

Exploration et classification de données



Eric Normandeau - 2015-01-27

Présentations du club

Format

- Similaire aux Interlabs IBIS
- Dimension analyses (analyses, programmes, code...)
- Exposer son expérience personnelle
- Pas simplement présenter des articles

Disponibilité

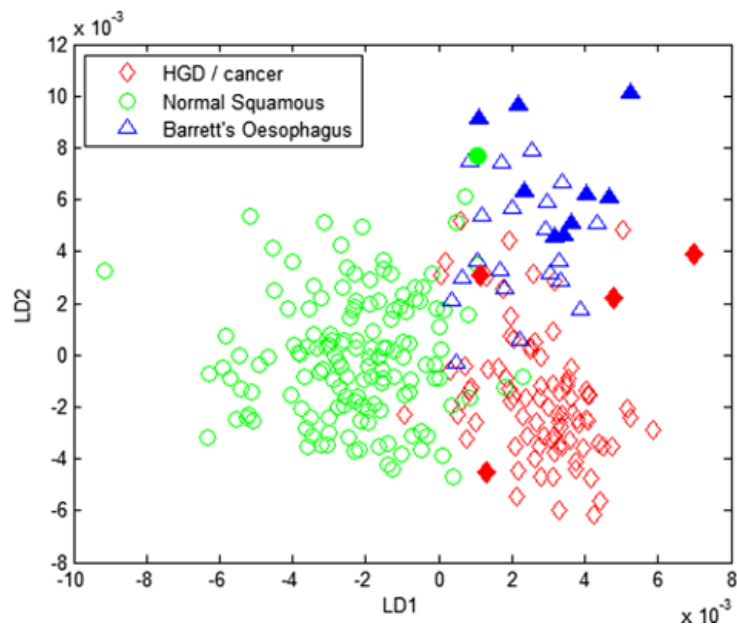
- Présentations et le matériel mis en ligne
- [Page GitHub du Club de Bioinfo de l'IBIS](#)

Exploration et classification de données

Objectifs

- Découvrir 3 techniques statistiques
- Exploration et de classification de données
- PCA, LDA, Random Forest
- Tester ces techniques dans R
- Faire un survol superficiel

PCA



PCA

Description

- Explorer les sources de variances principales.
- Aucune connaissances *a priori* des données.
- Untrained, non-supervisé.
- Visualisation.

Utilité en biologie/génétique/génomique

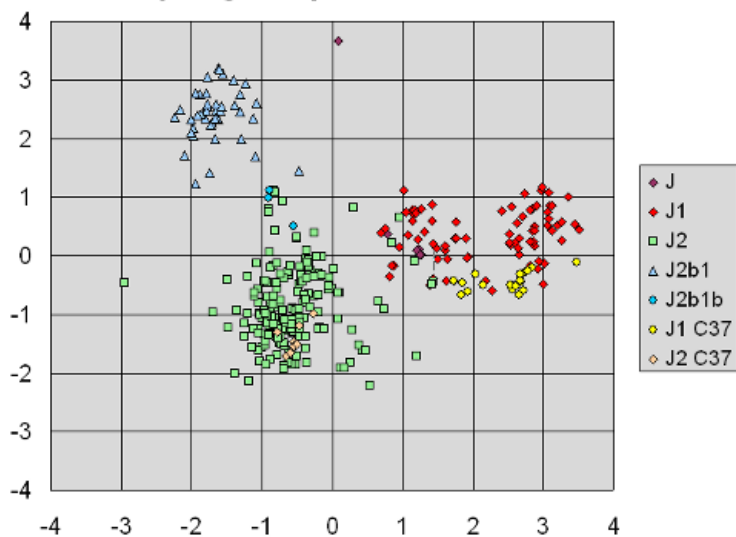
- Variables qui permettent de regrouper des échantillons.
- Importance de différentes sources de variation.
- Visualiser le regroupement ‘naturel’ des données.

Limites

- Complicé d’avoir une *p-valeur* des variables par axe.
- Plutôt, descriptif.

LDA

Haplogroup J - 37 STRs



LDA

Description

- Prédire l'appartenance d'échantillons.
- Connaissance *a priori* de la classification.
- Trained, supervisé.
- Problème de classification.
- Visualisation.

Utilité en biologie/génétique/génomique

- Classification d'individus en catégories
- Ex : sain, infecté.
- Ex : Individus comme bio-marqueurs de pollution.

Limites

- Grand nombre d'échantillons d'entraînement.
- Moins approprié/puissant pour grand nombre de variables.
- Données à classer doivent être similaires aux données test.

Random Forest



Random Forest

Description

- Classification ou régression.
- Connaissance *a priori* de la classification.
- Trained, supervisé.
- Problème de classification ou de régression.
- Visualisation.

Utilité en biologie/génétique/génomique

- Très approprié si le nombre de variables est très élevé
- Ex : expression de gènes ou marqueurs SNPs.
- Classification d'individus en catégories
- Ex : Utilisation d'individus comme bio-marqueurs de pollution.
- Identifier variables importantes (valeur d'importance).
- Prédiction de réponse par régression.

Limites

- Méthode encore plutôt nouvelle.
- Valeur d'importance réduite pour marqueurs génétiques liés.
- Données à classer doivent être similaires aux données test.