Forestier Quentin & Herzig Melvyn

MLG – 18.04.2022

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 classes, 1 neurone.
  + Sinon, autant de neurones que de classes.
* Les poids synaptiques sont attribués aléatoirement.
* La fonction d’activation choisie est la tangente hyperbolique

# 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

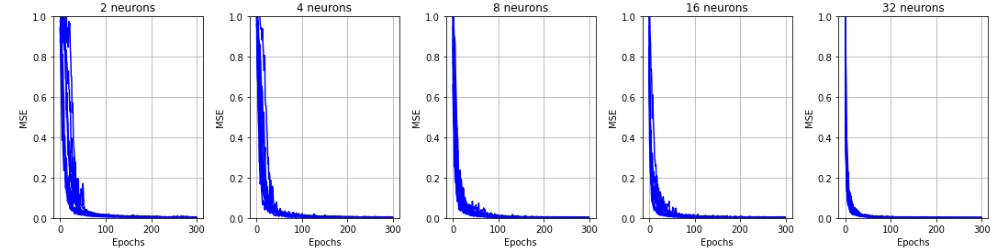
|  |  |  |
| --- | --- | --- |
| Total | Homme | Femme |
| 72 (100%) | 36 (50%) | 36 (50%) |
|  |  |  |

### Choix du nombre d’époques

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

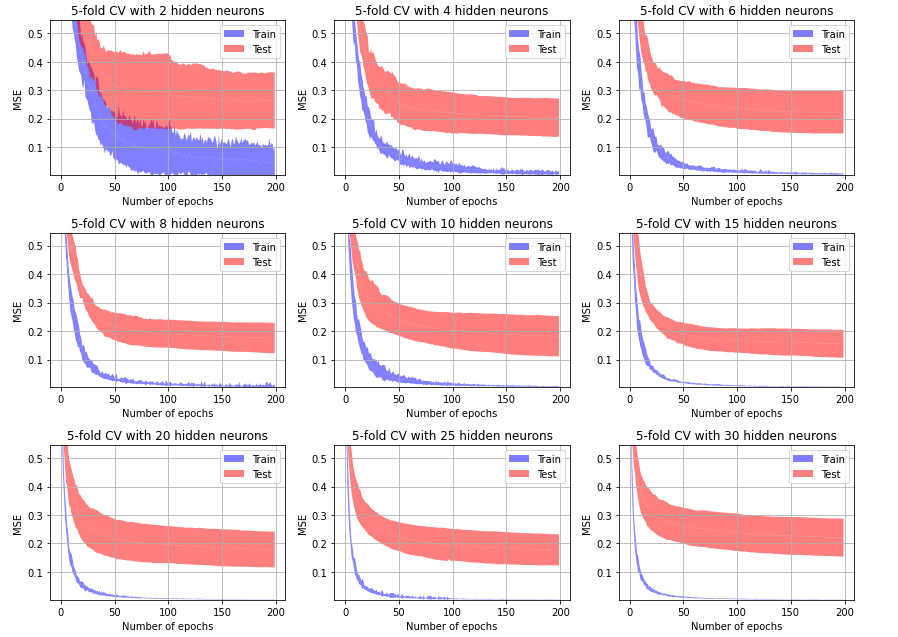
**Valeurs initiales**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 et plus épars.

### Performances finales

Avec un réseau de 13 neurones en entrée, 15 neurones dans la couche cachée et 1 neurone en sortie. Avec un threshold à 0, 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. En général, les voix d’hommes sont bien différentes des voix de femmes.

# Expérience 2 - Hommes Vs Femmes avec une voix naturelle et synthétique

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

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

### Nombres d’observations

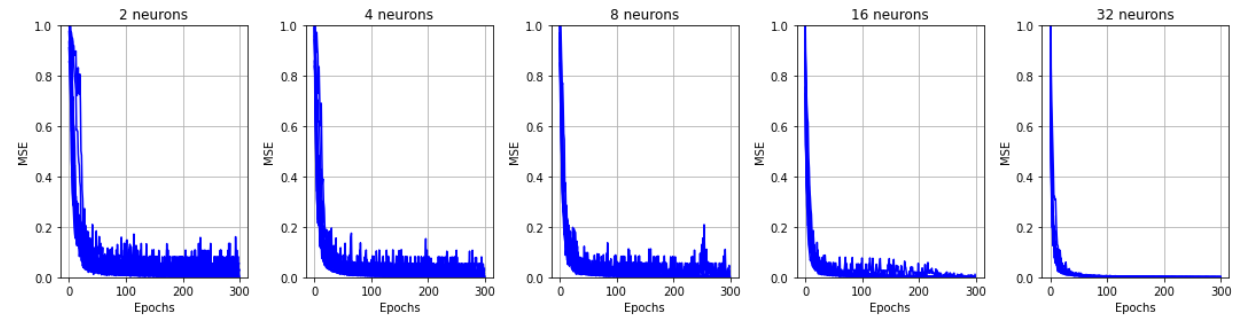
|  |  |  |
| --- | --- | --- |
| Total | Homme | Femme |
| 144 (100%) | 72 (50%) | 72 (50%) |
|  |  |  |

### 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 | 0.0008 | 0.5 |



Pour les configurations entre 2 à 8 neurones, nous constatons que la courbe a tendance à diminuer, mais elle se stabilise aux alentours de 200 époques. Cependant, pour la configuration à 16 neurones, la courbe se stabilise vers 230 époques. Pour garder un peu de marge, nous allons travailler avec 250 époques. Nous avons également décidé de réduire légèrement le learning rate, afin que les courbes soient plus lisses.

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

Il était très compliqué de voir sur le graphe quelle configuration était la meilleure. Mais l’image ci-dessus nous montre qu’il y a un meilleur taux de test sur la configuration à 8 neurones.

### Performances finales

Avec un réseau de 13 neurones en entrée, 8 neurones dans la couche cachée et 1 neurone en sortie. Avec un threshold à 0, nous obtenons les performances finales suivantes :

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | Prédiction | | | | Vérité |  | H | F | | H | 69 | 3 | | F | 3 | 69 | | |  |  | | --- | --- | | MSE training: 0.00256 | MSE test: 0.13773 | | Precision: 0.958 (H) 0.958 (F) | Recall: 0.958 (H) 0.958 (F) | | F-Score: 0.958 (H) 0.958 (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. Il est maintenant possible de dire que l’ajout des voix synthétiques ne change pas les résultats de l’expérience précédente.

# Expérience 3 - Hommes Vs Femmes Vs Enfants avec une voix naturelle

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

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

### Nombres d’observations

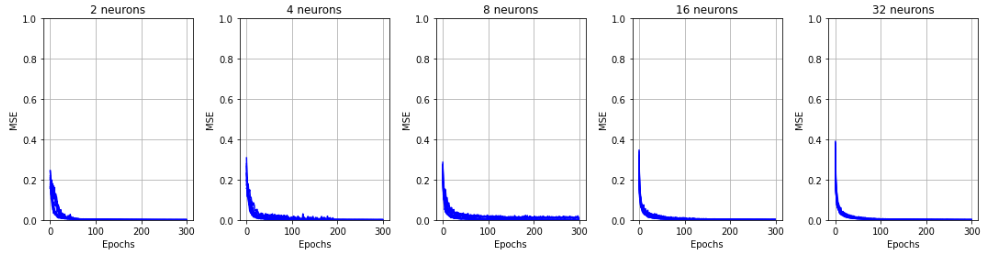
|  |  |  |  |
| --- | --- | --- | --- |
| Total | Homme | Femme | Enfant |
| 180 (100%) | 36 (20%) | 36 (20%) | 108 (60%) |
|  |  |  |  |

### Choix du nombre d’époques

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

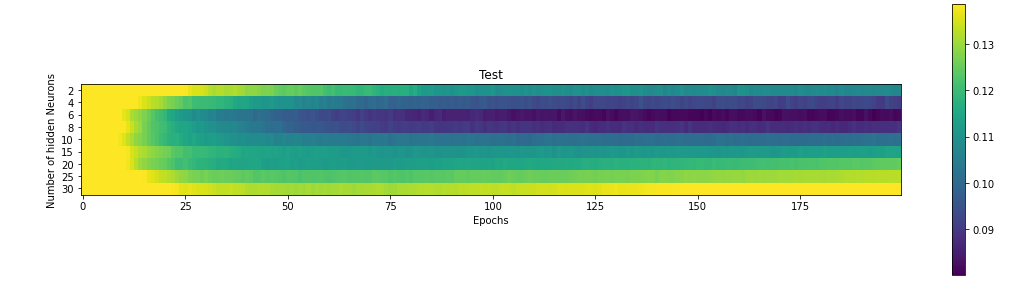
**Valeurs initiales**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 200 époques. 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 6 neurones. Dans les autres situations, le taux d’erreur de test est plus élevé et plus épars. De plus, avec 10+ neurones, nous remarquons de l’overfitting après 100 époques : plus les époques avancent plus le taux d’erreur de test se dégrade.

### Performances finales

Avec un réseau de 13 neurones en entrée, 6 neurones dans la couche cachée et 3 neurones en sortie. Avec un threshold de 0.55, nous obtenons les performances finales suivantes :

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Prédiction | | | | | | Vérité |  | H | F | E | | H | 34 | 2 | 2 | | F | 1 | 23 | 13 | | E | 1 | 14 | 94 | | |  |  | | --- | --- | | MSE training: 0.0271 | MSE test: 0.0933 | | Precision: 0.94(H) 0.59 (F) 0.86 (E) | Recall: 0.89 (H) 0.67 (F) 0.86 (E) | | F-Score: 0.91 (H) 0.62 (F) 0.86 (E) | | |

À ce stade nous pouvons faire deux remarques :

* Il y a plus d’observations dans la matrice que le nombre réel d’observations. Ce problème est dû au threshold. Bien que nous l’ayons calibré le plus précisément, il aurait fallu modifier la fonction qui calcule la matrice pour ne considérer que la sortie maximale des trois neurones comme étant la réponse.
* Les hommes se différencient facilement des femmes et des enfants. Toutefois, il semble que la barrière entre les femmes et les enfants soit plus floue. Ce qui donne des résultats légèrement moins bons. (1/3 femme = enfants et 1/10 enfants = femmes)

# Expérience 4 – Voix naturelle VS synthétique pour hommes, femmes et enfants

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

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

### Nombres d’observations

|  |  |  |
| --- | --- | --- |
| Total | Naturelle | Synthétique |
| 360 (100%) | 180 (50%) | 180 (50%) |
|  |  |  |

### Choix du nombre d’époques

Ensuite, 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 | 800 | 2,4,8,16,32 | 0.0003 | 0.6 |

Une image contenant intérieur

Description générée automatiquement

On voit une différence nette entre la configuration à 32 neurones et les autres. La courbe est devenue rapidement basse, et n’oscille que très peu. Même après 800 époques, la courbe a toujours tendance à baisser pour les configurations 2 à 8 neurones. Cependant, nous avons décidé de fixer une limite à 600 époques, qui nous a paru être un bon compromis. Nous avons également beaucoup baissé le learning rate, car la courbe avait tendance à partir dans tous les sens.

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

Une image contenant texte, shoji

Description générée automatiquement

On peut remarquer que le meilleur taux d’erreur est lors de la configuration a 30 neurones. Il atteint la valeur minimale, malgré son épaisseur.

### Performances finales

Avec un réseau de 13 neurones en entrée, 30 neurones dans la couche cachée et 1 neurone en sortie. Avec un threshold à 0, nous obtenons les performances finales suivantes :

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | Prédiction | | | | Vérité |  | N | S | | N | 172 | 8 | | S | 3 | 177 | | |  |  | | --- | --- | | MSE training: 0.00132 | MSE test: 0.11461 | | Precision: 0.983 (N) 0.957 (S) | Recall: 0.955 (N) 0.983 (S) | | F-Score: 0.969 (N) 0.970 (S) | | |

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. On remarque donc qu’il est tout à fait possible de détecter une voix synthétique d’une voix naturelle.

# Conclusion