字符串标识符进行独热编码，每个向量的大小为 15000。

一对多 (one-vs.-all)

给出一个有 N 个可能解决方案的分类问题，一对多解决方案包括 N 个独立的二元分类器——每个可能的结果都有一个二元分类器。例如，一个模型将样本分为动物、蔬菜或矿物，则一对多的解决方案将提供以下三种独立的二元分类器：

- 动物和非动物
- 蔬菜和非蔬菜
- 矿物和非矿物

在线推断 (online inference)

按需生成预测。可与离线推断对照阅读。

运算 (Operation/op)

TensorFlow 图中的一个节点。在 TensorFlow 中，任何创建、控制或损坏张量的步骤都是运算。例如，矩阵乘法是一个把两个张量作为输入、生成一个张量作为输出的运算。

优化器 (optimizer)

梯度下降算法的特定实现。TensorFlow 的基类优化器是 `tf.train.Optimizer`。不同的优化器（`tf.train.Optimizer` 的子类）对应不同的概念，如：

- 动量 (Momentum)
- 更新频率 (AdaGrad = ADAptive GRADient descent ; Adam = ADAptive with Momentum ; RMSProp)
- 稀疏性 / 正则化 (Ftrl)
- 更复杂的数学 (Proximal 及其他)

你甚至可以想象 NN-driven optimizer。

异常值 (outlier)

与大多数值差别很大的值。在机器学习中，下列都是异常值：

- 高绝对值的权重。
- 与实际值差距过大的预测值。
- 比平均值多大约 3 个标准差的输入数据的值。

异常值往往使模型训练中出现问题。

输出层 (output layer)

神经网络的「最后」一层。这一层包含整个模型所寻求的答案。

过拟合 (overfitting)

创建的模型与训练数据非常匹配，以至于模型无法对新数据进行正确的预测。

P

pandas

一种基于列的数据分析 API。很多机器学习框架，包括 TensorFlow，支持 pandas 数据结构作为输入。参见 pandas 文档。

参数 (parameter)

机器学习系统自行训练的模型的变量。例如，权重是参数，它的值是机器学习系统通过连续的训练迭代逐渐学习到的。可与超参数对照阅读。

参数服务器 (Parameter Server/PS)

用于在分布式设置中跟踪模型参数。

参数更新 (parameter update)

在训练过程中调整模型参数的操作，通常在梯度下降的单个迭代中进行。

偏导数 (partial derivative)

一个多变量函数的偏导数是它关于其中一个变量的导数，而保持其他变量恒定。例如， $f(x, y)$ 对于 x 的偏导数就是 $f(x)$ 的导数， y 保持恒定。 x 的偏导数中只有 x 是变化的，公式中其他的变量都不用变化。

分区策略 (partitioning strategy)

在多个参数服务器中分割变量的算法。

性能 (performance)

具有多种含义：

- 在软件工程中的传统含义：软件运行速度有多快 / 高效？
- 在机器学习中的含义：模型的准确率如何？即，模型的预测结果有多好？

困惑度 (perplexity)

对模型完成任务的程度的一种度量指标。例如，假设你的任务是阅读用户在智能手机上输入的单词的头几个字母，并提供可能的完整单词列表。该任务的困惑度 (perplexity, P) 是为了列出包含用户实际想输入单词的列表你需要进行的猜测数量。

困惑度和交叉熵的关系如下：

$$P = 2^{-crossentropy}$$

流程 (pipeline)

机器学习算法的基础架构。管道包括收集数据、将数据放入训练数据文件中、训练一或多个模型，以及最终输出模型。

正类 (positive class)

在二元分类中，有两种类别：正类和负类。正类是我们测试的目标。（不过必须承认，我们同时测试两种结果，但其中一种不是重点。）例如，医疗测试中正类可能是「肿瘤」，电子邮件分类器中的正类可能是「垃圾邮件」。可与负类对照阅读。

精度 (precision)

分类模型的一种指标。准确率指模型预测正类时预测正确的频率。即：

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

预测 (prediction)

模型在输入样本后的输出结果。

预测偏差 (prediction bias)

揭示预测的平均值与数据集中标签的平均值的差距。

预制评估器 (pre-made Estimator)

已经构建好的评估器。TensorFlow 提供多个预制评估器，包括 DNNClassifier、DNNRegressor 和 LinearClassifier。你可以根据指导 (<https://www.tensorflow.org/extend/estimators>) 构建自己的预制评估器。

预训练模型 (pre-trained model)

已经训练好的模型或模型组件（如嵌入）。有时，你将预训练嵌入馈送至神经网络。其他时候，你的模型自行训练嵌入，而不是依赖于预训练嵌入。

先验信念 (prior belief)

训练开始之前你对数据的信念。例如，L2 正则化依赖于权重值很小且正常分布在 0 周围的信念。

Q

队列 (queue)

实现队列数据结构的 TensorFlow 操作。通常在输入 / 输出 (I/O) 中使用。

R

秩 (rank)

机器学习领域中包含多种含义的术语：

- 张量中的维度数量。比如，标量有 1 个秩，向量有 1 个秩，矩阵有 2 个秩。（注：在这个词汇表中，「秩」的概念和线性代数中「秩」的概念不一样，例如三阶可逆矩阵的秩为 3。）
- 机器学习问题中类别的序数位置，按从高到低的顺序给类别分类。比如，行为排序系统可以把狗的奖励按从高（牛排）到低（甘蓝）排序。

评分者 (rater)

为样本提供标签的人，有时也叫「标注者」。

召回率 (recall)

分类模型的一个指标，可以回答这个问题：模型能够准确识别多少正标签？即：

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

修正线性单元 (Rectified Linear Unit/ReLU)

一种具备以下规则的激活函数：

- 如果输入为负或零，则输出为 0。
- 如果输入为正，则输出与输入相同。

回归模型 (regression model)

一种输出持续值（通常是浮点数）的模型。而分类模型输出的是离散值，如「day lily」或「tiger lily」。

正则化 (regularization)

对模型复杂度的惩罚。正则化帮助防止过拟合。正则化包括不同种类：

- L1 正则化
- L2 正则化
- dropout 正则化
- early stopping（这不是正式的正则化方法，但可以高效限制过拟合）

正则化率 (regularization rate)

一种标量级，用 λ 来表示，指正则函数的相对重要性。从下面这个简化的损失公式可以看出正则化率的作用：

```
minimize(loss function +  $\lambda$ (regularization function))
```

提高正则化率能够降低过拟合，但可能会使模型准确率降低。

表征

将数据映射到有用特征的过程。

受试者工作特征曲线 (receiver operating characteristic/ROC Curve)

反映在不同的分类阈值上，真正类率和假正类率的比值的曲线。参见 AUC。

根目录 (root directory)

指定放置 TensorFlow 检查点文件子目录和多个模型的事件文件的目录。

均方根误差 (Root Mean Squared Error/RMSE)

均方误差的平方根。

S

Saver

负责存储模型检查点文件的 TensorFlow 对象。

缩放 (scaling)

特征工程中常用的操作，用于控制特征值区间，使之与数据集中其他特征的区间匹配。例如，假设你想使数据集中所有的浮点特征的区间为 0 到 1。给定一个特征区间是 0 到 500，那么你可以通过将每个值除以 500，缩放特征值区间。还可参见正则化。

scikit-learn

一种流行的开源机器学习平台。网址：www.scikit-learn.org

序列模型 (sequence model)

输入具有序列依赖性的模型。例如，根据之前观看过的视频序列对下一个视频进行预测。

会话 (session)

保持 TensorFlow 程序的状态（如变量）。

Sigmoid 函数 (sigmoid function)

把 logistic 或 多项式回归输出（对数几率）映射到概率的函数，返回的值在 0 到 1 之间。sigmoid 函数的公式如下：

$$y = \frac{1}{1 + e^{-\sigma}}$$

其中 σ 在 logistic 回归问题中只是简单的：

$$\sigma = b + w_1 x_1 + w_2 x_2 + \dots w_n x_n$$

在有些神经网络中，sigmoid 函数和激活函数一样。

softmax

为多类别分类模型中每个可能的类提供概率的函数。概率加起来的总和是 1.0。例如，softmax 可能检测到某个图像是一只狗的概率为 0.9，是一只猫的概率为 0.08，是一匹马的概率为 0.02。（也叫作 full softmax）。

稀疏特征（sparse feature）

值主要为 0 或空的特征向量。比如，一个向量的值有 1 个 1、一百万个 0，则该向量为稀疏向量。再比如，搜索查询中的单词也是稀疏向量：在一种语言中有很多可以用的单词，但给定的查询中只用了其中的一些。

可与稠密特征对照阅读。

平方损失（squared loss）

线性回归中使用的损失函数（也叫作 L2 Loss）。该函数计算模型对标注样本的预测值和标签真正值之间差的平方。在平方之后，该损失函数扩大了不良预测的影响。即，平方损失比 L1 Loss 对异常值（outlier）的反应更加强烈。

静态模型（static model）

离线训练的模型。

稳态 (stationarity)

数据集中的一种数据属性，数据分布在一或多个维度中保持不变。通常情况下，维度是时间，意味着具备平稳性的数据不会随着时间发生变化。比如，具备平稳性的数据从 9 月到 12 月不会改变。

步 (step)

一个批量中的前向和后向评估。

步长 (step size)

学习速率 (learning rate) 乘以偏导数的值，即梯度下降中的步长。

随机梯度下降 (stochastic gradient descent/SGD)

批量大小为 1 的梯度下降算法。也就是说，SGD 依赖于从数据集中随机均匀选择出的一个样本，以评估每一步的梯度。

结构风险最小化 (structural risk minimization/SRM)

这种算法平衡两个目标：

- 构建预测性最强的模型（如最低损失）。
- 使模型尽量保持简单（如强正则化）。

比如，在训练集上的损失最小化 + 正则化的模型函数就是结构风险最小化算法。更多信息，参见 <http://www.svms.org/srm/>。可与经验风险最小化对照阅读。

摘要 (summary)

在 TensorFlow 中，特定步计算的值或值的集合，通常用于跟踪训练过程中的模型指标。

监督式机器学习 (supervised machine learning)

利用输入数据及其对应标签来训练模型。监督式机器学习类似学生通过研究问题和对应答案进行学习。在掌握问题和答案之间的映射之后，学生就可以提供同样主题的新问题的答案了。可与非监督机器学习对照阅读。

合成特征 (synthetic feature)

不在输入特征中，而是从一个或多个输入特征中衍生出的特征。合成特征的类型包括：

- 特征与自己或其他特征相乘（叫作特征交叉）。
- 两个特征相除。
- 将连续的特征放进 range bin 中。

由归一化或缩放单独创建的特征不是合成特征。

T

张量 (tensor)

TensorFlow 项目的主要数据结构。张量是 N 维数据结构（N 的值很大），经常是标量、向量或矩阵。张量可以包括整数、浮点或字符串值。

张量处理单元 (Tensor Processing Unit , TPU)

优化 TensorFlow 性能的 ASIC (application-specific integrated circuit , 专用集成电路)。

张量形状 (Tensor shape)

张量的元素数量包含在不同维度中。比如，[5, 10] 张量在一个维度中形状为 5，在另一个维度中形状为 10。

张量大小 (Tensor size)

张量包含的标量总数。比如，[5, 10] 张量的大小就是 50。

TensorBoard

展示一个或多个 TensorFlow 项目运行过程中保存的摘要数据的控制面板。

TensorFlow

大型分布式机器学习平台。该术语还指 TensorFlow 堆栈中的基础 API 层，支持数据流图上的通用计算。

尽管 TensorFlow 主要用于机器学习，但是它也适用于要求使用数据流图进行数值运算的非机器学习任务。

TensorFlow Playground

一个可以看到不同超参数对模型（主要是神经网络）训练的影响的平台。前往 <http://playground.tensorflow.org>，使用 TensorFlow Playground。

TensorFlow Serving

帮助训练模型使之可部署到产品中的平台。

测试集 (test set)

数据集的子集。模型经过验证集初步测试之后，使用测试集对模型进行测试。可与训练集和验证集对照阅读。

tf.Example

一种标准 protocol buffer，用于描述机器学习模型训练或推断的输入数据。

训练 (training)

确定组成模型的完美参数的流程。

训练集 (training set)

数据集子集，用于训练模型。可与验证集和测试集对照阅读。

真负类 (true negative , TN)

被模型正确地预测为负类的样本。例如，模型推断某封电子邮件不是垃圾邮件，然后该电邮真的不是垃圾邮件。

真正类 (true positive , TP)

被模型正确地预测为正类的样本。例如，模型推断某封电子邮件是垃圾邮件，结果该电邮真的是垃圾邮件。

真正类率 (true positive rate , TP rate)

召回率 (recall) 的同义词。即：

$$\text{TruePositiveRate} = \text{TruePositives} / (\text{TruePositives} + \text{FalseNegatives})$$

真正类率是 ROC 曲线的 y 轴。

U

无标签样本 (unlabeled example)

包含特征但没有标签的样本。无标签样本是推断的输入。在半监督学习和无监督学习的训练过程中，通常使用无标签样本。

无监督机器学习 (unsupervised machine learning)

训练一个模型寻找数据集（通常是无标签数据集）中的模式。

无监督机器学习最常用于将数据分成几组类似的样本。例如，无监督机器学习算法可以根据音乐的各种属性聚类数据。用这种方式收集的数据可以作为其他机器学习算法（如音乐推荐服务）的输入。聚类在难以获取真正标签的情景中非常有用。例如，在反欺诈和反滥用的情景中，聚类可以帮助人类更好地理解数据。

无监督机器学习的另一个例子是主成分分析（principal component analysis , PCA）。如，将 PCA 应用于包含数百万购物车内容的数据集中时，就有可能发现有柠檬的购物车往往也有解酸剂。可与监督式机器学习对照阅读。

V

验证集 (validation set)


数据集的一个子集（与训练集不同），可用于调整超参数。可与训练集和测试集对照阅读。

W

权重 (weight)

线性模型中的特征系数，或者深度网络中的边缘。线性模型的训练目标是为每个特征确定一个完美的权重。如果权重为 0，则对应的特征对模型而言是无用的。

宽模型 (wide model)

线性模型通常具备很多稀疏输入特征。我们称之为「宽」模型，因其具有大量与输出节点直接连接的输入，是一种特殊类型的神经网络。宽模型通常比深度模型更容易调试 (debug) 和检查。尽管宽模型无法通过隐藏层表达非线性，但它们可以使用特征交叉和 bucketization 等转换用不同方式对非线性建模。可与深度模型对照阅读。 SYNCED

原文链接 : <https://developers.google.com/machine-learning/glossary>

本文为机器之心编译，转载请联系本公众号获得授权。

✂️-----

加入机器之心 (全职记者/实习生) : hr@jiqizhixin.com

投稿或寻求报道 : content@jiqizhixin.com

广告&商务合作 : bd@jiqizhixin.com