Homework 1

使用 Support Vector Machine (SVM) 解决一个图像的二类别分类问题。数据集是从 FashionMNIST 数据集中采样出的两个类别（pullover和dress）。

数据集信息

训练集包含 6000 张图片，存储在 X_train_sampled.npy 文件中，每张图片的大小为 42×42；训练样本的标签存储在 y_train_sampled.npy 文件中。 测试集包含 2000 张图片，存储在 X_test_sampled.npy 文件中，每张图片的大小为 42×42；测试样本的标签存储在 y_test_sampled.npy 文件中。

请使用 get_data.py 代码文件中的函数来加载数据集。具体的用法可以参考 main.py 中的示例，以及 get_data.py 中的注释。

实验步骤

- **Step 1.** 提取图像的 Histogram Of Gradient (HoG) 特征（见 lecture2 的第 14 页）。对于一个样本 $x \in \mathbb{R}^{42\times 42}$, 提取到的 HoG 特征向量表示为 $h \in \mathbb{R}^{2646}$。

  提取 HoG 特征的代码见 HoG.py。特征提取的详细过程可以参考 main.py 中的示例，以及 HoG.py 中的注释。

- **Step 2.** 利用提取到 HoG 特征向量 h，尝试使用不同的 SVM 分类器进行分类。你可以使用**scikit-learn 库中的 SVC 类**来实现你的 SVM。参考文档请见链接: SVC 或者文件 **SVC_document.pdf**。**SVC_document.pdf 文件中已经对本次作业中可能用到的函数/属性进行了高亮，供同学们参考。**

  你需要实现三种 SVM (with outliers, 见lecture 2 的第 13 页) 分类器：

  1. Linear SVM；
  2. RBF kernel SVM；
  3. 其他任选一种核函数的 SVM, 比如 Polynomial kernel SVM。

你需要为SVM中的各个超参数找到合适的值，例如合适的 C 的值（ξ 的系数，相关的公式见 lecture 2 的 13 页）。 在具体实现中，scikit-learn库 SVC 类里面的 C 变量即为 ξ 的系数，见下面的示例代码。

```
from sklearn.svm import SVC
svm = SVC(kernel="linear", C=0.1)
```

你可以考虑尝试 $[10^{-3}, 10^{3}]$ 范围中的C 值，比如取$C = 10^{-3}, 10^{-2}, 10^{-1}, \ldots, 10^{2}, 10^{3}$。

请汇报以下结果:

1. 在不同超参数设置下，每种SVM在测试集上的分类准确率。
2. 对于 Linear SVM，请汇报参与参数 w 的计算的支持向量有哪些？（支持向量的定义见 lecture 2 的第 7 页）具体地，请回答：
   - 一共有几个支持向量参与了参数 w 的计算？ （即，有多少训练图片满足$\alpha_i > 0$？）
   - 请可视化出训练图片中分类信心最强的5个正样本和5个负样本。这里的分类信心是指 $y_i(w^\top h_i + b)$ 的值。

- 另外，请可视化出训练图片中$\alpha_i > 0$的5个支持向量样本（或outlier样本），可以是正样本，也可以是负样本。

Hint: 你可以通过SVC类的 `support_` 属性来获取所有支持向量的索引（即，训练样本中第几个样本是支持向量），还可以通过 `dual_coef_` 属性来获取所有支持向量的 $y_i\alpha_i$ 的值（注意这里不是 α_i，而是乘上了标签y_i 的，其中 $y_i \in \{-1, 1\}$）。对这些属性的详细介绍请见链接 SVC 或者文件 **SVC_document.pdf**。以下为如何获取这些属性的示例代码：

```
from sklearn.svm import SVC
svm = SVC(kernel="linear", C=0.1)

#########################################################
### Here is your code to train the SVM on training data
#########################################################

support_vector_indices = svm.support_  # shape: (num_of_support_vectors), 返回所有支持向量在训练样本中的索引（或者说序号，从0开始），比如返回[0,3,5]就表示训练样本中第0个，第3个，第5个样本是支持向量
num_support_vectors_each_class = svm.n_support_  # shape: (num_of_classes), 返回每个类别的支持向量的个数，本次作业是二分类问题，所以返回的是一个包含2个整数的array
dual_coefs_times_label = svm.dual_coef_  # shape: (num_of_classes-1, num_of_support_vectors), 本次作业是二分类问题，所以该array的shape为(1, num_of_support_vectors)，包含了所有支持向量对应的 y_i * alpha_i 的值
```

3. 对于RBF kernel SVM，通过尝试设置不同的C值（ξ的系数），汇报在不同C值下训练的SVM中 $\alpha_i > 0$的样本数量（即有多少样本是支持向量或outlier）

注意事项

请注意，你必须自己实现代码。任何抄袭的行为都可能导致挂科。

1. 此次作业只会大致比较提交上来的 SVM 分类器的性能（测试集上的分类准确率），性能高了半个百分点并不一定会获得更高的分数。 但是，如果你的分类准确率远低于其他同学，这可能意味着你的代码中有一些bug，从而影响得分。
2. 报告和源代码应分开提交，报告为pdf格式，代码打包成zip压缩文件。
3. 源代码中应当有清晰的注释。