Problème 1 : Études d'association pangénomiques

Le but des études d'associations pangénomiques est de déterminer quelles régions du génome sont associées à un caractère observable $Y \in \mathbb{R}$, appelé *phénotype* (par exemple la couleur des yeux, la taille, le groupe sanguin). On représente le génome par un vecteur $X = (X_1, X_2, \ldots, X_m)$ dont chacune des composantes $X_j \in \mathbb{R}$ caractérise une position spécifique.

Diverses techniques de génotypage permettent d'acquérir un échantillon de taille n des couples (X, Y). On fait alors m tests statistiques, évaluant chacun l'association entre X_j et Y, pour $j \in [\![1, m]\!]$.

La figure 1 représente les résultats d'une telle étude : chaque test j est représenté par un point, d'abscisse j et d'ordonnée $-\log_{10}$ de sa p-valeur.

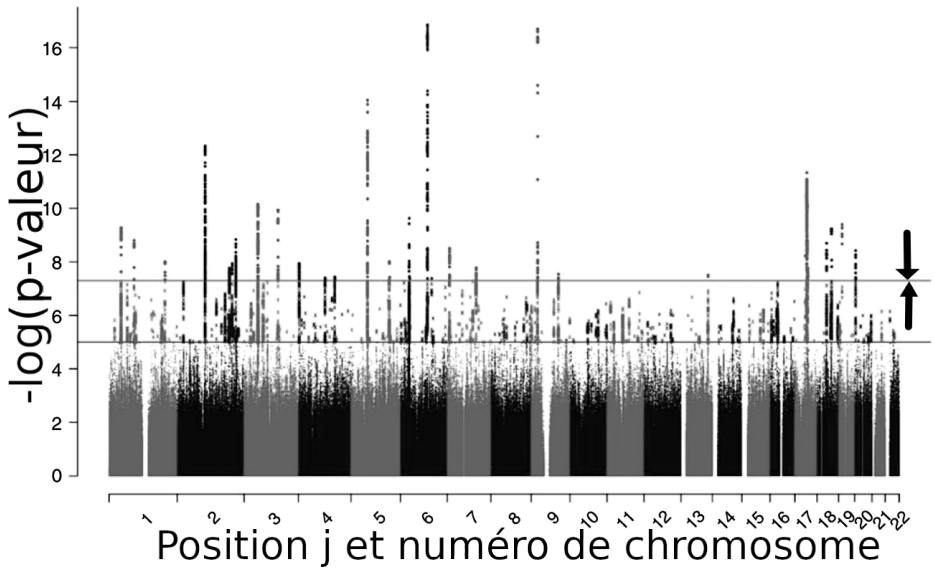

Figure 1 – Association entre différentes régions du génome et un phénotype.

1. (1 point) Quelle est l'hypothèse nulle du $j$ème test ?

Solution: La valeur de Y est indépendante de la valeur de X_j.

2. (1 point) Sur la figure 1, la ligne horizontale la plus haute (pointée par deux flèches) est tracée à l'ordonnée $y = -(\log_{10}(0.05) - \log_{10}(m))$. Pourquoi ?

Solution: Correction pour test d'hypothèses multiples à un niveau de signification $\alpha = 0.05$: les tests significatifs sont au-dessus de la ligne.

3. (4 points) La figure 1 est tirée d'une publication scientifique dans laquelle le phénotype Y est le revenu avant impôt du foyer de l'individu. Cette figure indique-t-elle que la capacité à gagner sa vie est déterminée par notre génome ? Discutez.

Pistes de réflexion : À quoi la corrélation entre X_j et Y peut-elle être due ? Quelle est l'hypothèse alternative du test j ?

Solution:

- Pourquoi se poser cette question ?
- Association via des maladies ? L'ethnicité ?
- Correlation $\neq$ causation
- Que certains X_j soient associées à Y n'implique pas qu'on peut parfaitement prédire Y à partir de ces X_j (pensez à un nuage de point de corrélation $R = 0.6$.

Problème 2 : Détection de spam

Nous voulons construire un modèle de détection de spams. Nous disposons d'un jeu de données $\mathcal{D} = \{(\vec{x}^{\,1}, y^1), (\vec{x}^{\,2}, y^2), \ldots, (\vec{x}^{\,n}, y^n)\}$, dans lequel, pour $i \in [\![1, n]\!]$, $\vec{x}^{\,i} \in \{0, 1\}^m$ est une représentation d'un email et $y^i \in \{0, 1\}$ vaut 0 si l'email est légitime et 1 s'il s'agit d'un spam.

On utilise pour $\vec{x}^{\,i}$ une représentation « *bag-of-words* » : m est le nombre de mots contenus dans un dictionnaire, et $x^i_j \in \{0, 1\}$ vaut 1 si le j-ème mot du dictionnaire apparaît dans le i-ème email et 0 sinon.

4. (1 point) De quel type de problème d'apprentissage statistique s'agit-il ?

Solution: Apprentissage supervisé, classification binaire.

Modélisation probabiliste

Nous supposons que $\mathcal{D}$ est un échantillon d'un couple (X, Y) dans lequel X est un vecteur aléatoire de dimension m et Y est une variable aléatoire discrète.

Nous faisons de plus l'hypothèse que les composantes de X sont indépendantes conditionnellement à Y : pour tout $j \neq k \in [\![1, m]\!]$, $\mathbb{P}(X_j = x_j | Y = y, X_k = x_k) = \mathbb{P}(X_j = x_j | Y = y)$.

5. ($\frac{1}{2}$ point) Cette hypothèse vous semble-t-elle réaliste ? Justifiez brièvement.

Solution: Au sein d'une classe, la présence du mot k ne change pas la probabilité de la présence du mot j. C'est peu réaliste, car les mots d'un même champ lexical n'ont aucune raison d'être indépendants, en particulier au sein d'une classe. Néanmoins, cette hypothèse va grandement simplifier la modélisation.

Nous pouvons maintenant écrire la loi jointe de (X, Y) comme

$$\mathbb{P}(X = (x_1, x_2, \ldots, x_m), Y = y) = \mathbb{P}(Y = y) \prod_{j=1}^{m} \mathbb{P}(X_j = x_j | Y = y),$$

et nous pouvons paramétrer cette loi jointe par $(2m+1)$ paramètres : $p = \mathbb{P}(Y = 1)$ et, pour $j \in [\![1, m]\!]$, q_{j0} et q_{j1} définis par $q_{j0} = \mathbb{P}(X_j = 1 | Y = 0)$ et $q_{j1} = \mathbb{P}(X_j = 1 | Y = 1)$.

6. (½ point) Supposons disposer d'estimations $\widehat{p}, \widehat{q_{j0}}$, et $\widehat{q_{j1}}$ de p, q_{j0}, et q_{j1}. Utilisez ces estimations pour exprimer $\mathbb{P}(Y = 0 | X = \vec{x})$ et $\mathbb{P}(Y = 1 | X = \vec{x})$.

   Remarquez que, pour $k \in \{0, 1\}$, $\mathbb{P}(X_j = x_j | Y = k) = q_{jk}^{x_j}(1 - q_{jk})^{1-x_j}$.

Solution: On utilise la loi de Bayes :

$$\mathbb{P}(Y = 1 | X = \vec{x}) = \frac{\mathbb{P}(X = \vec{x} | Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = \vec{x})} \approx \frac{\prod_{j=1}^{m} \left(\widehat{q_{j1}}^{x_j}(1 - \widehat{q_{j1}})^{(1-x_j)}\right) \widehat{p}}{\mathbb{P}(X = \vec{x})}$$

$$\mathbb{P}(Y = 0 | X = \vec{x}) \approx \frac{\prod_{j=1}^{m} \left(\widehat{q_{j0}}^{x_j}(1 - \widehat{q_{j0}})^{(1-x_j)}\right) (1 - \widehat{p})}{\mathbb{P}(X = \vec{x})}$$

7. (1 point) Comment prédire si $\vec{x}$ est un spam ?

Solution: On retourne l'étiquette « spam » si $\mathbb{P}(Y = 1 | X = \vec{x}) \geq \mathbb{P}(Y = 0 | X = \vec{x})$, autrement dit si

$$\prod_{j=1}^{m} \left(\widehat{q_{j1}}^{x_j}(1 - \widehat{q_{j1}})^{(1-x_j)}\right) \widehat{p} \geq \prod_{j=1}^{m} \left(\widehat{q_{j0}}^{x_j}(1 - \widehat{q_{j0}})^{(1-x_j)}\right) (1 - \widehat{p}).$$

8. (2 points) Écrivez la log-vraisemblance de $\mathcal{D}$ en fonction de (p, q_{j0}, q_{j1}).

Solution: La log-vraisemblance de $\mathcal{D}$ s'écrit

$$\begin{aligned}\ell(\mathcal{D}; p, q_{j0}, q_{j1}) &= \sum_{i=1}^{n} \ln \mathbb{P}(X = \vec{x}^{\,i}, Y = y^i) = \sum_{i=1}^{n} \ln \mathbb{P}(Y = y^i) + \sum_{j=1}^{m} \ln \mathbb{P}(X_j = x_j | Y = y) \\ &= \sum_{i=1}^{n} y^i \ln p + (1 - y^i) \ln(1-p) + \sum_{j=1}^{m} y^i \left(x_j^i \ln q_{j1} + (1 - x_j^i) \ln(1 - q_{j1})\right) + \\ &\qquad (1 - y^i) \left(x_j^i \ln q_{j0} + (1 - x_j^i) \ln(1 - q_{j0})\right).\end{aligned}$$

9. (2 points) Déterminez les estimations par maximum de vraisemblance des paramètres p, q_{j0}, et q_{j1}.

   Pour alléger les calculs, faites appel à l'estimation par maximum de vraisemblance du paramètre d'une Bernoulli, qui est la moyenne de l'échantillon.

   Notez $n_{\text{POS}} = \sum_{i=1}^{n} y^i$ le nombre de spams dans les données.

Solution: Il s'agit de trouver $\widehat{p}$, $\widehat{q_{j0}}$, et $\widehat{q_{j1}}$ qui maximisent

$$\begin{aligned}&\sum_{i=1}^{n} y^i \ln p + (1 - y^i) \ln(1-p) + \\ &\quad \sum_{j=1}^{m} y^i \left(x_j^i \ln q_{j1} + (1 - x_j^i) \ln(1 - q_{j1})\right) + (1 - y^i) \left(x_j^i \ln q_{j0} + (1 - x_j^i) \ln(1 - q_{j0})\right).\end{aligned}$$

On peut séparer l'optimisation par rapport à chacun des paramètres, et chacun prend la même forme que la maximisation de la log-vraisemblance d'une Bernoulli.

Ainsi

$$\begin{aligned}\widehat{p} &= \arg\max_{p \in]0,1[} \sum_{i=1}^{n} y^i \ln p + (1 - y^i) \ln(1-p) = \frac{1}{n} \sum_{i=1}^{n} y^i = \frac{n_{\text{POS}}}{n} \\ \widehat{q_{j0}} &= \arg\max_{q \in]0,1[} \sum_{i=1}^{n} (1 - y^i) \left(x_j^i \ln q + (1 - x_j^i) \ln(1 - q)\right) \\ &= \arg\max_{q \in]0,1[} \sum_{i \in \text{NEG}} x_j^i \ln q + (1 - x_j^i) \ln(1 - q) = \frac{\sum_{i=1}^{n} (1 - y^i) x_j^i}{n - n_{\text{POS}}} \\ \widehat{q_{j1}} &= \frac{\sum_{i=1}^{n} y^i x_j^i}{n_{\text{POS}}}\end{aligned}$$

10. (1 point) Comment interprétez-vous l'estimation par maximum de vraisemblance de q_{j1}?

Solution: C'est la proportion de spams contenant le mot j.

11. (1 point) En quoi cette méthode de détection de spams vous paraît-elle limitée ?

Solution:

- On ne prend pas en compte l'ordre des mots
- L'hypothèse d'indépendance conditionnelle n'est pas vérfiée
- Les mots rares ont des contributions faibles.

Réseau de neurones artificiels

Nous allons maintenant utiliser un perceptron multi-couche pour apprendre à détecter si un email est un spam.

12. (1 point) Quelle fonction d'activation choisir pour l'unité de sortie ?

Solution: Classification binaire : activation logistique.

13. (1 point) Supposez que le réseau de neurones contient 3 couches intermédiaires, utilisant chacune la fonction $\tanh$ comme fonction d'activation, et comportant chacune H neurones et une unité de biais. Combien de paramètres comporte le modèle décrit par ce réseau ?

Solution:

- $(m+1)H$ poids de la couche d'entrée à la première couche cachée ;
- $(H+1)H$ poids de la première à la deuxième couche cachée ;
- $(H+1)H$ poids de la deuxième à la troisième couche cachée ;
- $(H+1)$ poids de la troisième couche cachée à l'unité de sortie.

Donc un total de $H(m+1+2(H+1)+1)+1 = 2H^2+(m+4)H+1$ paramètres.

14. (3 points) Formulez l'apprentissage de ce réseau comme un problème de minimisation du risque empirique :

- quel est l'espace des hypothèses ?
- quelle est la fonction de perte ?
- quel est le problème d'optimisation ?

Solution:

- L'espace des hypothèses est donné par l'architecture du réseau :

$$\mathcal{F} = \Bigg\{ \vec{x} \mapsto \underbrace{\sigma\left(w_0^3 + \sum_{q=1}^{H} w_q^3 \tanh\left(w_0^2 + \sum_{l=1}^{H} w_{lq}^2 \tanh\left(w_0^1 + \sum_{k=1}^{H} w_{kl}^1 \tanh\left(w_0^0 + \sum_{j=1}^{m} w_{jk}^0\right)\right)\right)\right)}_{f_{w^0,w^1,w^2,w^3}(\vec{x})};$$
$$w^0 \in \mathbb{R}^{(m+1)\times H}, w^1 \in \mathbb{R}^{(H+1)\times H}, w^2 \in \mathbb{R}^{(H+1)\times H}, w^3 \in \mathbb{R}^{(H+1)} \Bigg\}$$

- La fonction de perte est l'entropie croisée L_H.
- Le problème d'optimisation est :

$$\underset{w^0\in\mathbb{R}^{(m+1)\times H}, w^1\in\mathbb{R}^{(H+1)\times H}, w^2\in\mathbb{R}^{(H+1)\times H}, w^3\in\mathbb{R}^{(H+1)}}{\arg\max} \sum_{i=1}^{n} L_H\left(y^i, f_{w^0,w^1,w^2,w^3}(\vec{x}^{\,i})\right)$$

15. (1 point) Sait-on résoudre ce problème d'optimisation ? Comment procéder en pratique ?

Solution: Non car non convexe. On utilise une méthode de gradient mais sans garantie.

16. (1 point) Ce perceptron multi-couche vous semble-t-il être un bon outil pour créer un détecteur de spam ?

Solution: Beaucoup trop de paramètres à apprendre (car m est très grand) ; problème d'optim difficile.

17. (1 point) Parmi les algorithmes d'apprentissage vus en cours, lequel vous semble adapté à ce problème de détection de spam et pourquoi ?

Solution: Forêts aléatoires : variables binaires, sortie binaire, chaque arbre est construit sur peu de mots.
Plus proches voisins peu adapté car dimension très grande.