Age regression from brain MRI

Ioannis Valasakis

I. Abstract

The objective for this coursework is to explore two different supervised learning approaches for age regression from brain MRI data. Data from 652 subjects were provided and their respective skull masks for efficient skull-stripping. A three-class brain segmentation using Gaussian Mixture Model was performed and evaluated. The three relative brain tissue volumes were calculated and three different regression approaches were used: Ridge Regression, Bayesian Ridge and Logistic Regression, using a two-cross validation approach to compare their results.

II. Methods

A. Volume-based regression using brain structure segmentation

A typical pipeline for brain regression across individuals includes: a) Co-registration (this step wasn't performed here) b) Resampling c) Skull-stripping, where the skull bone is removed from the brain image (in this case using a provided brain mask) d) Bias field correction (usually with the N4 method) e) Intensity normalisation f) Noise reduction.

The segmentation was defined as a ratio between each tissue volume and the total brain volume ratio for each subject, to include the variability factor. This forms the basis for quantification of tissue volume, further visualisation and analysis of anatomical features of the subject [1].

The Gaussian Mixture Model (GMM) was used for a known set of tissue classes, the three following: White Matter (WM), Gray Matter (GM) and Cerebrospinal Fluid (CSF). For each pixel in the image, features such as pixel intensity form patterns are classified based on the probability of them belonging to a pixel. The algorithm was set to get the mean GMM result for a reference segmentation and apply this furthermore to the rest of the segmentations, to ensure consinstency between labels.

Intensity normalisation was performed (z-score), where the mean image intensity from all pixels in an image was subtracted and divided by the standard deviation of intensities. Another very popular method is the one from [2], where the image intensities are piecewise linearly mapped onto a reference scale. Resampling was also performed with an isotropic 8mm resolution as target.

The data are modelled using a mixture of several components, where each of them possesses a simple parametric form. Each pixel's feature variation for each class need to be calculated, by assigning a probability density function (PDF), which is a convex combination, to the destination class.

To estimate the PDF parameters a parametric or a non-parametric approach can be followed: GMM, falls on the first case, as it is using Gaussian distributions [3]. For the optimisation process the Expectation-Maximization (EM) algorithm is used.

The GMM for $x \in \mathbf{R}^d$ is defined by its Kcomponents (Gaussian density with parameters $\underline{\mu}_k$and Σ_k). Each component is a multivariate Gaussian density with parameters $\theta_k = \{\mu_k,\ \Sigma_k\}$:

$$p_k(X|\theta_k) = \frac{1}{(2\pi)^{d/2}|\sum_k|^{1/2}} e^{-1/2(x-\mu_k)^t \sum_k^{-1}(x-\mu_k)}$$

The EM algorithm [4] is an iterative algorithm which takes a random initial estimate γ and iteratively updates it until it convergences. The iteration includes a E-step and M-step, described as follows:

E-step: For the random initial parameter γ, a membership weight $w_{ik}$is computed for each data point $x_i,\ 1 \leq i \leq N$ and all mixture components $1 \leq k \leq K$, such that it creates a matrix with dimensions $N \cdot K$ with membership weights and where the sum of each row is equal to one, i.e. the membership weights are defined such as $\sum_{k=1}^{K} w_{ik} = 1$

M-step: New parameter values are estimated using the membership weights and the data. With $L_k = \sum_{i=1}^{L} w_{ik}$ the effective number of assigned data points to the component k. The new mixture weights are:

$$a_k^{\text{new}} = \frac{L_k}{L}, 1 \leq k \leq K$$

and

$$\mu_k^{\text{new}} = \left(\frac{1}{L_k}\right) \sum_{i=1}^{L} w_{ik} \cdot x_i, 1 \leq k \leq K$$

This vector equation uses $d-$dimensional vectors to compute the updated mean with the fractional weight w_{ik} and similarly to an empirical covariance matrix (with an extra weight included):

$$\sum_k^{\text{new}} = \left(\frac{1}{L_k}\right) \sum_{i=1}^{L} w_{ik} \cdot (x_i - \mu_k^{\text{new}})(x_i - \mu_k^{\text{new}})^t, 1 \leq k \leq K$$

It is important that the order of the equations in M-step is followed as specified. After those calculations, the new membership weights should be re-calculated in the E-step and sequentially the re-calculation of the E-step parameters, and so on. Each E,M pair defines an iteration.

A Ridge Regression (RR) and Bayesian Ridge (BR) have been used for the age features regression including the extracted features from the previous step.

The regression predictor variables [5] are highly correlated and the RR was used to introduce a small bias factor (ridge

Ioannis Valasakis is with King's College London

penalty) to those variables, which is a form of regularisation and an extended linear regression. Using β-coefficients that have much lower values it attempts to minimise their mean square error and therefore the impact on the trained model. Given a response variable y_i which is continuous and a set of predictors z_{ij}, by minimising [6]:

$$\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j Z_{ij}\right)^2 + \lambda \sum_j \beta_j^2$$

the parameters $\beta_j s$ are estimated, with λ controlling the model's complexity. With $\lambda = 0$, ridge regression becomes a linear regression.

Bayesian Ridge is an extension of the RR [7] by complying to two conditions:

1) the error ϵ has a normal distribution with mean 0 and known variance matrix $\sigma^2 I$
2) the least square matrix has a prior normal distribution with known mean and variance matrix, with posterior probability able to obtain using Bayes Theorem

Linear Regression was also used as a matter of comparison between different models. Support Vector Machines (SVM and SVR for Support Vector Regression) with Radial Basis Function (RBF) kernel and Decision Tree regression was also implemented on the data set.

SVM creates a hyperplane in multidimensional space to separate various classes. The hyperplane is a decision boundary between classes and it is constructed iteratively to minimize a defined error. The main aim of SVM is to search for a maximum marginal hyperplane (MMH) that splits the dataset into classes, with the furthest possible orientation from its closest data points [8]. Radial Basis Function (RBF) kernel is used in support vector machine classification and it can map an input space in infinite dimensional space. For a given dataset:

$$(x_1, y_1), ..., (x_n, y_n), x_i \in \mathbf{R}^d, y_i \in (-1, +1)$$

With each x_i being a feature representation in a vector format and y_i the class label of the respective training compound i, the optimal hyperplane is defined as:

$$wx^T + b = 0$$

with w being the weight vector and b the bias. The aim is to train the SVM model to find those parameters such as the hyperplane separates the data and maximises $1 \div ||w||^2$. A kernel function can be added to add additional dimensions to the raw data therefore creating a linear problem in the resulting higher dimensional space.

Regression Decision Trees (RDT) attempt to segment the predictor space into regions. A set of splitting rules is defined and the prediction for a given observation is performed by using the mean of the data in the specific region. Those rules can be described by a tree analogy:

For a prediction of a response for class Y from inputs $X1, X2....X_p$ a binary tree can be created. For each internal node, a test is performed in an input, such as X_i. A decision to follow a sub-branch of that tree is made, depending on the previous test and when a leaf node is reached, a prediction is made. This prediction is an average of all the training data points which reach this specific leaf.

The decision tree algorithm is non-parametric and can deal with efficiency for large, complicated datasets without creating a difficult to handle parametric structure [9]. A simple decision tree is shown in the Fig. 1.

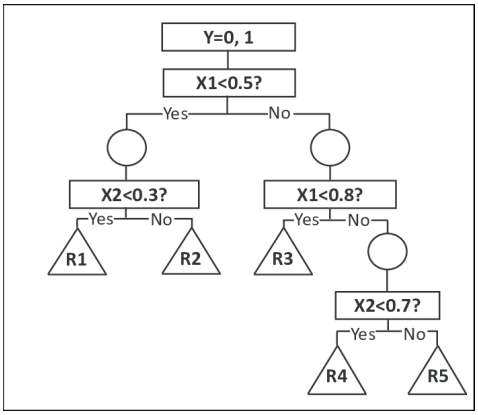

Figure 1. A simple decision tree with two parameter (binary) target variable Y

B. Image-based regression using grey matter maps

The gray matter maps were extracted from the given MRI scan data and using a common reference space they aligned to obtain spatially normalised maps. A state-of-the-art neuroimaging toolkit was used, named SPM12. The reference space corresponds to the MNI atlas.

That allow to have locations of voxels across subjects which correspond to the same anatomical locations and that allows each voxel location to be treated as a subjective individual feature. That means that those maps are very large and a dimensionality reduction method using PCA can be performed prior to training and regression. A regressor using a lower dimensionality can be tested first before the final regression and feature representation which will be obtained with PCA. Gaussian

Because the grey matter maps are spatially normalised, voxel locations across images from different subjects roughly correspond to the same anatomical locations. This means that each voxel location in the grey matter maps can be treated as an individual feature. Because those maps are quite large, there would be a very large number of features to deal with. A dimensionality reduction using PCA needs to be performed before training a suitable regressor on the low-dimensional feature representation obtained with PCA. It might also be beneficial to apply some preprocessing before running PCA, which should be explored. The implemented pipeline should be evaluated using cross-validation using the same data splits as in part A, so the two approaches can be directly compared.

Downsampling was used before the PCA reduction using local averaging. A Gaussian filter was also applied using an

experimentally found suitable $\sigma = 0.65$ such as to apply smoothing, reduce noise and to compensate for errors of the spatial normalisation that had been applied to the maps.

The dimensionality was reduced on the grey matter maps using the PCA method (already implement inside the SciKit Python module). By setting different parameter values for $n_{components} \in (0, 1)$the new dimensionality of data was determined. Furthermore, the PCA was applied to both the training and testing data by fitting the PCA model to X_{train} and applying a dimensionality reduction (by using the transform function) to X_{train}, X_{test}.

After the PCA, as in the previous part the age regression was performed with three different methods: Linear Regression, SVR with RBF kernel and DT with Adaboost.

III. Results

An overview of the patient given dataset is in the Fig. 2, Fig. 6 and Fig. 8, where the gender, age distribution and the age variability are shown.

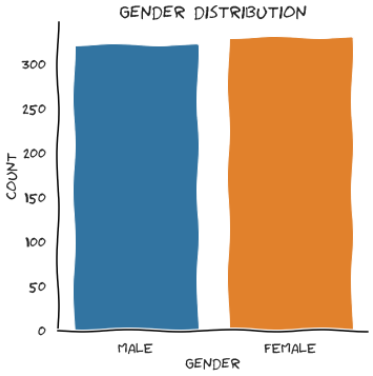

Figure 2. The gender distribution for the given subject dataset is very balanced, with a slight favour (1%) for female subjects.

The variability of the intensity of the scans was corrected using an intensity normalisation method. This standardisation of all the datasets performed satisfactory for this pipeline. The three brain areas were scanned using a label search and reordering was performed (if needed) as well for WM, GM and CSF labels.

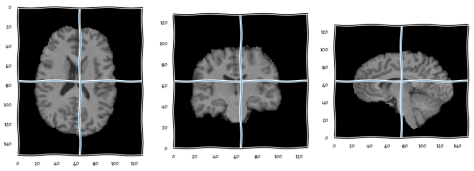

Figure 3. The skull-stripped brain, using the provided masks, for a test subject.

The images were sequentially segmented with a reference segmentation shown in Fig. 5. GMM was performed and for a chosen subject with the skull-stripped brain image shown in Fig. 3 the result of the components on the intensity histogram is shown in Fig. 4.

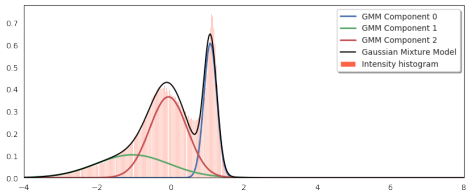

Figure 4. The GMM for one of the test subjects, with the overlapped components on the intensity histogram.

The features that were used in the regression were combinations the total volume GM, WM, CSF as well as a few different combinations (GM,CSF and WM,GM). From the resulting scoring, the most significant and accurate came from the total volume characteristic, therefore this is what is shown in the following figures. The included code has more examples and the ability to run further volume combinations.

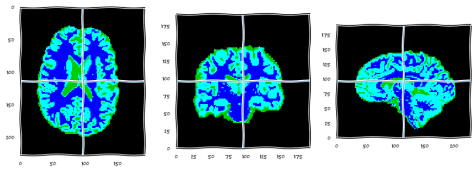

Figure 5. The segmentation result using GMM for the reference test subject.

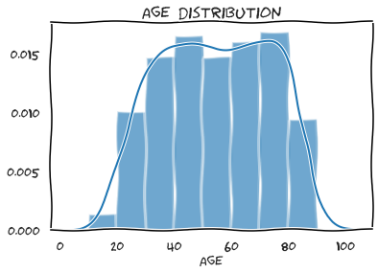

Figure 6. The age distribution on the given dataset with a distribution traced over.

For the regression, the cross-validated results are shown in the regression table in Fig. 7. The best results were achieved with SVM using RBF kernel, achieving an r2-score of 0.64, as well as the regression using DT. There were also trials using a combination of WM, GM or just one of WM, CSF, GM but those didn't achieve better results, achieving a maximum r2-score of 0.56 in the best case.

In the second part, the GM maps were smoothed using a Gaussian filter as described further in the methods. The result can be seen in the Fig. 9 and Fig. 10.

The regressions are shown in the following figures (Fig. 13, Fig. 12, Fig. 11, Fig. 14) for the image-based approach. Specifically for the last approach, the Adaptive Boosting was used

Regression	r2-score (mean, %)	RMSE (mean)
Linear Regression	55.95	12.5638
Bayessian Regression	56.05	12.5489
SVR with RBF	73	9.83623
Gradient Boosting	57.06	11.4631
Decision Tree	61.71	10.7365

Figure 7. The cross validated results for the image-based approach (using a 50% 2-fold training split, utilising scipy's K-Fold algorithm)

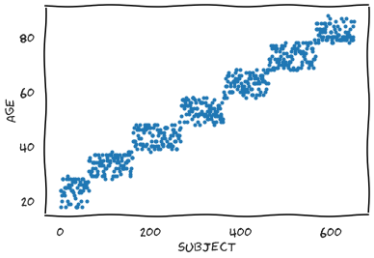

Figure 8. The age variability between the subject datasets which clearly follows a linear distributed path.

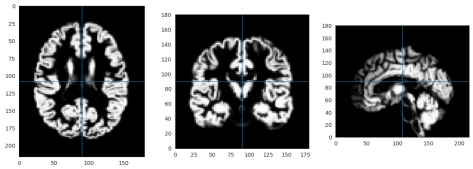

Figure 9. The GM images obtained from the subjects, after creating maps and being aligned to a common reference space

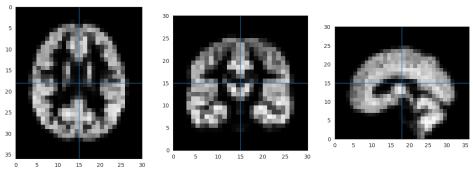

Figure 10. The downsampled and smoothed GM images using sigma=0.65

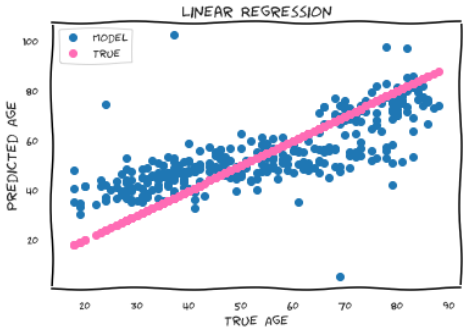

Figure 11. Linear Regression first pass

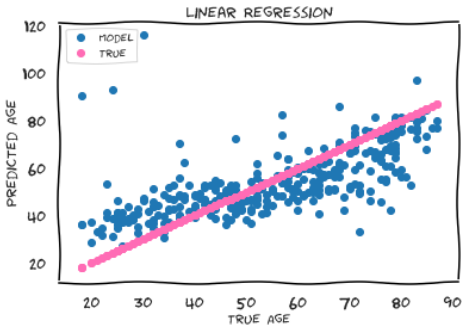

Figure 12. Linear regression, second pass with better prediction between the ages of 20-30 although still not optimal.

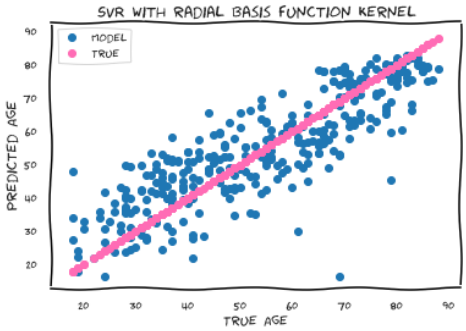

Figure 13. SVR regression with RBF kernel

on top of Decision Trees regression, where the performance of r2-score was improved by more than 15%.

For the volumetric approach, the linear regression is shown in Fig. 15 and the table with the improved scores over Fig. 16.

IV. 4 Conclusion

The best model prediction achieved an r2-score of 0.64 in the 2-fold split using SVR with RBF kernel. The advantage of SVM is that it is faster that the Bayesian regression, while offering good accuracy. It works with a clear margin of separation and with high dimensional space.

In a similar fashion, the Decision Tree regression is a very intuitive process and (even though in this case not exactly important) it doesn't require as much effort in the data preparation and preprocessing phase, including any normalisation. The complex relationships between the inputs and target can be simplified by dividing those input variable to subgroups. Decision trees are also robust to outliers and non-parametric. That said, it can be affected by small changes that will impact the structure of the decision tree and can be slower and having a more expensive training, if it was to be used in a bigger amount of data.

Nevertheless, Decision Tree using Adaboost can achieve a 20% up in performance and R2 score. The biggest improvements though, were seen with R2 values reaching 83%, using

Figure 14. Decision tree regression using Ada boost, with increased performance over the plain DT regression.

Figure 15. Linear Regression using the GM maps was much more sucesfull with a mean r2-score (%) of 88.21.

Regression	r2-score (mean, %)	RMSE (mean)
Linear Regression	88.21	6.50035
SVR with RBF	89.21	9.83623
Decision Tree	47.36	12.6496

Figure 16. The scores using the GM approach were much more improved in comparison. DT wasn't as sucesfull but that is more of an irregularity as it took very long time to run and didn't complete on time.

the PCA method for the pre-registered Gray Matter maps. Therefore, it can be concluded that this approach is much more promising and can be further bettered if configured for the specific dataset and by using and exploring more of the extracted features.

Given more time, further data selection process for feature representation (e.g. using Jacobians, Logarithm of Jacobians, Background tissues, Jacobian scaled WM/GM/BG) would have been made to explore if that would result in a more accurate prediction. Methods such as Gradient boosted trees are also very fitted for this kind of model prediction and they should further be implemented using the current dataset and be compared against the presented results.

REFERENCES

[1] J. Sethuraman, "A Constructive Definition of Dirichlet Priors," Tech. Rep., may 1991. [Online]. Available: https://doi.org/10.21236%2Fada238689

[2] L. G. Nyl and J. K. Udupa, "On standardizing the MR image intensity scale," *Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, dec 1999. [Online]. Available: https://doi.org/10.1002%2F%28sici%291522-2594%28199912%2942%3A6%3C1072%3A%3Aaid-mrm11%3E3.0.co%3B2-m

[3] M. A. Balafar, "Gaussian mixture model based segmentation methods for brain MRI images," *Artificial Intelligence Review*, vol. 41, no. 3, pp. 429–439, mar 2012. [Online]. Available: https://doi.org/10.1007%2Fs10462-012-9317-3

[4] P. Smyth and H. Tang, "The EM Algorithm for Gaussian Mixtures," Tech. Rep., 2019.

[5] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, feb 1970. [Online]. Available: https://doi.org/10.1080%2F00401706.1970.10488634

[6] S. A. Baldwin and M. J. Larson, "An introduction to using Bayesian linear regression with clinical data," *Behaviour Research and Therapy*, vol. 98, pp. 58–75, nov 2017. [Online]. Available: https://doi.org/10.1016%2Fj.brat.2016.12.016

[7] H. M. Nguyen, G. Kalra, T. J. Jun, and D. Kim, "A Novel Echo State Network Model Using Bayesian Ridge Regression and Independent Component Analysis," pp. 24–34, 2018. [Online]. Available: https://doi.org/10.1007%2F978-3-030-01421-6_3

[8] "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, jan 2018. [Online]. Available: https://doi.org/10.21873%2Fcgp.20063

[9] Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction." *Shanghai Arch Psychiatry*, vol. 27, pp. 130–5, Apr 2015.

[10] G. C. Monté-Rubio, C. Falcón, E. Pomarol-Clotet, and J. Ashburner, "A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods," *NeuroImage*, vol. 178, pp. 753–768, sep 2018. [Online]. Available: https://doi.org/10.1016%2Fj.neuroimage.2018.05.065