Due: Apr. 21, 2023 @ 11:59 p.m.

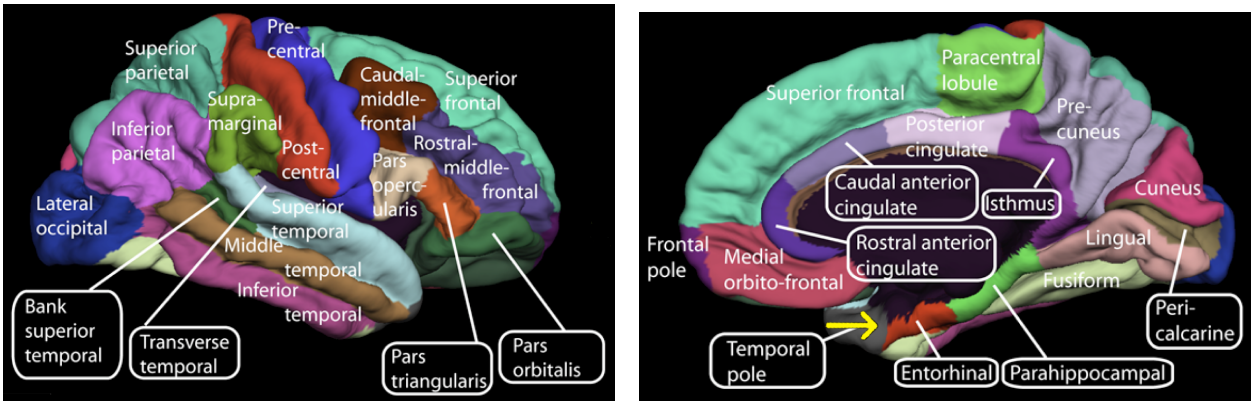
General notes to keep in mind:

- *All deliverables for the assignment must be submitted as a **single ZIP file per group** via the Brightspace D2L course shell. Submissions containing multiple ZIP files per group or those with a file that is not in the ZIP format will NOT be graded.*
- *The code submitted must be written purely using the Python programming language and it should execute within the Python 3.11.1 interpreter running on the Windows operating system (version 10 or above). The submitted code should NOT require external python modules other than scikit-learn 1.2.1 , matlplotlib 3.6.3, pandas 1.5.3 and their dependencies.*
- *Read the "Assignment code submission requirements" carefully and prepare the code accordingly. It is your responsibility to ensure that the submitted code executes. If the grader is unable to execute your code and/or your code does NOT adhere to the submission requirements, your code may not be graded.*
- *The written responses required to the questions in the assignments must be compiled into **single PDF** file named as `report.pdf`. You are encouraged to use LaTeX for typesetting your written responses, but however, the use of Microsoft Word™ or any other such programs is also acceptable.*

Alzheimer's Disease (AD) diagnosis using high-dimensional glucose metabolism measures

We revisit the problem of Alzheimer's Disease (AD) or the dementia of Alzheimer's type (DAT) diagnosis based on brain glucose metabolism changes. However, instead of just focusing on the glucose metabolism activity in the two particular brain regions, we will now consider glucose metabolism measurements taken across several regions of the cerebral cortex of the brain. The cerebral cortex is the outermost layer of the brain and it consists of around $14-16$ billion neurons making up approximately 40% of the mass of the brain. The cerebral cortex is divided into two hemispheres and its surface is highly folded. The "peaks" of the folds are called gyri and the "grooves" of the folds are called sulci [1]. Figure 1 illustrates the various regions within the cerebral cortex of a human brain also referred to as "cortical" brain regions.

The specific goal of this assignment is to develop a **support vector machine (SVM) classifier that can predict if an individual belongs to the stable normal controls (sNC) group or the stable DAT (sDAT) group** based on a high-dimensional glucose metabolism signature taken from several regions in the individual's brain.



(a) Lateral surface of the human cerebral cortex

(b) Medial surface of the human cerebral cortex

Figure 1: Anatomical regions of the cerebral cortex of the human brain. ©Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey CJ, et al. (2008) Mapping the Structural Core of Human Cerebral Cortex. PLOS Biology 6(7): e159.

[1] https://en.wikipedia.org/wiki/Cerebral_cortex

Data

The data for this assignment can be downloaded from here. The "training" dataset consists of glucose metabolism features taken from 14 cortical brain regions (see `fdg_pet.feature.info.txt` for a list of these regions) across 237 sNC and 237 sDAT individuals given in `train.fdg_pet.sNC.csv` and `train.fdg_pet.sDAT.csv` respectively. The `test.fdg_pet.sNC.csv` and `test.fdg_pet.sDAT.csv` files correspond to a "test" dataset with the same brain glucose metabolism features taken from another 415 sNC and 122 sDAT individuals respectively. Further, brain glucose metabolism features from yet another 100 sNC and 100 sDAT individuals has been gathered and will be used as the "independent test" dataset to perform a "blinded" validation of your submitted SVM model.

Performance reporting convention

Always report *accuracy*, *sensitivity*, *specificity*, *precision*, *recall* and *balanced accuracy* performance metrics when summarizing the performance ("Err") of a classification model.

Question 1 [25 marks]

Train a *linear SVM* classification model to discriminate between the sNC and sDAT individuals based on the brain glucose metabolism features. Use the "training" dataset and a cross-validation (CV) based grid-search approach to tune the regularization parameter "C" of the linear SVM model. Using the "best" "C" setting, re-train on the entire "training" dataset to obtain the final linear SVM classification model. Estimate the "Err" of this final model on the "test" dataset. Plot and discuss the performance of the models explored during the "C" hyperparameter tuning phase as function of "C".

Question 2 [25 marks]

Now train a non-linear SVM classifier with a *polynomial kernel* for the sNC vs sDAT classification task. Use the "training" dataset and a CV based grid-search approach to tune both the regularization parameter "C" as well as the degree "d" parameter of the polynomial kernel. Using the "best" "(C, d)" setting, re-train on the entire "training" dataset to obtain the final polynomial kernel SVM model. Estimate the "Err" of this final model on the "test" dataset. Compare the performance of the polynomial kernel SVM model with that of the above obtained final linear SVM model.

Question 3 [25 marks]

Repeat the above experiment by replacing the polynomial kernel with a *radial basis function (RBF) kernel*. Note that here the RBF kernel parameter "γ" needs to be tuned instead of the "d" parameter of the polynomial kernel. Compare the performance of the final RBF kernel SVM classification model with that of both the final polynomial kernel SVM and the linear SVM models.

Question 4 [25 marks]

Leveraging the experience you gained from the experiments thus far, design the "best" SVM classifier for discriminating between the sNC and sDAT groups using the glucose metabolism features derived from the 14 cortical brain regions. You may also employ any other strategies that we have not discussed in the course to train this "best" SVM classifier. The only restriction is that, you may not use datasets other than the ones provided as part of this assignment. Submit this "best" SVM classifier as the following method:

```
def ytest = diagnoseDAT(Xtest, data_dir):
"""Returns a vector of predictions with elements "0" for sNC and "1" for sDAT,
corresponding to each of the N_test features vectors in Xtest

Xtest     N_test x 2 matrix of test feature vectors

data_dir  full path to the folder containing the following files:
          train.fdg_pet.sNC.csv, train.fdg_pet.sDAT.csv,
          test.fdg_pet.sNC.csv, test.fdg_pet.sDAT.csv
"""
```

The above method will be evaluated on the "independent test" dataset by the grader to determine the classification performance. See note below regarding the grading rubric for this question.

Note on grading

The grading for Question 1, Question 2 and Question 3 will be based on the appropriateness of the submitted code and the written responses. The grading for Question 4 will be based on the relative performance of your trained model. The submission(s) with the best performing model (referred below as 1^{st} ranked model) in terms of *balanced accuracy* (rounded to 4 decimal places) will receive full marks on Question 4 (i.e., 25 marks). All other submissions will receive marks that are proportional to the decrease in performance of their model with respect to the 1^{st} ranked model. For example, if the balanced accuracy of the model of a given submission is 10% lower than the 1^{st} ranked model, then that submission will receive 22.5 marks for Question 4.