Forestier Quentin & Herzig Melvyn

MLG – 13.04.2022

Voice recognition

Labo 3



# Introduction

Ce laboratoire demande d’appliquer la méthodologie apprise durant les derniers travaux pratiques pour évaluer les performances d’un réseau de neurones entrainé.

Pour ce faire, il faut réaliser 4 expériences qui comparent des voix d’hommes, de femmes et d’enfants, naturelles ou synthétiques.

Au terme de chaque expérience, il faut pouvoir définir le model final à sélectionner pour traiter le problème associé.

La sélection du modèle consiste à trouver le modèle dont la complexité et la configuration sont appropriées.

## Contenu commun

Les caractéristiques suivantes, seront communes à chaque expérience :

* Une couche d’entrée de 13 neurones.
  + Pour chaque expérience, nous récupérons 13 valeurs MFCC par fenêtre.
  + Nous travaillions avec soit la moyenne, soit l’écart type.
* Une couche cachée de longueur variable que nous souhaitons optimiser.
  + [2, 4, 6, 8, 10, 15, 20, 25, 30]
* Une couche de sortie variable
  + Si deux classe, 1 neurone.
  + Sinon, autant de neurones que de classes.
* Les poids synaptiques sont attribués aléatoirement.
* La fonction d’activation choisie par défaut est la tangente hyperbolique
* Le « *threshold »* vaut 0.

# Expérience 1 - Hommes Vs Femmes avec une voix naturelle

Dans cet exemple, il s’agit d’entrainer le modèle à différencier des voix d’hommes naturelles à des voix de femmes naturelles.

Dans un premier temps, nous avons travaillé avec la moyenne des 13 valeurs MFCC.

### Nombres d’observations

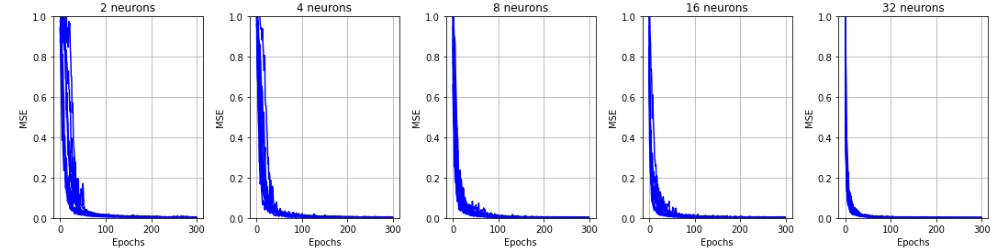
|  |  |  |
| --- | --- | --- |
| Total | Homme | Femme |
| 72 (100%) | 36 (50%) | 36 (50%) |
|  |  |  |

### Choix du nombre d’époques

Dans un premier temps, nous avons estimé combien d’époques étaient nécessaires pour trouver un faible taux d’erreur d’entrainement.

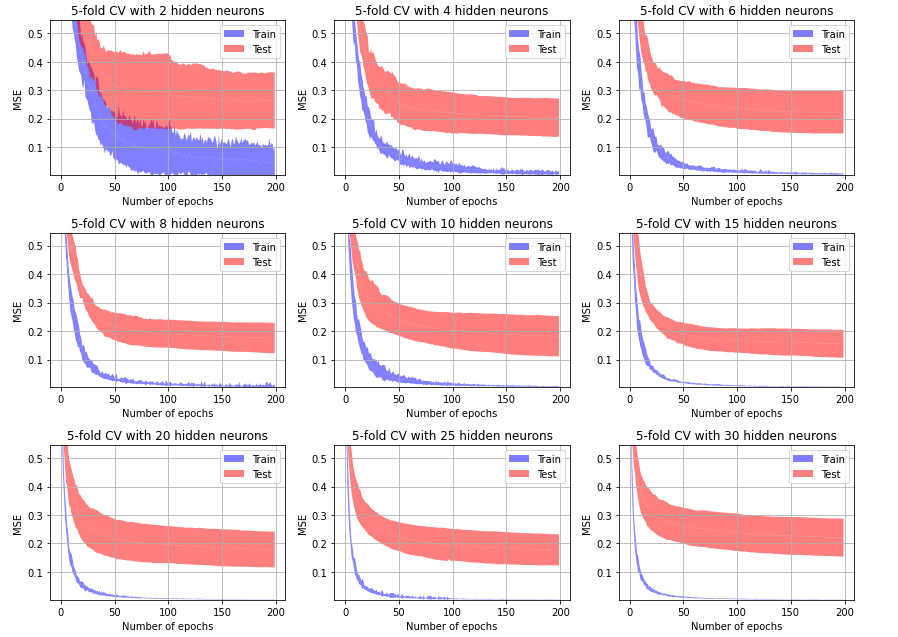
**Valeurs initiales**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N\_INITS** | **EPOCHS** | **N\_NEURONS** | **LEARNING\_RATE** | **MOMENTUM** |
| 10 | 300 | 2,4,8,16,32 | 0.001 | 0.5 |



Dans chaque configuration de 2 à 32 neurones dans la couche cachée, nous constatons que la courbe d’erreur se stabilise après 150 époques. En conséquence, nous allons travailler avec 200 itérations pour avoir une marge d’erreur. Les courbes n’étant pas extravagantes, nous avons conservé les mêmes valeurs pour le « *learning rate »* et le « *momentum »*.

### Choix du nombre de neurones dans la couche cachée



Grâce à un système de cross-validation, nous remarquons que le meilleur taux d’erreur de test est optimal avec 15 neurones, avec moins ou plus, le taux est légèrement moins bon.

### Performances finales

Avec un réseau de 13 neurones en entrée, 15 neurones dans la couche cachée et 1 neurone en sortie. Nous obtenons les performances finales suivantes :

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | Prédiction | | | | Vérité |  | H | F | | H | 35 | 1 | | F | 2 | 34 | | |  |  | | --- | --- | | MSE training: 0.00329 | MSE test: 0.17623 | | Precision: 0.945 (H) 0.971 (F) | Recall: 0.972 (H) 0.944 (F) | | F-Score: 0.957 (H) 0.957 (F) | | |

Grâce aux F-scores obtenus (>0.95), nous validons la configuration retenue. En conclusion, nous pouvons dire qu’il n’est pas nécessaire d’utiliser l’écart type. De plus, il semblerait que les voix d’hommes et de femmes peuvent être confondues.