MASTER 1 IA
MR. ALEXIS AREKION

LES METHODES STATISTIQUES EN INTELLIGENCE ARTIFICIELLE

INTRODUCTION

Dans cet essai nous testerons l'efficacité de plusieurs méthodes statistiques comme base de classifieurs que nous testerons sur la base IRIS. Nous nous intéresserons plus particulièrement aux méthodes bayésiennes selon plusieurs hypothèses sur les données. L'objectif étant bien d'expérimenter différents classifieurs Bayésiens l'ensemble du projet est orienté dans ce sens.

La classification de la base IRIS

La classification de la base IRIS est un problème très ancien et connu. Si aujourd'hui il semble dépassé compte tenu des nombreuses avancés faites en IA depuis sa création. Il a l'avantage d'être extrêmement simple et semble a priori un bon choix pour une première expérimentation de méthodes de classification se basant sur les statistiques.

1. Présentation de la base IRIS

La base IRIS est une base de donnée faites par des biologistes qui recense 150 fleurs avec respectivement, la longueur de leurs sépales, leur largeur, la longueur de leurs pétales, leur largeur et enfin leur espèce. Le but des classifieurs qui en découlent est de déterminer à laquelle de ces 3 espèces appartient une fleur à partir des 4 caractéristiques précédemment cités.

Les limites de la base IRIS

La base IRIS étant basé sur des mesures faites sur des organismes vivants elle semble intéressante pour nous initier à l'utilisation des méthodes statistiques que nous découvrons. Cependant elle présente 2 gros problèmes pour réellement tester l'efficacité d'un classifieur :

- Sa simplicité en effet le problème qu'elle pose est complètement dépassé à l'heure actuelle ainsi être performant sur ce problème ne veux pas dire qu'on est performant en général.
- La taille de sa base de donné en effet 150 est un nombre d'exemple ridiculement faible pour entraîner une ia et encore plus dans le cas de l'utilisation statistiques classiques.

Cependant malgré tout cela nous partirons de ce cadre : On découvre la base IRIS et venons de découvrir les méthodes statistiques ainsi nous allons à la fois prospecter sur la base IRIS mais également prendre les précautions nécessaires à l'expérimentation de nouvelles méthodes.

Faisabilité du problème

Étant donné que nous voulons expérimenter de nouvelles méthodes il paraît plus prudent de vérifier au préalable qu'il existe bien une solution au problème sur lequel nous allons l'expérimenter.

C'est pourquoi nous allons commencer dans un premier temps par tester la faisabilité du problème à l'aide d'un classifieur que nous connaissons déjà à savoir le ppv (plus proche voisin).

1. Le PPV

L'algorithme du plus proche voisin se base sur la distance entre les différents éléments à classifier, ainsi a chaque fois qu'il doit classer un élément A il vas chercher dans sa base d'apprentissage l'élément B le plus proche et classera A dans la même classe que B.

2. Mise en place du PPV

Afin de ne laisser aucune place au doute on testeras 1 000 fois le ppv sur la base IRIS avec à chaque fois la constitution d'une nouvelle base de test(⅓ de la base tiré aléatoirement) et d'une nouvelle base d'apprentissage (⅔ de la base tiré aléatoirement).

On en sortira une matrice contenant les pourcentages de réussite en généralisation de chacun des 1 000 lancers qu'on utilisera pour avoir la moyenne du pourcentage de réussite / la variance / le minimum ainsi que le maximum avec un graphique contenant l'ensemble des pourcentage de réussite et bien sur un histogramme de ceux ci.

On en tireras également une matrice de confusion "combiné" des 1 000 tirages, en somme une matrice de confusion sur les 1 000 tirages.

3. Résultats du PPV

Sur les 1 000 lancers ont à un pourcentage de réussite qui :

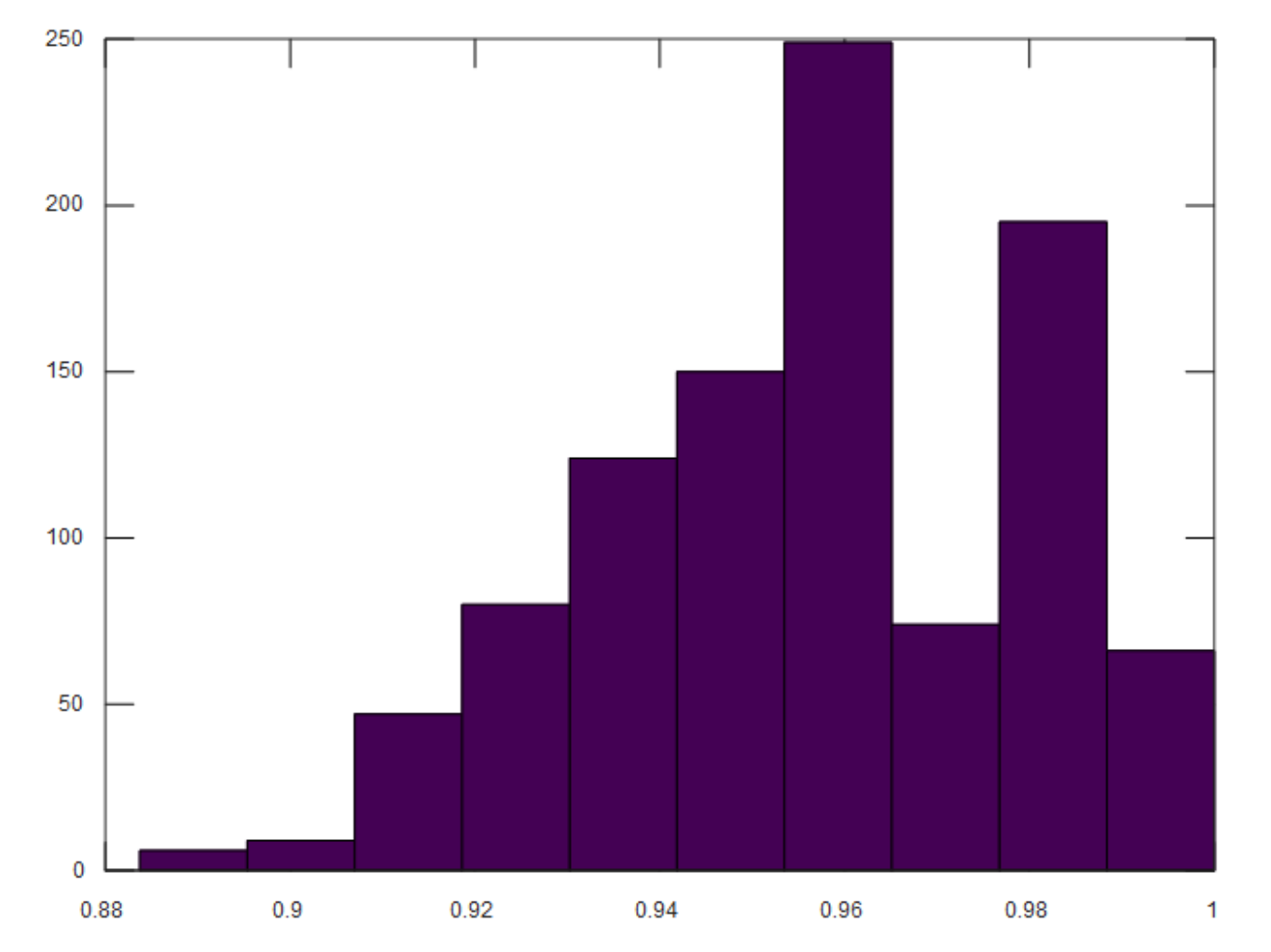
- En Moyenne est égal à `0.956772`ce qui est un bon résultat cela signifie donc qu'il existe des solution au problème. Ce qui implique que même sur un très mauvais tirage le ppv réussi dans presque 9 cas sur 10
- Avec une variance de `0.023182` ce qui implique des résultats relativement stables.
- Un minimum de `0.883721` ce qui implique que même sur un très mauvais tirage le ppv réussi dans presque 9 cas sur 10
- Un maximum de `1.000000` ce qui implique que sur un bon tirage le ppv est capable de classifier parfaitement les fleurs.

La Matrice de confusion "combiné"

	Classe 0 ppv	Classe 1 ppv	Classe 2 ppv
Classe 0 v	16487	0	0
Classe 1 v	0	15658	974
Classe 2 v	0	1171	15413

On remarque que le ppv ne se trompe jamais dans la classification de la classe 0 qu'on peut supposer plus facile à identifier que les autres classes

L’Histogramme du pourcentage de réussite des 1 000 lancers :



On remarque que la plupart des tirages donnent un pourcentage de réussite compris entre 0.93 et 1 et que le ppv ne descend que rarement sous les 90% de réussite

Certains tirages donnant des résultats quelque peux hasardeux on pourrait utiliser certaines techniques permettant d’optimiser au maximum notre base IRIS. Notamment le leave one out (technique consistant a prendre 1 seul exemple pour notre base de test a chaque fois, ceci afin de maximiser la taille de notre base d’apprentissage qui est potentiellement trop petite)

En conclusion le problème est faisable.

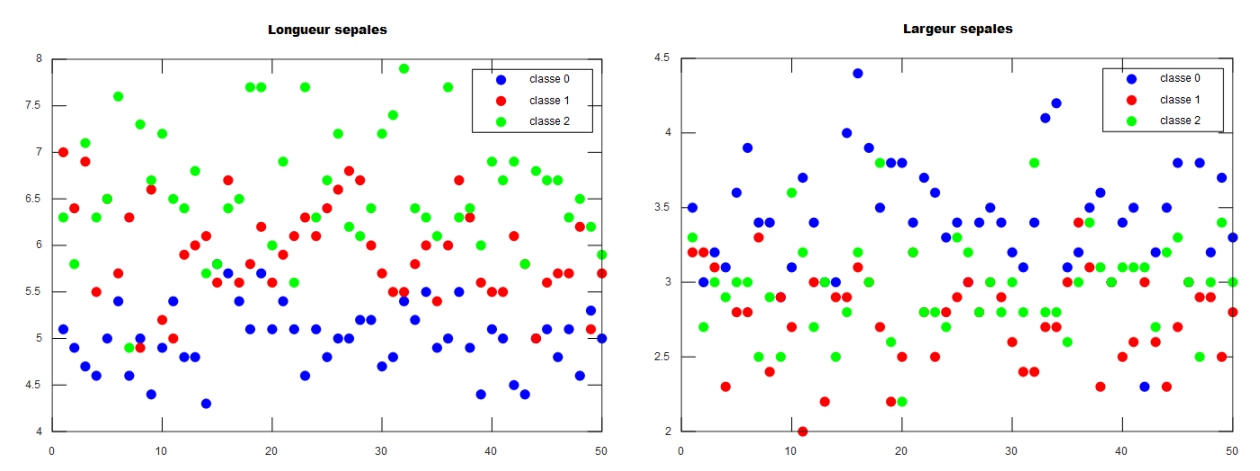
Analyse des donnés

Dans cette partie on cherche à savoir de quelle manière sont réparties nos donnés afin d'ajuster notre approche statistique par la suite.

1. Analyse des 4 caractéristiques indépendamment

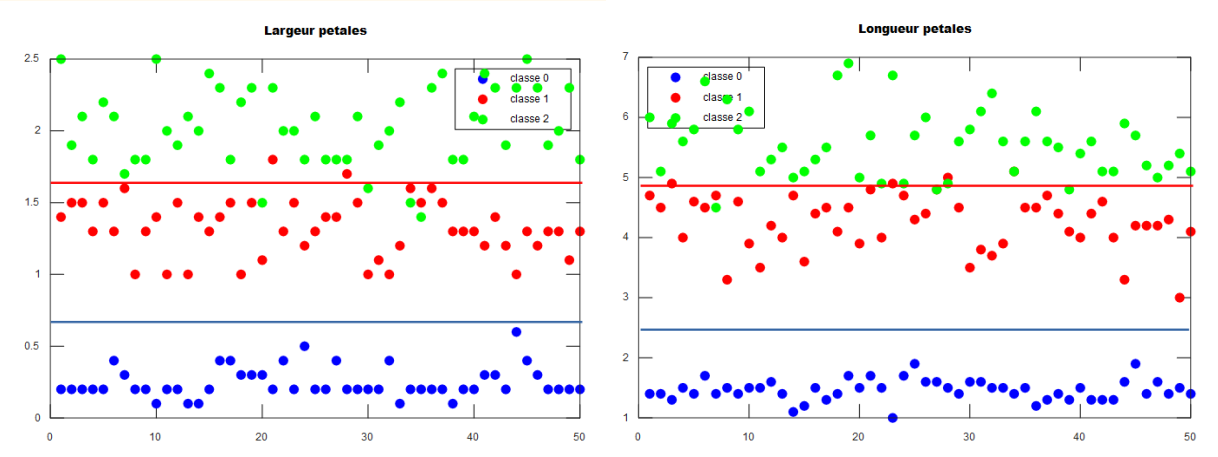
Dans un premier temps on regarderas les 4 caractéristiques indépendamment :

Dans un premier temps les sepales :



Que ce soit leur longueur ou leur largeur l'utilisation seule des mesures des sépales semble impossible du fait des classes qui ne semblent pas se distinguer les unes des autres.

Dans un second temps les pétales :



Ici contrairement aux sépales les mesures des pétales semblent bien distinctes pour la classe 0 et relativement distincte même si elle se confondent légèrement sur les classes 1 et 2 .

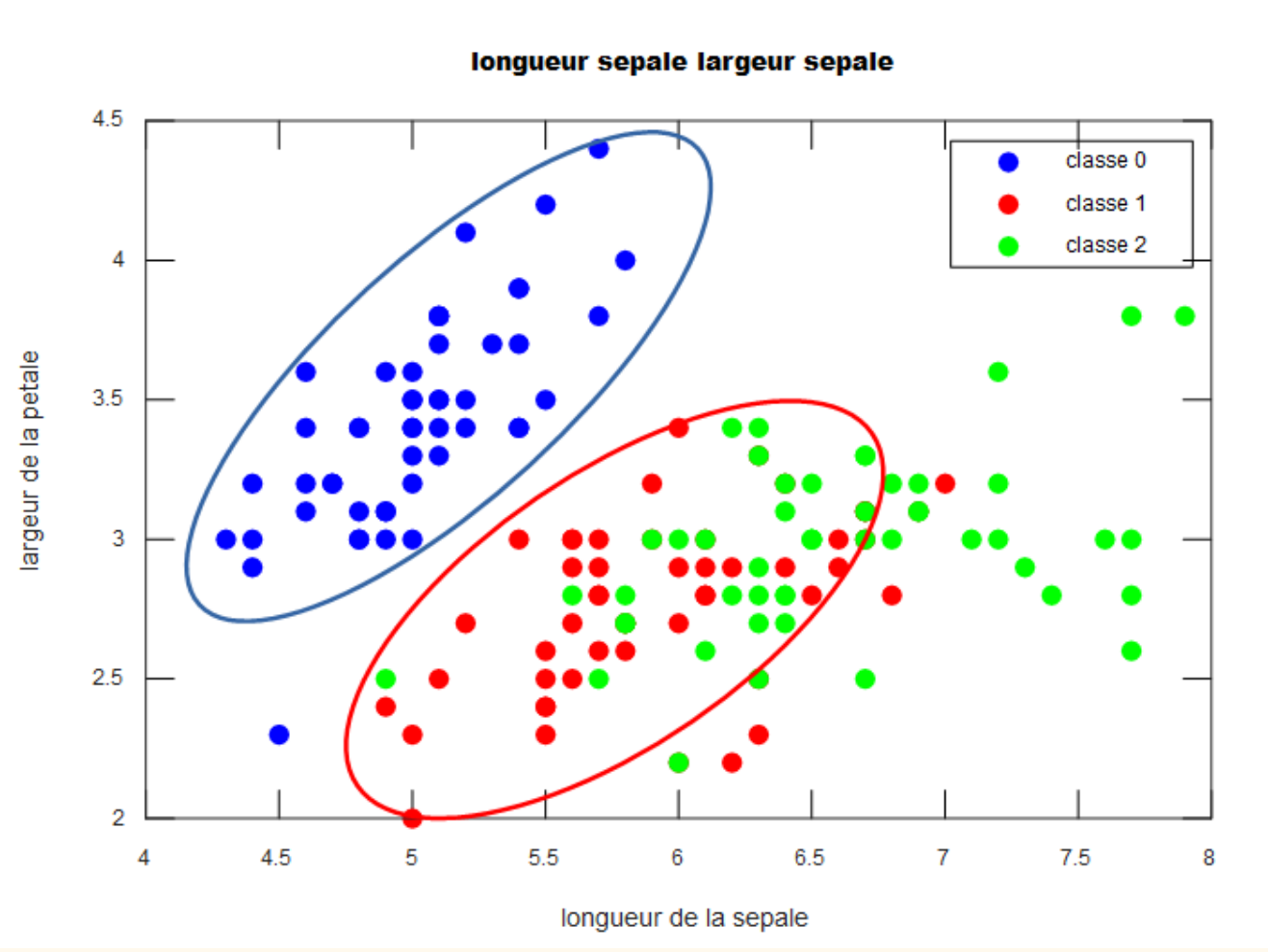
Si on ne voulait que distinguer que la classe 0 des classes 1 et 2 on pourrait penser à n'utiliser uniquement les mesures des pétales. Cependant pour classifier les 3 classes il ne semble pas très pertinent d'utiliser une seule des caractéristiques on peut cependant penser qu'utiliser une combinaison de plusieurs caractéristiques notamment la largeur et la longueur des pétales puisse être pertinent.

2. Analyse des rapports de dépendances entre les différentes caractéristiques

Dans cette seconde partie nous allons enquêter non plus sur les caractéristiques indépendamment mais sur les éventuels rapports de dépendances qu'elles pourraient avoir.

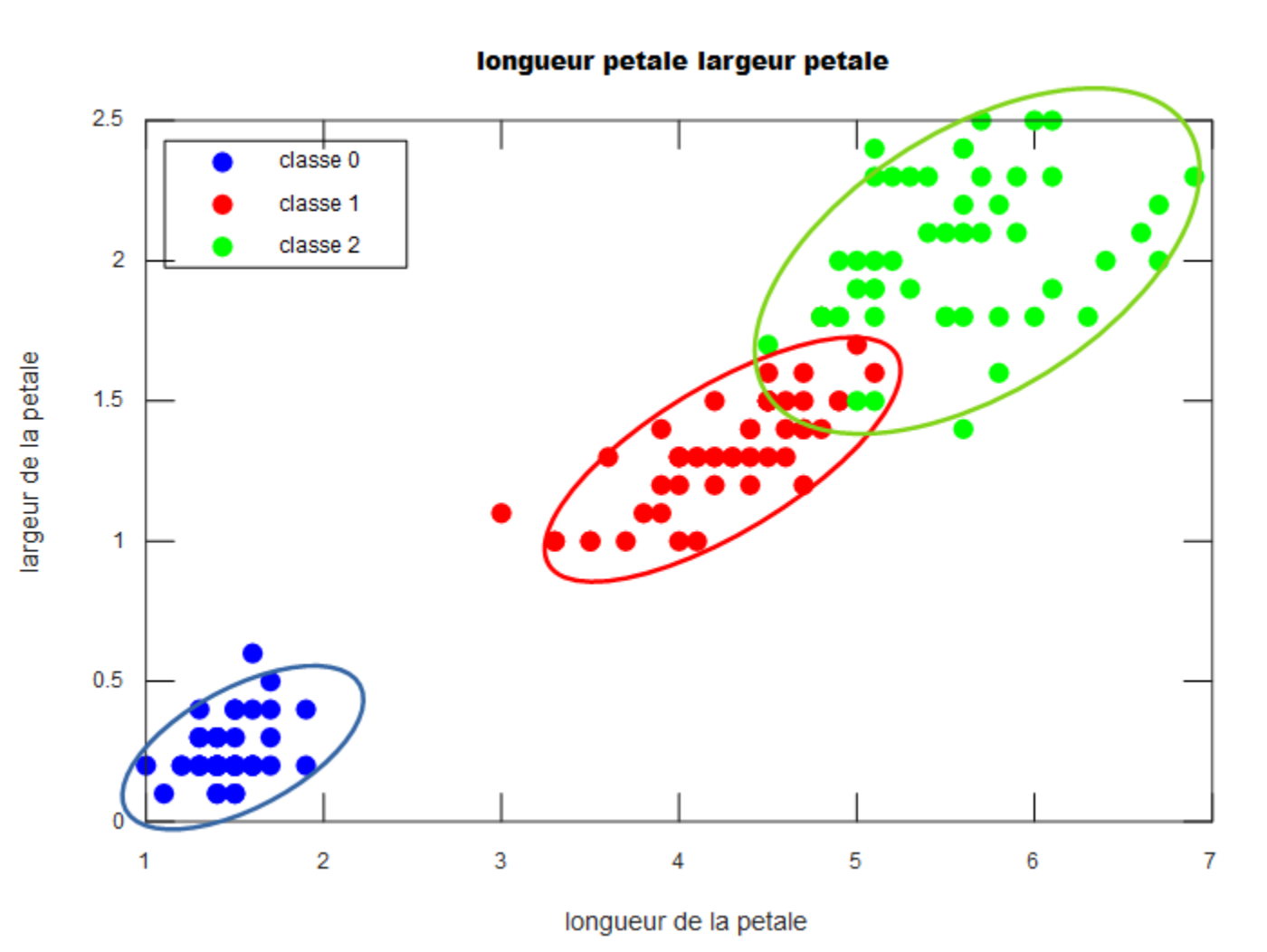
A prioris s'agissant de mesures de plantes dans la natures ont peut supposer qu'il existe une certaine "proportionnalité" entre les différentes caractéristiques ainsi il serait logique qu'une fleur plus large que ses consoeurs serait également plus longue par exemple.

Dans un premier temps on ne s'intéresse qu'au mesures des sépales :



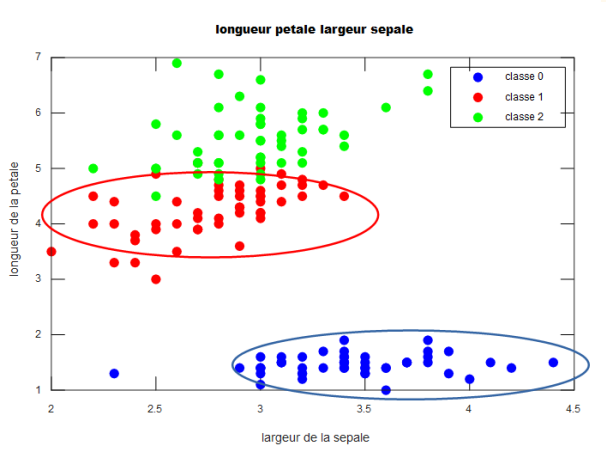
Bien que ce ne soit clair que pour la classe 0 il semble il y avoir un lien entre la longueur et la largeur des sépales, de plus contrairement à l'observation des largeurs puis longueurs des sépales seules l'observation de ces 2 caractéristiques en même temps semble séparer la classe 0 de ses consoeurs.

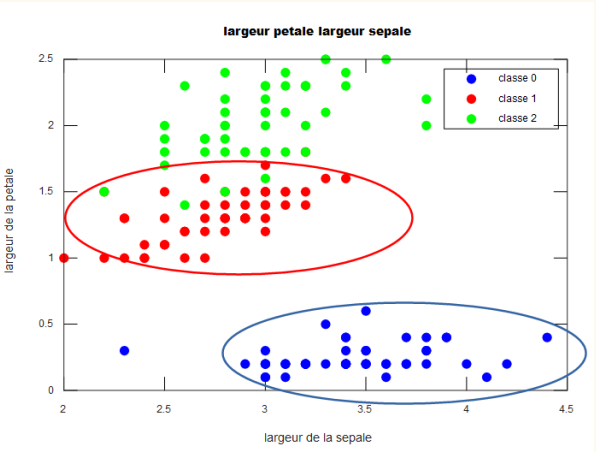
Observation des mesures des pétales :

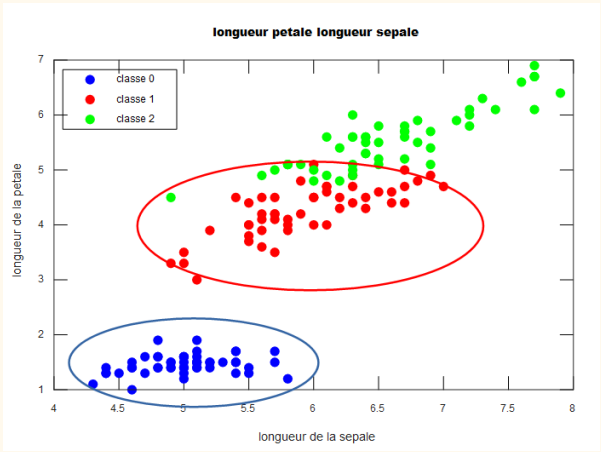


Ici bien plus que pour les sépales, la longueur et la largeur des pétales est lié de plus observer les 2 mesures des pétales semble être efficace pour distinguer les classes en elles (même si la frontière entre les classes 1 et 2 semble un peux flou).

Observation des autres couples de caractéristiques :







Comme on peut le voir les autres couples de caractéristiques ne semblent à prioris pas avoir de liens entre elles.

En conclusion les 4 caractéristiques ne sont pas indépendantes puisqu'il y a une relation entre la largeur et la longueur des sépales ainsi qu'entre la largeur et la longueur des pétales, ce qui confirme les suppositions qu'on a pu faire avant de réaliser ces observations.

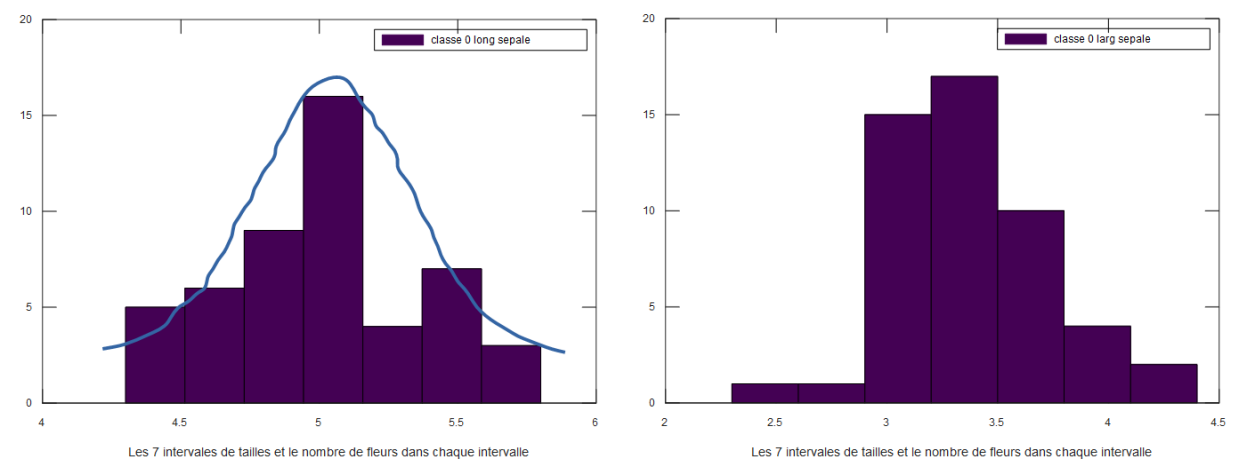
A noter : il existe des tests mathématiques permettant de vérifier rigoureusement l'indépendance de plusieurs paramètres on aurait pu les utiliser au lieu de faire ces observations cependant l'objectif étant non pas d'uniquement déterminer l'indépendance des mesures entre elles mais bien d'analyser les données de la base il était plus pertinent de procéder à ces observations.

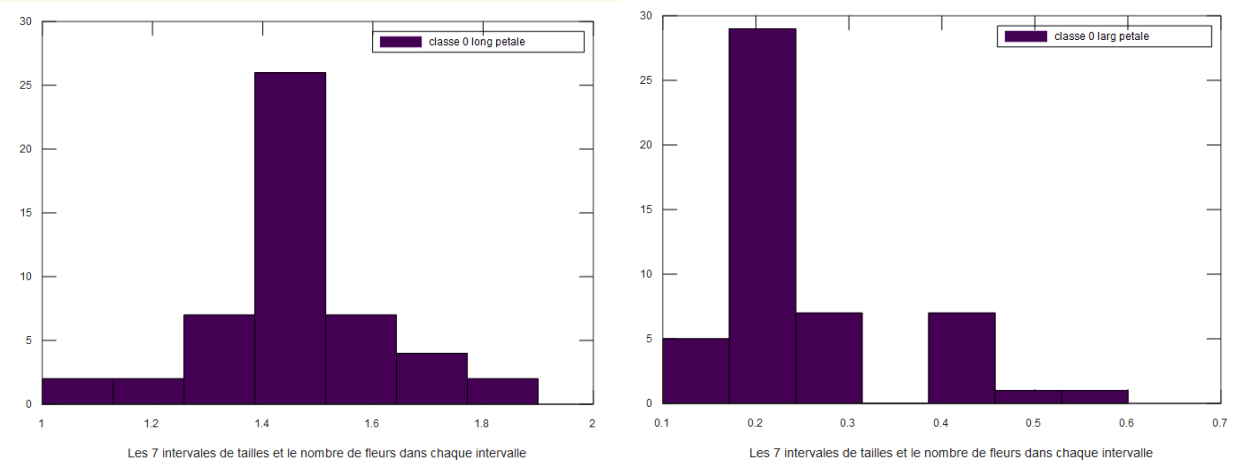
3. Analyse des histogrammes

Dans cette partie nous allons procéder à l'observation des histogrammes des différentes caractéristiques des différentes classes. Ces observations peuvent nous apporter beaucoup d'informations sur la répartition des densités de probabilités ce qui pourrait nous permettre d'utiliser certains modèles mathématiques.

S'agissant de mesures faites sur un organisme vivant dans la nature et donc soumis à la sélection naturelle, il semble probable de trouver des répartition de probabilité gaussiennes.

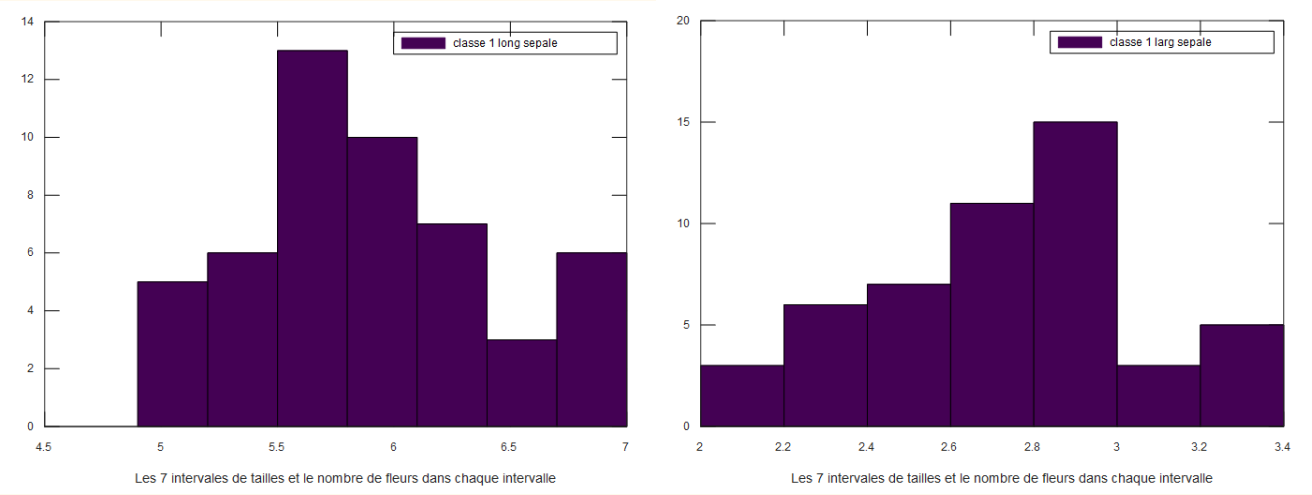
Observation des histogrammes des 4 caractéristiques individuellement pour la classe 0 :

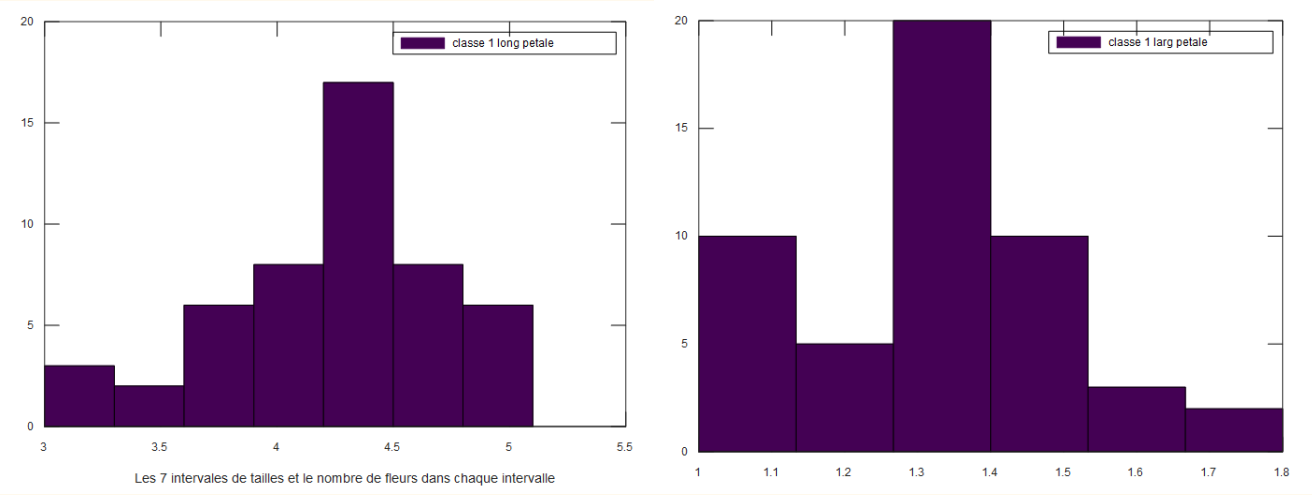




Ici particulièrement pour les mesures des pétales, la répartition des densités de probabilité de la classe 0 semble suivre une loi gaussienne.

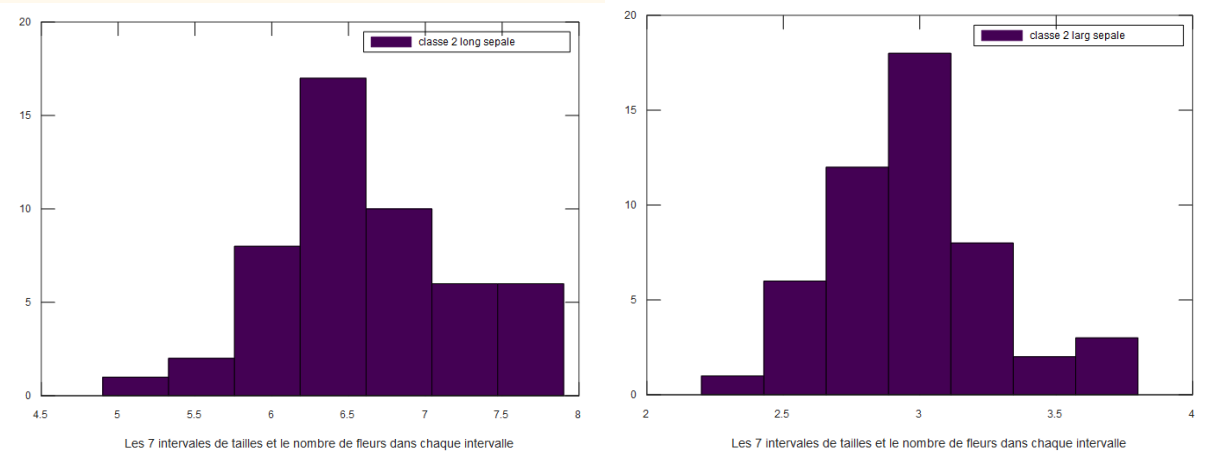
Observation des histogrammes des 4 caractéristiques individuellement pour la classe 1 :

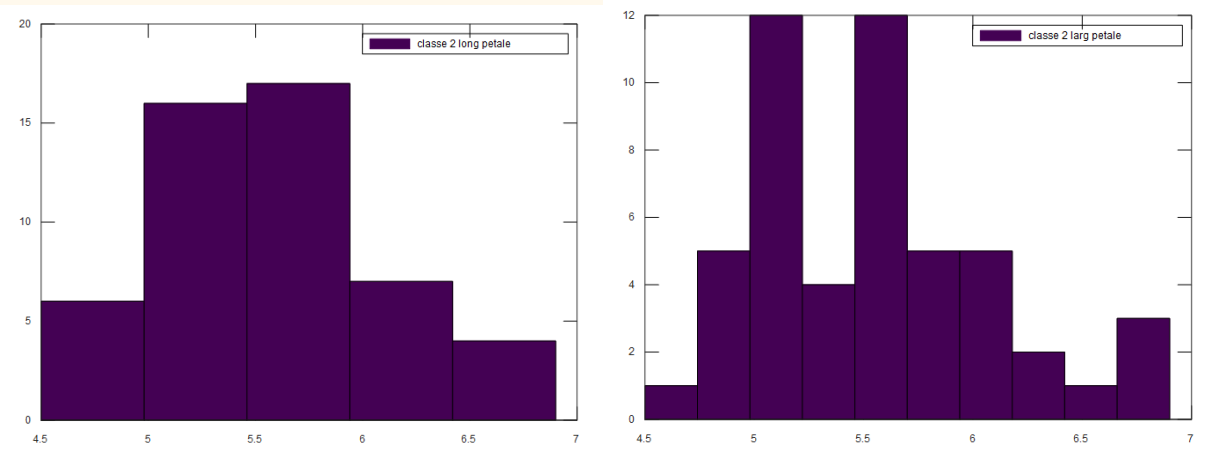




Ici comme pour la classe c0, la répartition des densités de probabilité de la classe 1 semble suivre une loi gaussienne, particulièrement sur les mesures des pétales.

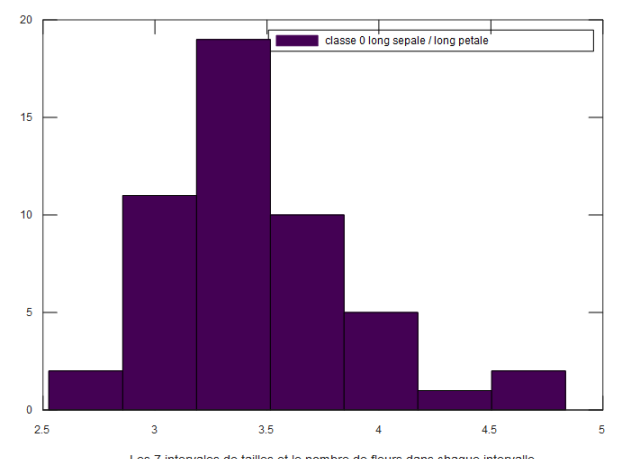
Observation des histogrammes des 4 caractéristiques individuellement pour la classe 2 :

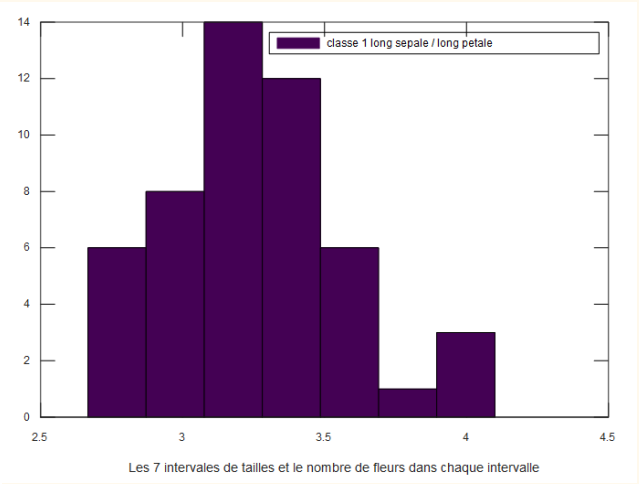


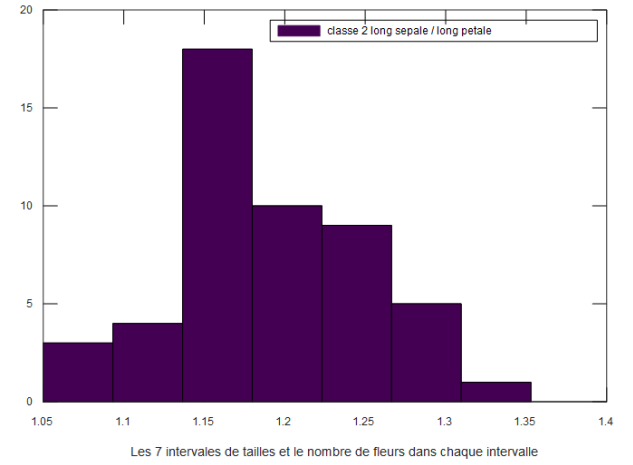


Ici comme pour les classe c0 et c1, la répartition des densités de probabilité de la classe 2 semble suivre une loi gaussienne, particulièrement sur les mesures des sépales cette fois ci.

Observation des histogrammes de la longueur des sépales / la longueur des pétales pour les 3 classes :







Comme on pouvait s'y attendre les histogrammes des 3 classes semblent suivre une loi gaussienne ici aussi.

Conclusion

Les classes ne semblent pas être mono-modales, il semble donc essentiel de ne pas se baser sur une seule des caractéristiques pour notre classifieur.

Les caractéristiques ne sont pas indépendantes il semble donc plus judicieux d'utiliser les méthodes ne se basant sur une indépendance stricte des caractéristiques entre elles.

La densité de probabilité des différentes caractéristiques semble suivre l'hypothèse gaussienne pour chaque classe, on peut donc vraisemblablement se baser sur celle ci.

Les méthodes de classification Bayésiennes

Dans un premier temps il convient de préciser que la base IRIS ne compte que très peux d'exemple ce qui est incompatible avec l'utilisation de statistiques "conventionnelles", c'est entre autre pour cela qu'on vas devoir utiliser la statistiques Bayésiennes.

1. Présentation de la statistiques Bayésiennes

La statistique bayésienne est une approche statistique fondée sur le Théorème de Bayes qui est le suivant : $$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Où P(A|B) est la probabilité conditionnelle de A sachant B. (Par exemple la probabilité qu'un homme ai des cheveux serais P(Cheveux | Homme))

Cette méthode présente 2 gros avantages qui répondent parfaitement à nos besoins :

-Elle demande beaucoup moins d'exemples que la statistique fréquentielle, ce qui colle parfaitement avec notre petite base d'exemple.

-Elle ne demande pas d'hypothèse de départ ce qui nous permet d'explorer librement nos donnés.

2. Les statistiques comme base d'un classifieur

Le fonctionnement de notre classifieur Bayésien est simple : le classifieur calcule pour chaque fleur la probabilité P(Cx|carac) qu'elle appartienne à chacune des 3 classes en fonction des caractéristiques de la fleur puis estimera que la fleur appartient à la classe dont la probabilité P(Cx|carac) est la plus haute.

Ainsi le but des des classifieurs Bayésiens que nous mettrons en place sera de déterminer ces probabilités P(Cx|carac), pour cela chacuns de ces classifieurs se basera sur les donnés de leurs bases d'apprentissage respectives ainsi que sur diverses lois de probabilités.

3. Pertinence et mise en place des différents classifieurs

A - Classifieur Bayésien, hypothèse de densités de probabilité gaussiennes, estimation des paramètres de moyenne et de variance, hypothèse d'indépendance des caractéristiques.

Ici bien que l'hypothèse de densités de probabilité gaussiennes semble bien respecté, les caractéristiques ne sont pas indépendantes. Fort heureusement il existe le classifieur Bayésien naïf qui nous permet de passer outre le non respect de cette hypothèse.

1. Mise en place du Bayésien naïf

Comme vus précédemment le but du classifieur est de calculer P(Cx|carac) la probabilité d'appartenance a chaque classe selon les caractéristiques des fleurs qu'il devras classer. Ce que nous donne cette méthode c'est P(carac|Cx) grâce à la fonction normpdf qui en découle.

A noter que P(carac|Cx) est en réalité le produit de P(carac1|Cx)* P(carac2|Cx) * P(carac3|Cx) * P(carac4|Cx) en effet le Bayésien naïf ne calculant que les probabilité P(caracy|Cx) d'avoir une caractéristique a la fois sachant la classe, on doit passer par ce petit calcul pour avoir la probabilité P(carac|Cx).

Hors selon le théorème de Bayes P(carac|Cx) = (P(Cx|carac) * P(carac)) / P(Cx), les classes étant réparties de manière égale (on a autant de fleurs dans chaque classe dans notre base), nous n'avons pas besoin de prendre en compte P(Cx)

De plus étant donné que l'on veut comparer a chaque fois P(C1|carac) P(C2|carac) et P(C3|carac) il n'est pas nécessaire non plus de prendre en compte le facteur P(carac) (qui équivaut à une constante dans le cadre de cette comparaison).

Ainsi dans notre cas P(Cx|carac) ⇔ P(carac|Cx) on peut donc utiliser P(carac|Cx) pour notre classifieur.

De ce fait pour classer une fleur on choisira la classe dont la probabilité P(carac|Cx) est la plus haute.

Pour tester notre classifieur on utilisera la même méthode que pour le ppv à savoir :

On testeras 1 000 fois le Bayésien naïf sur la base IRIS avec à chaque fois la constitution d'une nouvelle base de test(⅓ de la base tiré aléatoirement) et d'une nouvelle base d'apprentissage (⅔ de la base tiré aléatoirement).

On en sortira une matrice contenant les pourcentages de réussite en généralisation de chacun des 1 000 lancers qu'on utilisera pour avoir la moyenne du pourcentage de réussite / la variance / le minimum ainsi que le maximum avec un graphique contenant l'ensemble des pourcentage de réussite et bien sur un histogramme de ceux ci.

On en tirera également une matrice de confusion "combiné" des 1 000 tirages, en somme une matrice de confusion sur les 1 000 tirages.

2. Résultats du Bayésien naïf

Sur les 1 000 lancers ont à un pourcentage de réussite qui :

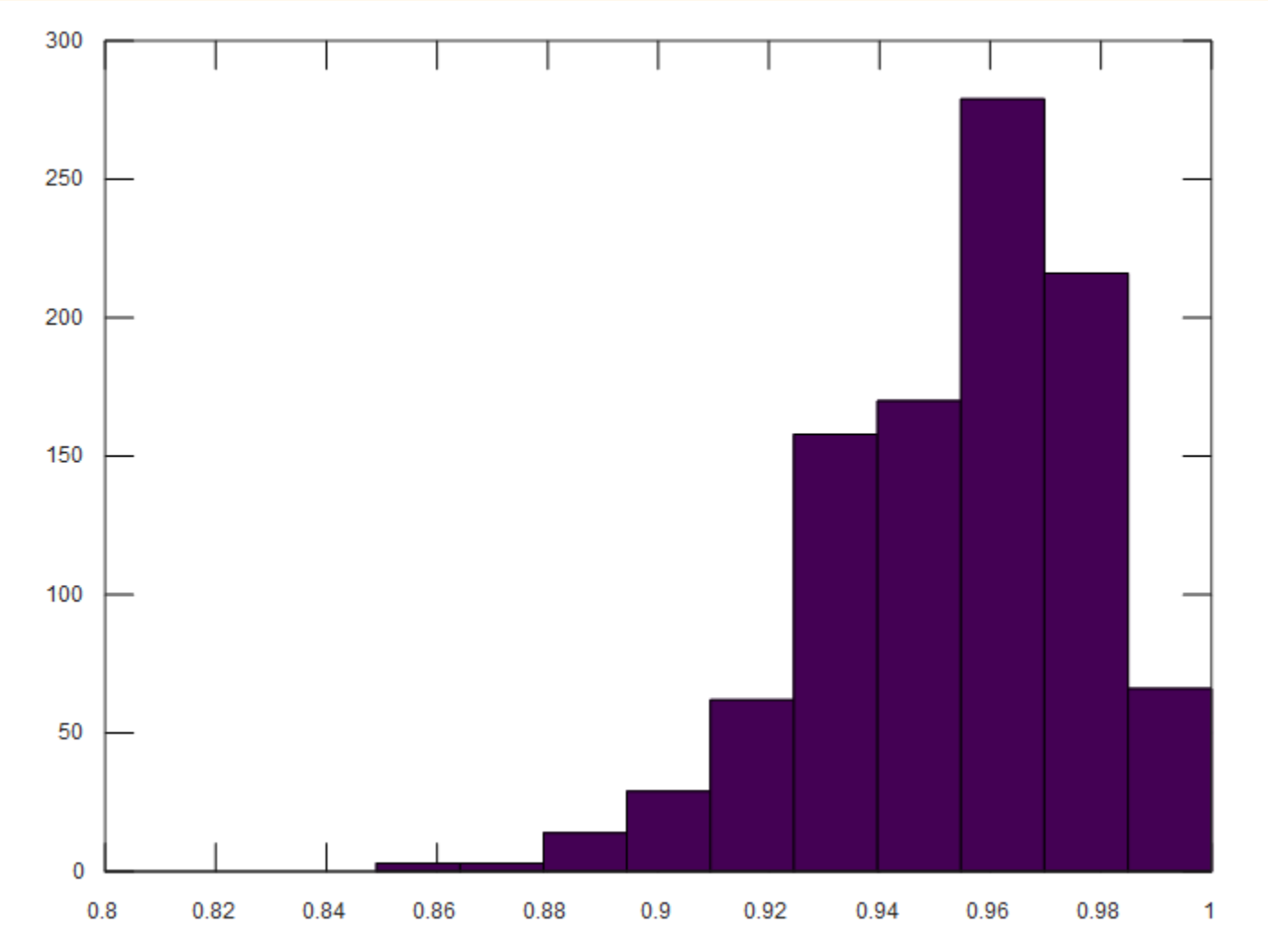
- En Moyenne est égal à `0.954538` ce qui est un bon résultat.
- Avec une variance de `0.025482` ce qui implique des résultats relativement stables.
- Un minimum de `0.849057`.
- Un maximum de `1.000000` ce qui implique que sur un bon tirage le Bayésien naïf est capable de classifier parfaitement les fleurs.

La Matrice de confusion "combiné"

	Classe 0 ppv	Classe 1 ppv	Classe 2 ppv
Classe 0 v	16488	0	0
Classe 1 v	0	15440	1037
Classe 2 v	0	1215	15478

On remarque que tout comme le ppv il ne se trompe jamais dans la classification de la classe 0 qu'on peut supposer plus facile à identifier que les autres classes.

L’Histogramme du pourcentage de réussite des 1 000 lancers :



On remarque que la plupart des tirages donnent un pourcentage de réussite compris entre 0.93 et 1 et que le Bayésien naïf ne descend que rarement sous les 90% de réussite.

En conclusion le Bayésien naïf est un bon classifieur pour la base IRIS.

B - Classifieur Bayésien, hypothèse de densités de probabilité d'une loi normale multivariée, estimation de la matrice de covariance.

Ici contrairement au cas du Bayésien naïf on prend bien en compte la non indépendance des caractéristiques que l'on a observé lors de l'analyse des données. On s'attend donc a avoir de meilleurs résultats.

1. Mise en place de la loi normale multivariée

Comme vus précédemment le but du classifieur est de calculer P(Cx|carac) la probabilité d'appartenance a chaque classe selon les caractéristiques des fleurs qu'il devras classer. Ce que nous donne cette méthode c'est P(carac|Cx) grâce à la fonction mvnpdf qui en découle.

Ici contrairement au Bayésien naïf la loi normale multi-varié nous donne directement P(carac|Cx).

Ainsi pour les mêmes raisons que le Bayésien naïf, pour classer une fleur on choisira la classe dont la probabilité P(carac|Cx) est la plus haute.

Pour tester notre classifieur on utilisera la même méthode à savoir :

On testeras 1 000 fois le Bayésien basé sur la loi normale multi-varié sur la base IRIS avec à chaque fois la constitution d'une nouvelle base de test(⅓ de la base tiré aléatoirement) et d'une nouvelle base d'apprentissage (⅔ de la base tiré aléatoirement).

On en sortira une matrice contenant les pourcentages de réussite en généralisation de chacun des 1 000 lancers qu'on utilisera pour avoir la moyenne du pourcentage de réussite / la variance / le minimum ainsi que le maximum avec un graphique contenant l'ensemble des pourcentage de réussite et bien sur un histogramme de ceux ci.

On en tirera également une matrice de confusion "combiné" des 1 000 tirages, en somme une matrice de confusion sur les 1 000 tirages.

2. Résultats de la loi normale multivariée

Sur les 1 000 lancers ont à un pourcentage de réussite qui :

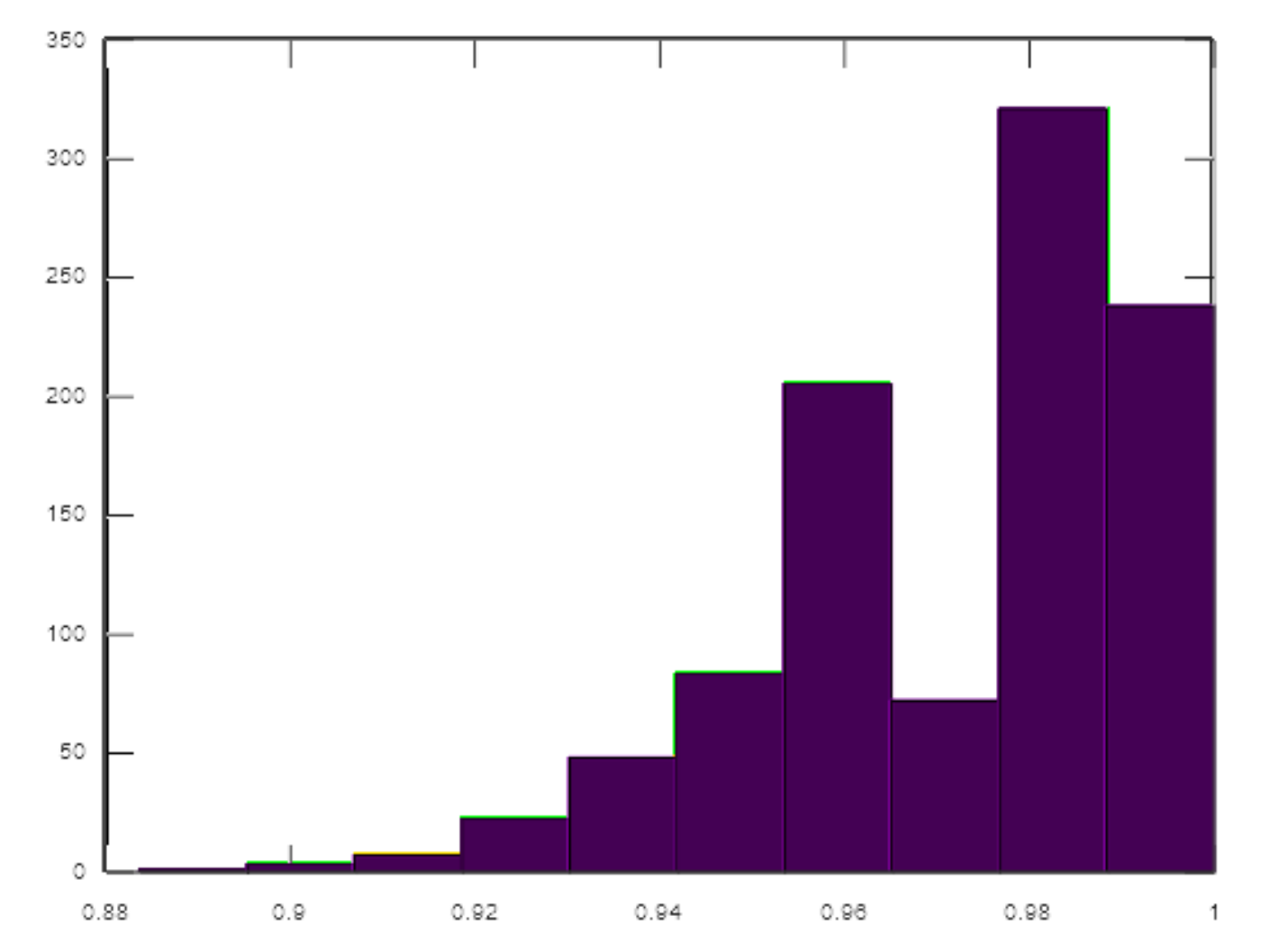
- En Moyenne est égal à `0.973454` ce qui est un meilleur résultat que le Bayésien naïf.
- Avec une variance de `0.021131` ce qui implique des résultats relativement stables.
- Un minimum de `0.883721.`
- Un maximum de `1.000000` ce qui implique que sur un bon tirage il est capable de classifier parfaitement les fleurs, c'est même plus qu'un bon tirage puisqu'il arrive à 100% de réussite dans 238 cas sur 1 000 soit quasiment 1 tirage sur 4.

La Matrice de confusion "combiné"

	Classe 0 ppv	Classe 1 ppv	Classe 2 ppv
Classe 0 v	16521	0	0
Classe 1 v	0	15703	1009
Classe 2 v	0	313	16213

On remarque une grande amélioration de sa précision sur le classement de la classe 2 comparativement au bayésien naïf et au ppv. On peut donc supposer que dans cadre du classement de la classe 2 considérer les caractéristiques comme non indépendantes soit un bon choix.

L’Histogramme du pourcentage de réussite des 1 000 lancers :



On remarque une amélioration globale des résultats puisque désormais la plupart des tirages donnent un pourcentage de réussite compris entre 0.96 et 1 et qu’on ne descend quasiment jamais sous les 90%.

En conclusion comme on pouvait s'y attendre le classifieur Bayésien basé sur la loi normale multivarié est encore meilleur que le Bayésien naïf ou le ppv.