数据集详细说明参看我的 CSDN 博客链接：

https://blog.csdn.net/qq_38384924/article/details/97128744

数据集下载：请自行下载到本地

数据集的解释：

数据集的每一行表示一条网络连接数据样本，每个样本有 42 个属性，最后一个属性为攻击类别；在训练集和测试集中每条连接被标注为 normal 或 attack（异常行为），异常行为被细分为四大类共 39 种攻击。

注意：训练集中共出现了 22 个攻击类型，而剩下的 17 种只在测试集中出现，目的是检验分类器模型的泛化能力，对未知攻击类型的检测能力是评价入侵检测系统好坏的重要指标。

四大类攻击：

（1）DoS（denial-of-service）拒绝服务攻击,有：

apache2 ,back ,land ,mailbomb ,neptune ,pod ,processtable ,smurf ,teardrop , u d pstorm

（2）R2L（Remote to Local） 远程用户攻击，有：

ftp_write ,guess_passwd, imap, multihop ,named ,phf,sendmail ,snmpgetattack ,s nmpguess , spy ,warezclient ,warezmaster ,worm, xlock, xsnoop

（3）U2R（User to Root） U2R 攻击又叫提权攻击，有：

buffer_overflow ,httptunnel ,loadmodule,perl ,ps ,rootkit ,sqlattack ,xterm

（4）Probe（Surveillance or Probe）攻击采用端口扫描的方式获得信息,有:

ipsweep ,mscan ,nmap ,portsweep ,saint ,satan

特征描述：

每条数据都用 41 维特征向量表示，在这 41 维特征数据中，共有 9 维特征是离散的。其余都是连续型特征。

第 1-9 位特征表示 TCP 连接的基本特征。

第 10-22 位特征表示 TCP 连接的内容特征。

第 23-31 位特征表示基于时间的网络流量统计特征。

第 32-41 位特征表示基于主机的网络流量统计特征。

以下为 42 个属性：

1、duration：连接持续时间，以秒为单位，连续类型。范围是 [0, 58329]。

2、protocol_type：协议类型，离散类型，共有 3 种：TCP, UDP, ICMP。

3、service：目标主机的网络服务类型，离散类型，共有 70 种

4、flag：连接正常或错误的状态，离散类型，共 11 种。

5、src_bytes：从源主机到目标主机的数据的字节数，连续类型，范围是 [0, 1379963888]。

6、dst_bytes：从目标主机到源主机的数据的字节数，连续类型，范围是 [0. 1309937401]。

7、land：若连接来自/送达同一个主机/端口则为 1，否则为 0，离散类型，0 或 1。

8、wrong_fragment：错误分段的数量，连续类型，范围是 [0, 3]。

9、urgent：加急包的个数，连续类型，范围是[0, 14]。

10、hot：访问系统敏感文件和目录的次数，连续，范围是 [0, 101]。

11、num_failed_logins：登录尝试失败的次数。连续，[0, 5]。

12、logged_in：成功登录则为 1，否则为 0，离散，0 或 1。

13、num_compromised：compromised 条件（**）出现的次数，连续，[0, 7479]。

14、root_shell：若获得 root shell 则为 1，否则为 0，离散，0 或 1。root_shell 是指获得超级用户权限。

15、su_attempted：若出现”su root” 命令则为 1，否则为 0，离散，0 或 1。

16、num_root：root 用户访问次数，连续，[0, 7468]。
17、num_file_creations：文件创建操作的次数，连续，[0, 100]。
18、num_shells：使用 shell 命令的次数，连续，[0, 5]。
19、num_access_files：访问控制文件的次数，连续，[0, 9]。
20、num_outbound_cmds： 一个 FTP 会话中出站连接的次数，连续，0。数据集中这一特征出现次数为 0。
21、is_host_login：登录是否属于“hot”列表（***），是为 1，否则为 0，离散，0 或 1。
22、is_guest_login：若是 guest 登录则为 1，否则为 0，离散，0 或 1。
23、count：过去两秒内，与当前连接具有相同的目标主机的连接数，连续，[0, 511]。
24、srv_count：过去两秒内，与当前连接具有相同服务的连接数，连续，[0, 511]。
25、serror_rate：过去两秒内，在与当前连接具有相同目标主机的连接中，出现“SYN” 错误的连接的百分比，连续，[0.00, 1.00]。
26、srv_serror_rate：过去两秒内，在与当前连接具有相同服务的连接中，出现“SYN” 错误的连接的百分比，连续，[0.00, 1.00]。
27、rerror_rate：过去两秒内，在与当前连接具有相同目标主机的连接中，出现“REJ” 错误的连接的百分比，连续，[0.00, 1.00]。
28、srv_rerror_rate：过去两秒内，在与当前连接具有相同服务的连接中，出现“REJ” 错误的连接的百分比，连续，[0.00, 1.00]。
29、same_srv_rate：过去两秒内，在与当前连接具有相同目标主机的连接中，与当前连接具有相同服务的连接的百分比，连续，[0.00, 1.00]。
30、diff_srv_rate：过去两秒内，在与当前连接具有相同目标主机的连接中，与当前连接具有不同服务的连接的百分比，连续，[0.00, 1.00]。
31、srv_diff_host_rate：过去两秒内，在与当前连接具有相同服务的连接中，与当前连接具有不同目标主机的连接的百分比，连续，[0.00, 1.00]。
32、dst_host_count：前 100 个连接中，与当前连接具有相同目标主机的连接数，连续，[0, 255]。
33、dst_host_srv_count：前 100 个连接中，与当前连接具有相同目标主机相同服务的连接数，连续，[0, 255]。
34、dst_host_same_srv_rate：前 100 个连接中，与当前连接具有相同目标主机相同服务的连接所占的百分比，连续，[0.00, 1.00]。
35、dst_host_diff_srv_rate：前 100 个连接中，与当前连接具有相同目标主机不同服务的连接所占的百分比，连续，[0.00, 1.00]。
36、dst_host_same_src_port_rate：前 100 个连接中，与当前连接具有相同目标主机相同源端口的连接所占的百分比，连续，[0.00, 1.00]。
37、dst_host_srv_diff_host_rate：前 100 个连接中，与当前连接具有相同目标主机相同服务的连接中，与当前连接具有不同源主机的连接所占的百分比，连续，[0.00, 1.00]。
38、dst_host_serror_rate：前 100 个连接中，与当前连接具有相同目标主机的连接中，出现 SYN 错误的连接所占的百分比，连续，[0.00, 1.00]。
39、dst_host_srv_serror_rate：前 100 个连接中，与当前连接具有相同目标主机相同服务的连接中，出现 SYN 错误的连接所占的百分比，连续， [0.00, 1.00]。
40、dst_host_rerror_rate：前 100 个连接中，与当前连接具有相同目标主机的连接中，出现 REJ 错误的连接所占的百分比，连续，[0.00, 1.00]。
41、dst_host_srv_rerror_rate：前 100 个连接中，与当前连接具有相同目标主机相同服务的连接中，出现 REJ 错误的连接所占的百分比，连续，[0.00, 1.00]。
42、label（类别）