GENDER RECOGNITION BY VOICE

Project for the course

MATH-412 Statistical Machine Learning,
Ecole Polytechnique Federale de Lausanne,
Fall 2017

Authors: Adrien Besson, Hippolyte Lefebvre, Grigorios Kaklamanos
Professor: Dr. Emeric Thibaud

December 22, 2017

Contents

1 Introduction

In the last decades, automatic gender recognition (AGR) from speech has grown much interest thanks to the digitization of an extensive number of applications and the development of mobile platforms [1]–[8].

The applications of AGR have increased consequently. Indeed, in general, the accuracy of gender-dependent systems is higher than the one of gender-independent systems [4]. Thus, AGR improves the prediction of other speaker traits such as age [9] and emotional state [10], [11]. It can also facilitate speech recognition by gender-based normalization [12] and is a key feature for more natural and personalized dialog systems such as Siri.

The AGR techniques are based on statistical features extracted from the speech signals such as maximum, minimum and average frequency measured in a time span. These features translate physiological differences between males and females like the length of the vocal chords or the glottal shape [13]. Among all the features, it appears that the fundamental frequency plays a crucial role in gender classification as described in many studies [1], [14], [15]. In recent works, the use of the fundamental frequency coupled with spectral components such as Mel-frequency spectral components [16] or relative spectral perceptual linear predictive coefficients [5] have demonstrated best AGR performances even in noisy environments.

In this project, we study different state-of-the-art classification methods applied to the task of gender recognition by voice. The study is based on a dataset of features extracted from 3168 subjects available on Kaggle[1] and described in details in Chapter 2. A preliminary exploratory data analysis is performed in Chapter 3 which gives many hints for the evaluation of the classification methods, described in Chapter 4. Eventually, the best classification model is tested on 11 voices recorded by the authors in Chapter 5.

[1] `https://www.kaggle.com/primaryobjects/voicegender`

2 Description of the Dataset

2.1 General considerations

The voice gender dataset[1] consists of features extracted from 3168 recorded voice samples, collected from male and female speakers. The features have been extracted from the voice recordings using tuneR[2] and seewave[3], two acoustic analysis packages of R.

The dataset takes the form of a ".csv" file where each row is composed of the following acoustical features:

- **meanfreq:** mean frequency (in kHz)
- **sd:** standard deviation of frequency
- **median:** median frequency (in kHz)
- **Q25:** first quartile (in kHz)
- **Q75:** third quartile (in kHz)
- **IQR:** interquartile range (in kHz)
- **skew:** skewness of the spectrum
- **kurt:** kurtosis
- **sp.ent:** spectral entropy
- **sfm:** spectral flatness
- **mode:** mode frequency
- **centroid:** frequency centroid
- **peakf:** peak frequency (frequency with highest energy)
- **meanfun:** average of fundamental frequency measured across acoustic signal
- **minfun:** minimum fundamental frequency measured across acoustic signal
- **maxfun:** maximum fundamental frequency measured across acoustic signal
- **meandom:** average of dominant frequency measured across acoustic signal
- **mindom:** minimum of dominant frequency measured across acoustic signal
- **maxdom:** maximum of dominant frequency measured across acoustic signal
- **dfrange:** range of dominant frequency measured across acoustic signal
- **modindx:** modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- **label:** male or female

The features are all quantitative and represent frequency characteristics of the voices.

[1] `https://www.kaggle.com/primaryobjects/voicegender`
[2] `https://cran.r-project.org/web/packages/tuneR/tuneR.pdf`
[3] `https://cran.r-project.org/web/packages/seewave/seewave.pdf`

2.2 Description of the features

Before starting the data analysis, it is important to perfectly understand the features involved in the exercise. This will be very useful in a preprocessing step, since it will allow us to remove similar or linear combination of features. It will also be a great asset when it will come to the analysis of the most important features involved in the classification process.

As already pointed out in Section 2.1, the extracted features are all related to the spectrum.

Frequency-related features The feature "meanfreq" corresponds to the mean frequency, *i.e.* a weighted average of the frequencies present in the spectrum by the amplitude of the spectral components:

$$\mu_f = \sum_{i=1}^{N} f_i y_i, \tag{2.1}$$

where N is the number of frequency components of the spectrum, f_i is the i-th frequency and y_i is the relative amplitude of the i-th component of the spectrum. As described p.163 of the seewave documentation, it is equal to the feature "centroid". The feature "sd" is the standard deviation of the frequencies, calculated as:

$$\sigma_f = \sqrt{\sum_{i=1}^{N} y_i \left(f_i - \mu_f\right)^2}. \tag{2.2}$$

The feature "median" is the median frequency calculated as the frequency where the spectrum is divided into frequency intervals of same energy. The calculation of the quartiles "IQ25" and "IQ75" is based on the same criterion. The interquartile range "IQR" is calculated as the difference between the first and the first quartile.

The feature "mode" characterizes the dominant frequency of the spectrum, *i.e.* the frequency with the highest amplitude. It is very similar to "peakf", the peak frequency, which corresponds to the frequency with the highest energy. The fundamental frequency is the lowest frequency of the spectrum.

The features "meanfun", "minfun", "maxfun", "meandom", "maxdom" and "mindom" are calculated as the mean, minimum and maximum values of the fundamental and dominant frequencies respectively, where the mean, maximum and minimum are taken across many time segments which compose a given voice recording. The feature "dfrange" is the difference between "maxdom" and "mindom" and "modindx" is the mean absolute difference of the fundamental frequencies.

In addition to the frequency-related features, we can find measures on the shape of the spectrum which may give very interesting additional information.

Skewness of the spectrum The skewness of the spectrum is a measure of its asymmetry around the mean frequency. It is calculated as follows:

$$S = \frac{1}{\sigma_f^3} \frac{\sum_{i=1}^{N} \left(f_i - \mu_f\right)^3}{N-1}. \tag{2.3}$$

From (2.3), it is clear that the sign of S gives information of the left or right asymmetry of the spectrum while the absolute value of S gives the strength of the asymmetry.

Kurtosis The kurtosis is a measure of the "tailedness" of a probability distribution. It is calculated as the fourth order moment of the frequency distribution, described below:

$$K = \frac{1}{\sigma_f^4} \frac{\sum_{i=1}^{N} (f_i - \mu_f)^4}{N-1}. \tag{2.4}$$

When $K = 3$, the frequency distribution is normal. When $K < 3$, the frequency distribution is said to be *platikurtic*, it has fewer components around the means than in the tails compared to a normal distribution. When $K > 3$, the distribution is said to be *leptokurtic* and has more components around the mean than in the tails compared to a normal distribution.

Shannon spectral entropy The Shannon entropy is used to discriminate whether the voice signal is noisy or pure [17]. It is calculated as follows:

$$H = \frac{-\sum_{i=1}^{N} y_i \log_2 (y_i)}{\log_2 (N)}. \tag{2.5}$$

If the signal is pure, then all the energy is concentrated in one frequency component, *i.e.* $y_i = \delta_{ij}$ and $H = 0$. If the signal is a white noise, then $y_i = 1/N$, $\forall i \in \{1, ..., N\}$ and $H = 1$.

Spectral flatness The spectral flatness is rather similar to the spectral entropy. It is measured as the ratio between the geometric mean and the arithmetic mean of the spectral components:

$$F = N \frac{\sqrt[N]{\prod_{i=1}^{N} y_i}}{\sum_{i=1}^{N} y_i}. \tag{2.6}$$

In case of a white noise, the spectrum is flat and $H = 1$. In case of a pure tone, the geometrical mean is equal to zero and $H = 0$.

2.3 Cleaning the dataset

From the description of the features given in Section 2.2, a first cleaning of the dataset may be achieved before starting the analysis. Indeed, several features are exactly the same or a linear combination of other features:

- The features "meanfreq" and "centroid" are exactly similar. So "centroid" has been removed;
- The following relationship holds: "IQR" = "Q75" − "Q25". "IQR" has been removed;
- The following relationship holds: "dfrange" = "maxdom" − "mindom". "dfrange" has been removed.

3 Exploratory Data Analysis

As a preliminary step, we propose to perform an exploratory data analysis. This will give us some hints about the dataset, *e.g.* the most important features, their correlation etc.

In order to have a first overview of the features, summary statistics are reported in Table 3.1. It can be noticed that the frequencies have low values, which makes sense since they are expressed in kHz. The mean fundamental frequency is about 143 Hz which is coherent with the fundamental frequencies of males and females reported in the literature [18].

Regarding the shape of the spectrum, the mean value of "skew" indicates an average right-asymmetry of the spectrum. The mean value of "kurt" shows that the frequency distribution is leptokurtic. About the flatness of the spectrum, the mean values of "sfm" and "sp.ent" seem to have an inconsistent behaviour since one is above 0.5 and the other is below. However, the high standard deviation of "sfm" makes the analysis rather difficult.

Table 3.1 Summary statistics of the features of the dataset

	meanfreq	sd	median	Q25	Q75	skew	kurt	sp.ent	sfm
mean	0.181	0.0571	0.186	0.140	0.225	3.14	36.6	0.895	0.408
std	0.0299	0.0167	0.0364	0.0487	0.0236	4.24	134	0.0450	0.178
min	0.0394	0.0184	0.0110	0.000229	0.0429	0.142	2.07	0.739	0.0369
25 %	0.164	0.0420	0.170	0.111	0.209	1.65	5.67	0.863	0.258
50 %	0.185	0.0592	0.190	0.140	0.226	2.20	8.32	0.902	0.396
75 %	0.199	0.0670	0.211	0.176	0.244	2.93	13.7	0.929	0.534
max	0.251	0.115	0.261	0.247	0.273	34.7	1310	0.982	0.843

	mode	meanfun	minfun	maxfun	meandom	mindom	maxdom	modindx
mean	0.165	0.143	0.0368	0.259	0.829	0.0526	5.05	0.174
std	0.0772	0.0323	0.0192	0.0301	0.525	0.0633	3.52	0.119
min	0.00	0.0556	0.00977	0.103	0.00781	0.00488	0.00781	0.00
25 %	0.118	0.117	0.0182	0.254	0.420	0.00781	2.07	0.0998
50 %	0.187	0.140	0.0461	0.271	0.766	0.0234	4.99	0.139
75 %	0.221	0.170	0.0479	0.277	1.18	0.0703	7.01	0.210
max	0.280	0.238	0.204	0.279	2.96	0.459	21.87	0.932

Regarding the distribution of the samples, there are 1584 recording of male voices and 1584 recording of female voices. So the classes are perfectly balanced.

Let us have a look to the correlation between the features. The correlation matrix, displayed in Figure 3.1, exhibits high correlations between "skew" and "kurt" and between "sp.ent" and "sfm", which make sense since they quantify similar quantities. It can also be noticed that "meanfreq", "median", "Q25", "Q75" are highly correlated which is self-evident given their definitions.

In the state-of-the-art, it appears that the fundamental frequency is a key feature for AGR, as stated

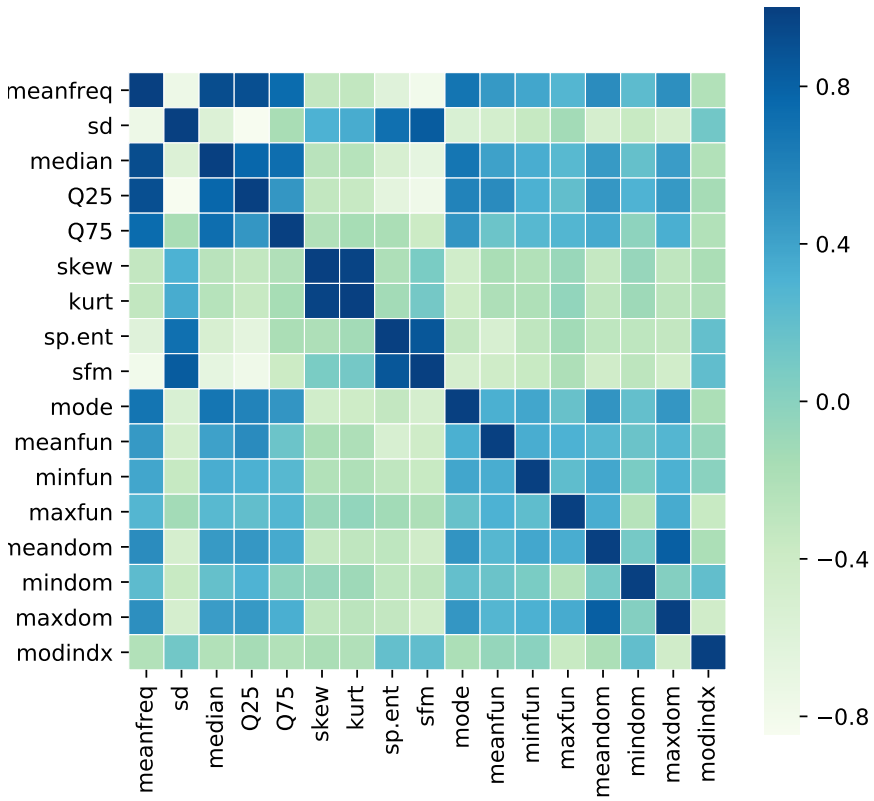
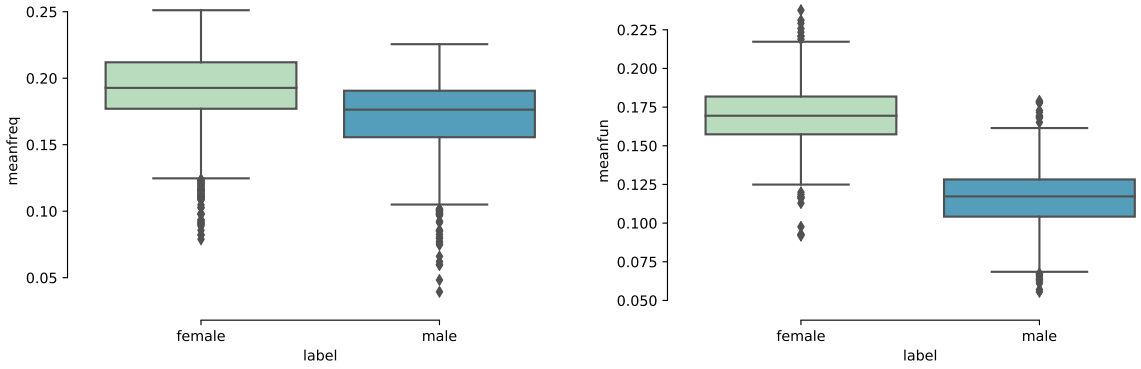


Figure 3.1 Correlation matrix of the dataset.

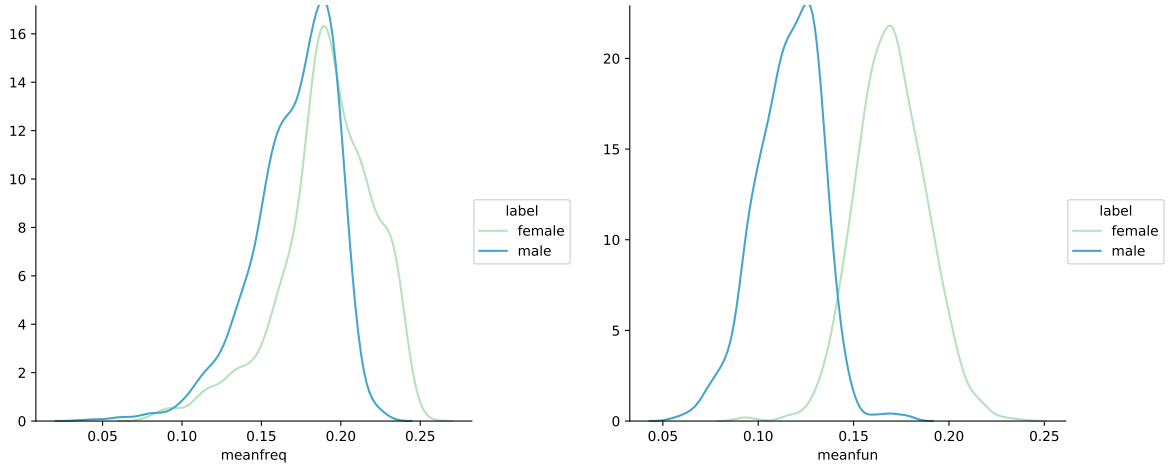
in Chapter 1. Intuitively, we also think that the mean frequency should be a good classifier. In order to analyze this, Figs. 3.2a and 3.2b represent the box plots of "meanfreq" and "meanfun" respectively. It can be noticed that "meanfun" is indeed a key feature for classification since the overlap between male and female distributions is very low. Regarding "meanfreq", the overlap is bigger than for "meanfun".

Figs. 3.3a and 3.3b represent the male and female distributions with respect to "meanfreq" and "meanfun" respectively. They substantiate the analysis made with the box plot, *i.e.* that "meanfun" is a key component in AGR and is a far better classifier than "meanfreq".



(a) (b)

Figure 3.2 Box plots for (a)-“meanfreq” and (b)-“meanfun” features.



(a) (b)

Figure 3.3 Distribution of “male” and “female” for (a)-“meanfreq” and (b)-“meanfun” features.

4 Evaluation of the Best Classification Method

4.1 Considered classification methods

The goal of our analysis is to classify $y \in \{female, male\}$ given the data matrix $\mathbf{X}$ of our 17 predictors. We are interested in determining not only the model with the best predictive performance, but also the most significant features of human voice. Therefore, we decided not to pre-process our data with a dimensionality reduction technique, such as Principal Component Analysis (PCA). Transforming the feature space with PCA would prevent us from extracting the importance of each single predictor. Furthermore, we argue that sensitivity and specificity are of equal importance in our setting and, thus, our objective is to minimize the total misclassification error. Since our classes are perfectly balanced, we use a threshold of 0.5 for our Bayes plug-in estimator. In our analysis we implement the following statistical learning methods to predict the class of $y \in \{female, male\}$.

- **Logistic regression:** models the posterior probability of response $y \in \{female, male\}$, given the predictors $\mathbf{X}$, using the logistic function;
- **Regularized logistic regression:** Ridge (ℓ_2) and Lasso (ℓ_1) penalties can be used for variable shrinkage and contribute in avoiding overfitting. Especially, Lasso regularization performs variable selection and is therefore a useful approach to examine which features are significant.
- **Linear discriminant analysis (LDA):** LDA makes the assumption that the conditional densities $f(\mathbf{X}|y = male)$ and $f(\mathbf{X}|y = female)$ are both multivariate Gaussian with a common covariance matrix. It belongs to a family of techniques that use linear boundaries to separate classes in classification problems. If the assumption of normality is realistic, then LDA is expected to provide better results than logistic regression;
- **Quadratic discriminant analysis (QDA):** QDA is a similar method to LDA. The multivariate normality assumption remains, but unlike LDA, QDA assumes that each class has its own covariance matrix. If QDA has better predictive performance than LDA, it is an indication that a linear boundary is not the optimal to separate the 2 classes.
- **k-nearest neighbors (kNN):** kNN is a non-parametric approach which is highly flexible and uses non-linear decision boundaries. Before the implementation of kNN, it is crucial to scale the data since this method relies on the euclidean distance between observations;
- **Classifications trees:** Classification trees stratify the feature space recursively into simple regions and assign the label "female" or "male" using the majority vote. Bagging, random forests and gradient tree boosting are also implemented with the aim of improving predictive performance;
- **Support vector machines (SVM):** SVM is also a non-parametric method that performs well in classification problems where there is a clear margin of separation. We experiment with 2 kernels;

the linear and the radial basis function (RBF).

For an exhaustive description of the classification methods, please refer to [19].

4.2 Classification based on the fundamental frequency

Based on the exploratory data analysis described in Chapter 3, we propose to perform the classification based on the "meanfun" feature alone. This will give a baseline for further analysis described in the remaining of this Chapter.

To perform the analysis, we randomly split the dataset into a training set (80 %) and a test set (20 %). The training set is used to fit the models and for best parameter selection if needed. The parameter selection is based on 10-fold cross-validation. The test set is used to compute the classification error estimate used to compare the models. The classification error considered in the study is the $0-1$ loss. The experiments have been performed on Python 3[1] with a single seed number for reproducibility of the results.

Table 4.1 Classification Error Estimate of the Methods for Classification With "meanfun" Feature

Type	Methods	
Max. Likelihood	Logistic reg.	0.0536
	Logistic reg. - Ridge	0.0505
	LDA	**0.0489**
	QDA	**0.0489**
Trees	Tree	0.0505
	Bagging	0.0505
	XGBoost	0.0505
SVM	Linear	0.0505
	Gaussian	**0.0489**
x	kNN	0.0520

The results, summarized in Table 4.1, show that the classification error is already very low when considering only the "meanfun" feature. Indeed the average classification error of the different classifiers is 0.0505. Thus, "meanfun" is a very good feature for classification which is in accordance with the exploratory data analysis described in Chapter 3.

Regarding the classifiers, it can be seen that some of the ones described in Section 4.1 are not mentioned in Table 4.1. Indeed, they do not make sense when considering only one feature for classification. About the relative performance of the different classifiers, the results are homogeneous around the average classification error and all the classifiers, even the simplest ones, perform well.

Such a result may be justified by the very low overlap between the "male" and "female" distributions observed in Chapter 3. Indeed, it can be noticed on Figure 3.3 that even a simple threshold of "meanfun" around a value of 0.15 would perform relatively well.

[1] `https://github.com/AdriBesson/Statistical_learning_course/tree/develop/project`

4.3 Classification based on a 80/20 split of the dataset

4.3.1 Experimental settings

To perform the analysis, we randomly split the dataset into a training set (80 %) and a test set (20 %). The training set is used to fit the models and for best parameter selection if needed. The test set is used to compute the classification error and to compare the models. The classification error is the $0-1$ loss.

The models compared in the study are the ones described in Section 4.1. The errors are computed for 5 seed numbers in order to study the variability of the models with respect to the training/test sets and their initialization. The results are reported in Table 4.2 and the discussions in the remainder of this Section 4.3 are based on the test error for the seed number 1.

Table 4.2 Classification Error Estimate of the Methods for Different Seed Numbers With 80/20 Split

Type	Methods	Seed number					Mean	Std.
		1	2	3	4	5		
Max. Lik.	Log. reg.	0.0158	0.0347	0.0315	0.0237	0.0221	0.0256	0.00760
	Log. reg. - ℓ_2	0.0158	0.0315	0.0315	0.0189	0.0284	0.0252	0.00740
	Log. reg. - ℓ_1	0.0315	0.0363	0.0379	0.0284	0.0300	0.0328	0.00408
	LDA	0.0315	0.0410	0.0379	0.0284	0.0268	0.0331	0.00611
	QDA	0.0347	0.0347	0.0347	0.0268	0.0363	0.0334	0.00377
Trees	Tree	0.0379	0.0426	0.0315	0.0284	0.0300	0.0341	0.00596
	Pruned Tree	0.0394	0.0473	0.0347	0.0363	0.0300	0.0375	0.00645
	Bagging	0.0237	0.0410	0.0142	**0.0174**	0.0284	0.0249	0.0105
	Random Forest	0.0189	0.0347	**0.0126**	0.0205	0.0205	0.0215	0.00809
	XGBoost	0.0189	**0.0268**	**0.0126**	0.0205	**0.0189**	**0.0196**	0.00506
SVM	Linear	**0.0142**	0.0315	0.0284	0.0189	0.0252	0.0237	0.00705
	Gaussian	0.0205	0.0284	**0.0126**	**0.0174**	0.0205	0.0199	0.00575
x	kNN	0.0300	0.0347	0.0252	0.0379	0.0300	0.0315	0.00486

4.3.2 Maximum-likelihood methods

Logistic regression If we run logistic regression on our 17 features, we obtain an error rate of 0.0158. The reported p-values indicate that, at 5 % significance level, the coefficients of the following features are significantly non-zero: “Q25”, “Q75”, “sp.ent”, “sfm”, “meanfun” and “minfun”. We observe that the lowest p-value corresponds to “meanfun”, which shows the importance of this particular feature in voice analysis.

Regularized logistic regression We observe that Ridge regularization gives the same results as logistic regression. However, Lasso regularization performs worse providing a test error of 0.0315. Lasso shrinks the coefficients of 9 predictors to zero, letting only non-zero the following ones: “Q25”, “Q75”, “skew”, “sfm”, “mode”, “minfun”, “maxfun” and “meanfun”. The largest coefficient corresponds to “meanfun”, confirming its significance.

[2] `https://cran.r-project.org/web/packages/MVN/index.html`

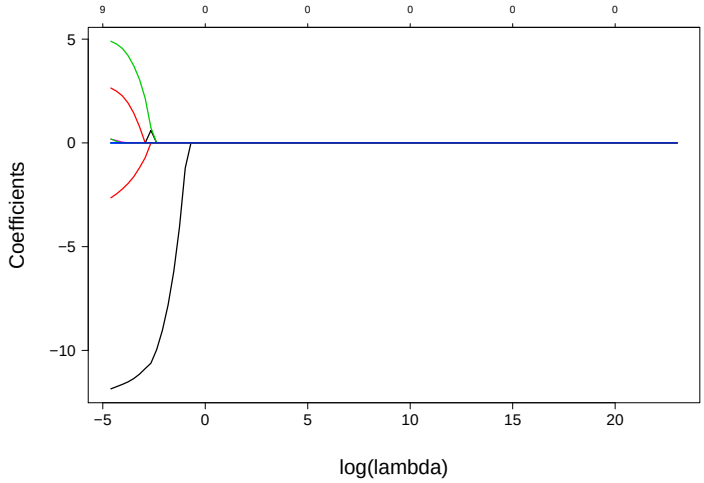


Figure 4.1 Parameters shrinkage for logistic regression with Lasso regularization.

Linear discriminant analysis LDA gives a test error of 0.0315. In order to check whether the model assumptions are actually satisfied, we resort to Mardia's test[2] for multivariate normality. Mardia's test is based on the multivariate extensions of skewness and kurtosis measures. Under the null hypothesis, our sample is drawn from a multivariate normal. We run Mardia's test on the two data matrices corresponding to male and female classes and conclude that normality assumption is violated. Furthermore, the two covariance matrices are not quite similar and one can argue that the assumption of a common covariance matrix is unrealistic. To sum up, LDA seems to perform well, despite the lack of normality. We suspect that the reason lies in some particularity of our data set.

Quadratic discriminant analysis As we have mentioned above, the two covariance matrices of our two classes (male/female) are not sufficiently close. Hence we expect QDA to perform better than LDA. However, this is not the case, since QDA yields a slightly higher test error.

4.3.3 k-Nearest neighbors

We use 10-fold cross validation with one-standard-error rule to find that the optimal value for parameter k, is $k = 9$, as described on Figure 4.2. Then we implement 9-NN, which yields a test error of 0.03. We observe a slight improvement compared to LDA and QDA, which can be attributed to the high flexibility of kNN. We reckon that the linear boundary is not the most appropriate to separate the 2 classes.

4.3.4 Tree-based methods

Classification trees We begin with fitting a large unpruned tree to our data. We use the cross-entropy impurity measure to grow our tree and obtain a test error of 0.038. Although its predictive performance is quite good in our case, its high complexity prevents it from being interpretable as Figure 4.3a manifests.

After fitting a large tree, it is sensible to prune it in order to improve its interpretability and avoid overfitting. We prune our tree using 10-fold cross validation to determine the cost-complexity parameter, and misclassification error as the loss function. Although the test error has slightly increased, the pruned tree can now be used to convey meaningful results as shown on Figure 4.3b. We also observe that "meanfun" is used in the first node, confirming, for one more time, its integral role.

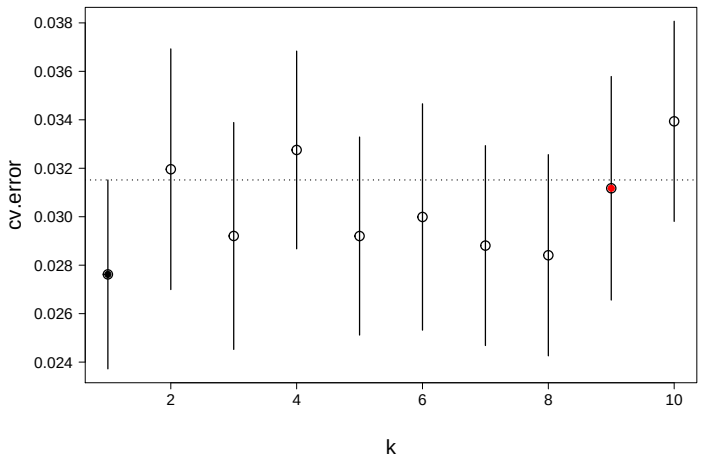
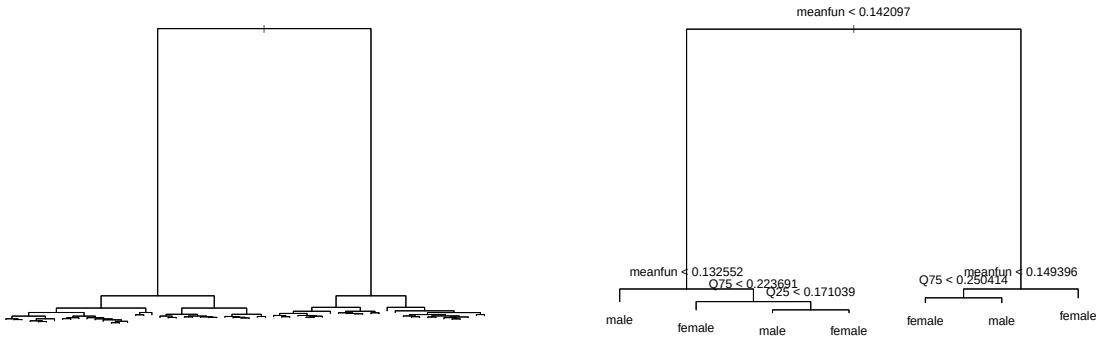


Figure 4.2 Best parameter selection for kNN based on the one-standard error rule.



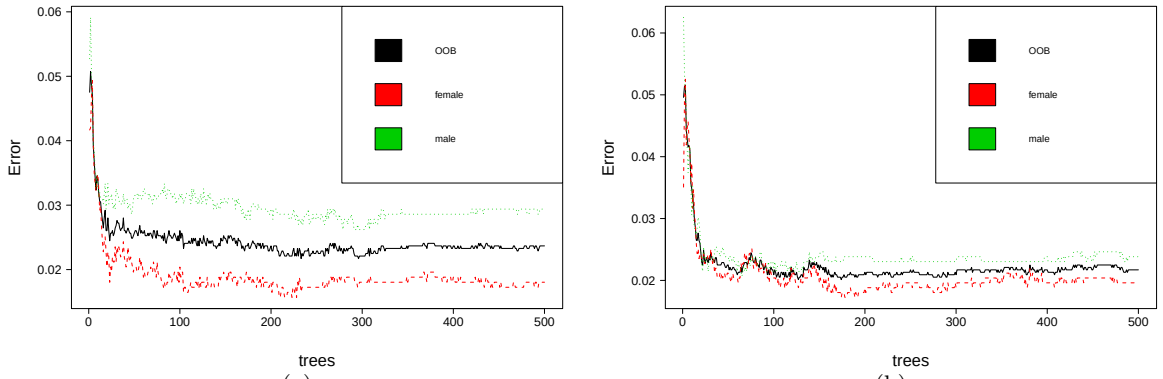
(a) (b)

Figure 4.3 (a)-Unpruned and (b)-pruned tree.

Bagging Unpruned trees have low bias but suffer from high variance. Bagging can substantially reduce the variance of unstable procedures like trees, and improve predictive performance. Bagging indeed improves accuracy of our trees, since we obtain a test error of 0.0237. Figure 4.4a depicts the evolution of class-specific training error and Out-of-Bag error estimate as the number of bagged trees increases.

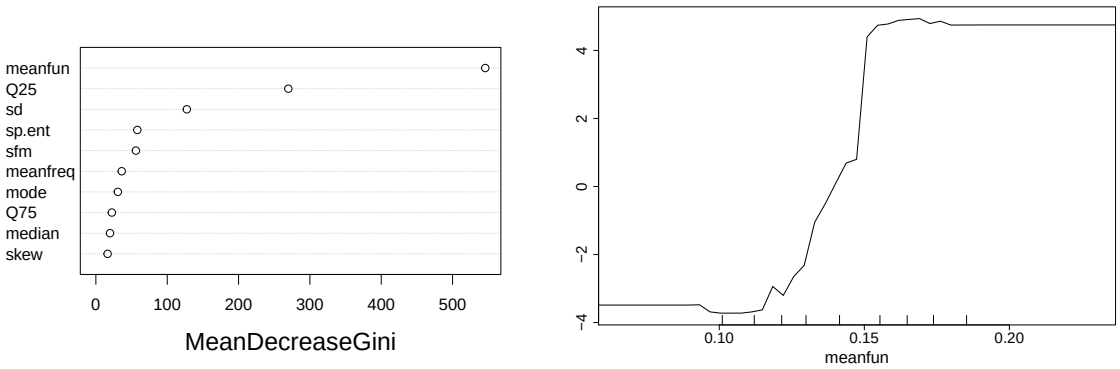
Random forests We proceed our analysis with running a random forest. We are particularly interested in this method, since random forests provide measures of variable importance. Mean decrease accuracy and mean decrease gini metrics, the latter being displayed on Fig. 4.5a, indicate that "meanfun" is the most influential feature, followed by "Q25". The test error is 0.0189, which confirms the superb predictive performance of random forests. Fig. 4.4b illustrates the efficiency of random forests, compared to bagging, in optimizing training error. Fortunately, it does not translate into overfitting in our case.

Partial dependence plots are useful to interpret variable importance in complex "black box" methods, such as random forests. These plots illustrate the marginal effect of the selected feature after integrating out the other features. In Fig. 4.5b, we present the partial dependence plot on "meanfun". The shape of the curve indicates that the fitted random forest uses a clear threshold that separate the two classes (female/male). Consequently, we can argue that the steepness of the curve in the middle pinpoints the importance of "meanfun".



(a) (b)

Figure 4.4 Out-of-bag error estimate of (a)-bagging and (b)-random forest.



(a) (b)

Figure 4.5 (a)-Feature importance based on the mean decrease of gini metric; (b)-Partial dependence with respect to the "meanfun" feature.

Gradient tree boosting In our quest of determining the model with the best accuracy, we then implement gradient tree boosting by using XGBoost R package[3]. We choose stumps, which are simple trees with two leaves, to be the slow learners. We obtain a test error of 0.0189, which confirms the superiority of boosting compared to trees and bagged trees.

4.3.5 Support vector machines

SVM with a linear kernel We apply SVM with a simple linear kernel using 10-fold cross validation for parameter selection. We obtain a test error of 0.0142.

SVM with a RBF kernel Finally, we implement SVM with a gaussian radial kernel. 10-fold cross validation is used to determine the best combination of the kernel parameter γ and the margin parameter C.

4.3.6 Conclusion

The analysis conducted above, using all available features, confirms that "meanfun" plays indeed an integral role in gender recognition by voice. Variable importance metrics of random forests highlight that "meanfun" is substantially more important than the rest of the features. Simultaneously, the results

[3] `https://cran.r-project.org/src/contrib/Archive/xgboost/`

obtained from (Lasso) logistic regression, pinpoint the fact that some predictors are not significant and, thus, not useful in improving predictive performance.

Nevertheless, we observe that using all features instead of only "meanfun", decreases the test errors of discussed models. Since we are also interested in specifying the model with the lowest classification error, it makes sense to take into account all 17 variables. However, as Table 4.2 indicates, model performance seems to be highly dependent on the initial training/test error split. Therefore, we cannot draw a conclusion about the best classifier and we have to refine our approach of model assessment.

4.4 Classification based on a 50/50 split of the dataset

4.4.1 Experimental settings

In order to decrease the variance, we suggest to use another strategy described hereafter. First we perform a 50/50 split of the dataset and we obtain two subsets of same dimension, denoted as $S1$ and $S2$. We use $S1$ to perform best parameter selection based on 10-fold cross validation. We use $S2$ for model comparison based on averaging of the classification error estimates on 5 folds, where the classification error estimate is computed as the $0-1$ loss. The advantage of such an approach is that the averaging induced by the 5 folds should reduce the variance, while keeping a relatively low bias because of the size of $S2$.

As in Section 4.3, the errors are computed for 5 seed numbers in order to study the variability of the models and the results are reported in Table 4.3.

Table 4.3 Classification Error Estimate of the Methods for Different Seed Numbers With 50/50 Split

Type	Methods	Seed number					Mean	Std.
		1	2	3	4	5		
Max. Lik.	Log. reg.	0.0221	0.0262	0.0310	0.0261	0.0291	0.0269	0.00339
	Log. reg. - ℓ_2	0.0196	0.0214	0.0305	0.0248	0.0278	0.0248	0.00449
	Log. reg. - ℓ_1	0.0284	0.0294	0.0368	0.0348	0.0364	0.0332	0.00397
	LDA	0.0291	0.0269	0.0342	0.0309	0.0313	0.0305	0.00274
	QDA	0.0280	0.0257	0.0370	0.0362	0.0365	0.0327	0.00542
Trees	Tree	0.0284	0.0317	0.0387	0.0454	0.0468	0.0382	0.00814
	Pruned Tree	0.0229	0.0296	0.0356	0.0454	0.0567	0.0355	0.0133
	Bagging	0.0209	0.0244	0.0319	0.0253	0.0302	0.0266	0.00445
	Random Forest	0.0191	0.0206	0.0305	0.0248	0.0290	0.0248	0.00500
	XGBoost	**0.0170**	**0.0149**	**0.0259**	**0.0197**	**0.0227**	**0.0201**	0.00439
SVM	Linear	0.0209	0.0269	0.0286	0.0229	0.0291	0.0249	0.00362
	Gaussian	0.0184	0.0188	0.0272	0.0211	0.0234	0.0218	0.00361
x	kNN	0.0273	0.0312	0.0326	0.0410	0.0382	0.0340	0.00551

4.4.2 Results and best classification method

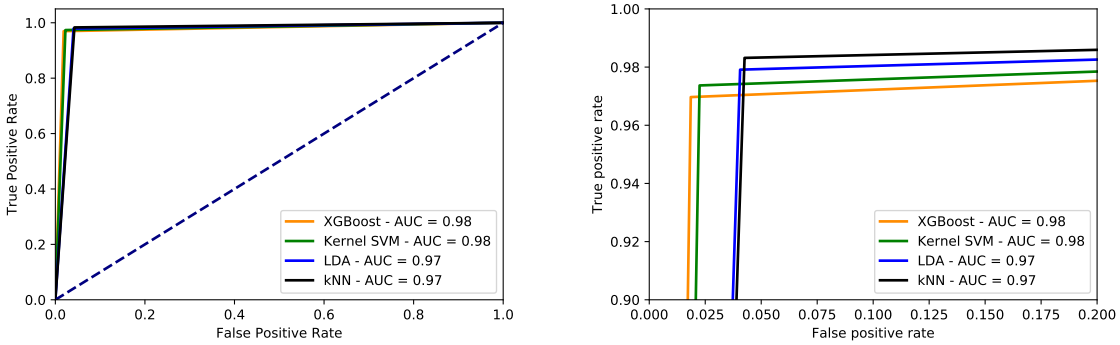
From Table 4.3, it can be noticed that the proposed strategy:

- gives very similar results to the 80/20 split regarding the classification error estimates of the different methods;

- tends to decrease the variance of the error estimate with respect to the seed number. Indeed, the average standard deviation of the error estimate over the different estimators is 0.00523 while it was 0.00636 with the 80/20 split.

Surprisingly, the variance reduction does not work for all the methods. For instance, QDA has a higher variance with the proposed strategy than with the 80/20 split. This can be explained by the fact that a 5-fold approach for test error estimation is not sufficient to significantly reduce the variance. However, we believe that, given the amount of data, increasing the number of folds may have a significant impact on the bias.

Nevertheless, the proposed strategy allows us to identify that XGBoost slightly outperforms other tree-based methods as well as the SVM with the RBF kernel. It is therefore chosen as the best classification model for our gender classification task.



(a) (b)

Figure 4.6 (a)-ROC curves for XGBoost (orange), Kernel SVM (green), LDA (blue) and kNN (black); (b)-ROC curves in a region around the transition.

The receiver operating characteristics (ROC) curves of XGBoost, Kernel SVM, LDA and kNN ,displayed on Fig. 4.6a, highlight the fact that the different classification methods are close. A finer analysis at the transition, displayed on Fig. 4.6b, shows that the superiority of XGBoost does not come from a higher true positive rate than the other methods, but from its ability to reach high true positive rates for lower false positive rates than the other methods.

5 Application of the Best Classifier on Acquired Voice Recordings

One main disadvantage of the dataset provided for the study is that there is no information about the way the voices have been recorded. This prevents us from further characterization such as the robustness of the classification against noise.

To study this aspect, we propose to test the best classifier on a small dataset of voice recordings that we have acquired. The dataset is made of 11 voice recordings, 3 from Hippolyte, 3 from Adrien, 2 from Hippolyte's girlfriend and 2 from Adrien's girlfriend. 3 recordings, *e.g.* the two from Adrien's girlfriend and 1 from Hippolyte have been acquired in a noisy environment, *i.e.* in a crowded place of EPFL.

A ".csv" file containing the features is generated from the recordings using available R scripts[1].

Regarding the classifier, XGBoost, which has been identified as the best classifier in Section 4.4 is fitted on the whole dataset of 3168 voices. Then it is used to classify the acquired voices, based on the features extracted from the ".csv" file.

The classification error on the set of acquired voice is 0 for XGBoost, which means that it does not do any mistake, even in the noisy cases.

As an indication, we also test the kNN classifier using the same process as before, and we obtain the same error as XGBoost.

Thus, it seems that the proposed classification method is robust to small amount of noise. However, further experiments need to be performed to validate the preliminary results.

[1] https://github.com/primaryobjects/voice-gender/blob/master/sound.R

6 Conclusion

In this work, we are interested in the problem of automatic gender recognition from voice. Given a dataset of spectral features extracted from 3168 labelled voice recordings, the objective is to design the most efficient gender classifier.

We compare an extensive number of state-of-the-art methods, *i.e.* logistic regression, regularized logistic regression, linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbors, decision tree, random forests, bagging, gradient boosting trees (XGBoost), linear and kernel-based (Gaussian) support vector machines. To perform the comparison, we achieve a 50/50 split of our dataset. The first subset is used for best parameter selection and the second subset for model comparison based on a 5-fold averaging of the test error to reduce the variance. We show that all the methods perform very well, with an average classification error of 2.98 % and demonstrate that XGBoost gives the lowest classification error, with an average of 2.01 %. We then perform a preliminary test of its robustness against small amount of noise, based on our own dataset of 11 voice recordings, acquired in a crowded place of EPFL. We show that XGBoost is robust to small amount of noise since it classifies well all the voices.

In addition to this comparison, we perform an in-depth study of the features involved in the classification since we believe that it is interesting to see which spectral information is most relevant for gender recognition. It appears that the most important feature is by far the mean fundamental frequency, which translates physiological differences between males and females, *e.g.* glottal shape and size of the vocal chords. Indeed, with this feature alone, the average classification error of the different models is already 5.05 %. Apart from it, statistical spectral features such as the first and third frequency quartile, the minimum and maximum fundamental frequencies are also important. More interestingly, the flatness of the spectrum seems to import more than the asymmetry or the kurtosis.

Thus, we manage to both identify the best classifier by establishing a rigorous comparison process and to understand the high importance of some features by performing an in-depth data analysis. In order to provide a really efficient method for real-word applications, the next step is to perform an in-depth study of the robustness of the classifier against many sources of noise such as whispering, background noise, and quantization.

Bibliography

[1] K. Wu and D. G. Childers, "Gender recognition from speech. Part I: Coarse analysis", *J. Acoust. Soc. Am.*, vol. 90, pp. 1828–1840, 1991.

[2] D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis", *J. Acoust. Soc. Am.*, vol. 90, pp. 1841–1856, 1991.

[3] D. Childers, Ke Wu, K. Bae, and D. Hicks, "Automatic recognition of gender by voice", in *1988 IEEE Int. Conf. Acoust. Speech Signal Process.*, 1988, pp. 603–606.

[4] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications", *J. Intell. Inf. Syst.*, vol. 24, pp. 179–198, 2005.

[5] Y.-m. Zeng, Z.-y. Wu, T. Falk, and W.-y. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech", in *2006 Int. Conf. Mach. Learn. Cybern.*, 2006, pp. 3376–3379.

[6] V. N. Sorokin and I. S. Makarov, "Gender recognition from vocal source", *Acoust. Phys.*, vol. 54, pp. 571–578, 2008.

[7] Fl. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, *et al.*, "Comparison of four approaches to age and gender recognition for telephone applications", in *2007 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, pp. 1089–1092.

[8] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines", in *2008 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 1605–1608.

[9] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks", in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, 2015, pp. 34–42.

[10] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications", *IEEE Trans. Emerg. Top. Comput.*, vol. 1, pp. 244–257, 2013.

[11] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information", in *Eur. Signal Process. Conf.*, 2004.

[12] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech", in *1996 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 1996, pp. 339–341.

[13] I. R. Titze, "Physiologic and acoustic differences between male and female voices", *J. Acoust. Soc. Am.*, vol. 85, pp. 1699–1707, 1989.

[14] H. Hollien and E. Malcik, "Evaluation of gross-sectional studies of adolescent voice change in males", *Speech Monogr.*, vol. 34, pp. 80–84, 1967.

[15] M. S. F. Poon and M. L. Ng, "The role of fundamental frequency and formants in voice gender identification", *Speech, Lang. Hear.*, vol. 18, pp. 161–165, 2015.

[16] M. Gupta, S. S. Bharti, and S. Agarwal, "Support vector machine based gender identification using voiced speech frames", in *2016 Fourth Int. Conf. Parallel, Distrib. Grid Comput.*, 2016, pp. 737–741.

[17] R. R. Nunes, M. P. de Almeida, and J. W. Sleigh, "Spectral entropy: A new method for anesthetic adequacy", *Rev. Bras. Anestesiol.*, vol. 54, 2004.

[18] H. Traunmüller, "The frequency range of the voice fundamental in the speech of male and female adults", Tech. Rep., 1994.

[19] Tr. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2001.