Chapitre 6

Ouvertures

6.1 Régression logistique

Modèle

On se pose maintenant dans le cas où une variable qualitative Y a 2 modalités : 1 ou bien 0. Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\beta$ ne prend pas des valeurs simplement binaires. Si l'on ne connait que Y, on pourra estimer le paramètre Π de la loi de Bernoulli, $\mathbb{P}(Y = 1) = \Pi$ et $\mathbb{P}(Y = 0) = 1 - \Pi$, en calculant la moyenne empirique de $\mathbb{P}(Y = 1)$. On va cependant s'intéresser ici au cas où Y est lié à n observations en dimension p.

On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X. On suppose alors que :

$$\ln \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{6.1}$$

où les X_j sont les p composantes de X. Il est intéressant de remarquer que $\ln(\mathbb{P}/(1 - \mathbb{P}(.)))$ est une fonction de $\mathbb{P}(.)$ strictement croissante qui :

— tend vers $-\infty$ quand $\mathbb{P}(.)$ se rapproche de 0,
— vaut 0 pour $\mathbb{P}(.) = 0.5$,
— tend vers $+\infty$ quand $\mathbb{P}(.)$ se rapproche de 1.

On en conclue que Y à plutôt des chances de valoir 0 si $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$ est négatif, et que Y à plutôt des chances de valoir 1 $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$ est positif. Notons aussi que le modèle de régression est dit logistique car la loi de probabilité est modélisée à partir d'une loi logistique. Ce modèle est extremement populaire en apprentissage automatique car il se montre performant quand on n'a que deux classes à distinguer, et il passe facilement à l'échelle.

Apprentissage des β_j

Après transformation de l'équation, on obtient :

$$\mathbb{P}(Y = 1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$

que l'on notera pour une observation i, $i = 1, \ldots, n$:

$$p(y_i = 1|x_i^1, \ldots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_i^j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_i^j}} .$$

Pour résoudre estimer les β_j à l'aide d'un jeu de n observations $y_i, x_i^1, \ldots, x_i^p$, $i = 1, \ldots, n$ et de la méthode du maximum de vraisemblance, on note la contribution à la vraisemblance de l'observation i :

$$\big(p(y_i = 1|x_i^1, \ldots, x_i^p)\big)^{y_i} \cdot \big(1 - p(y_i = 1|x_i^1, \ldots, x_i^p)\big)^{1-y_i}$$

qui vaut $p(y_i = 1|x_i^1, \ldots, x_i^p)$ si $y_i = 1$ et qui vaut $p(y_i = 0|x_i^1, \ldots, x_i^p)$ si $y_i = 0$. La vraisemblance des observations s'écrit alors :

$$L(\beta) = \prod_{i=1}^n \left[\big(p(y_i = 1|x_i^1, \ldots, x_i^p)\big)^{y_i} \cdot \big(1 - p(y_i = 1|x_i^1, \ldots, x_i^p)\big)^{1-y_i} \right]$$

Les paramètres β_j qui maximisent cette quantité sont les estimateurs du maximum de vraisemblance de la régression logistique. Ils seront estimés typiquement en utilisant une méthode itérative. Pour des raisons numériques la log-vraisemblance, *i.e.* $n^{-1} \log(L)$, sera aussi maximisé plutôt que L.

Prédiction

Une fois les β_j appris, on se réfère à Eq. (6.1) et son interprétation pour prédire le label d'un y_0 en fonction de x_0^j observés. On calculera simplement $\beta_0 + \sum_{j=1}^p \beta_j x_0^j$. Si le signe est positif alors $\hat{y_0} = 1$ et si le signe est négatif alors $\hat{y_0} = 0$.

Sélection de modèle

Notons enfin qu'il est possible et même classique de sélectionner un modèle en régression logistique en pénalisant les β_j lors de la maximisation de la Log-vraisemblance, typiquement avec une méthode de type Lasso. On résoudra alors le problème suivant :

$$\hat{\beta} = \arg\max_{\beta} \left(n^{-1} \log\left(L(\beta)\right) - \lambda |\beta|_1 \right)$$

où comme pour la régression linéaire multiple, on sera amené à trouver un λ qui offrira un bon compromis entre pouvoir prédictif et explicabilité du modèle.

6.2 Méthode Partial Least Squares

Intuition

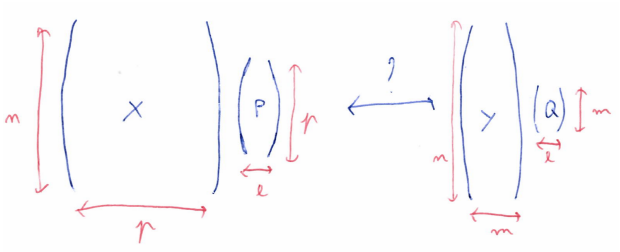
On a vu dans le cours de statistique que l'Analyse en Composantes Principales (ACP) était un outil essentiel pour explorer un ensemble d'observations $X_i = (x_i^1, \ldots, x_i^p)$, $i = 1, \ldots, n$ regroupées en ligne dans une matrice $\mathbf{X}$. L'ACP consiste en effet à maximiser la variance des projections des observations X_i, ce qui permet entre autres d'expliquer comment les variables interagissent

entre elles. Plus spécifiquement, le 1er vecteur propre v_1 est celui qui maximise la variance des projections des X_i. En supposant que les X_i sont centrés (et idéalement réduits), cela signifie que

$$v_1 = \underset{v \text{ t.q. } |v|_2=1}{\arg\max} \sum_{i=1}^{n} (X_i v)^2$$

Le 2ème vecteur propre v_2 est choisi suivant le même principe, une fois enlevée l'influence de v_1 dans $\mathbf{X}$; et ainsi de suite.

L'idée de la méthode *Partial Least Squares* (PLS) est relativement similaire, mais maintenant on s'intéresse au lien entre $\mathbf{X}$ et une matrice $n \times m$ de réponses $\mathbf{Y}$. Pour chaque observation X_i de $\mathbf{X}$, la matrice $\mathbf{Y}$ contient une réponse Y_i en dimension m. Si $m = 1$, on a les mêmes données d'entrée que dans le cadre de la régression linéaire multiple (Section 2.2). L'approche d'analyse est cependant totalement différente : On cherche les transformations linéaires $\mathbf{P}$ et $\mathbf{Q}$ de $\mathbf{X}$ et de $\mathbf{Y}$ (si $m > 1$), respectivement, telles que : La 1ère colonne de $\mathbf{P}$ est celle qui projete les X_i de manière à séparer au mieux les y_i projetés par la première colonne de $\mathbf{Q}$; et ainsi de suite. Cette idée est shématisée ci-dessous :



Modèle

La méthode PLS (Partial Least Squares) repose toujours sur une hypothèse de modèle linéaire $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, où $\mathbf{U}$ modélise le bruit. L'approche utilisée pour estimer le lien entre $\mathbf{Y}$ et les variables explicatives de $\mathbf{X}$ est cependant différente de celle du modèle linéaire classique. En particulier, le modèle sur le bruit est totalement différent et va dépendre de la covariance entre des combinaisons linéaires de $\mathbf{X}$ et $\mathbf{Y}$.

Plus spécifiquement, on suppose :

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F}$$

où

- $\mathbf{X}$ est la matrice $n \times p$ de prédicteurs. Elle est supposée centrée/réduite,
- $\mathbf{Y}$ est la matrice $n \times m$ de réponses. Elle est supposée centrée/réduite,
- $\mathbf{P}$ et $\mathbf{Q}$ sont respectivement des matrices $p \times l$ et $m \times l$ de projection. Leurs colonnes sont orthonormés.
- $\mathbf{T}$ et $\mathbf{U}$ sont les projections de $\mathbf{X}$ et de $\mathbf{Y}$ respectivement par $\mathbf{P}$ et $\mathbf{Q}$. Elles sont de taille $n \times l$.
- $\mathbf{E}$ et $\mathbf{F}$ sont des termes d'erreur de même taille que $\mathbf{X}$ et $\mathbf{Y}$. Ils sont supposés *i.i.d.* et distribués suivant une loi normale.

Les projections de $\mathbf{X}$ et de $\mathbf{Y}$ dans $\mathbf{T}$ et $\mathbf{U}$ sont aussi toutes deux de même taille $n \times l$ avec $l \leq p$. La PLS consiste alors à calculer les projecteurs $\mathbf{P}$ et $\mathbf{Q}$ qui maximisent la covariance entre et $\mathbf{T}$ et $\mathbf{U}$. On dénote $\bar{\mathbf{T}}_j$ et $\bar{\mathbf{U}}_j$ la moyenne des valeurs des colonnes j de $\mathbf{T}$ et $\mathbf{U}$. On maximise alors $\sum_{j=1}^{l}\sum_{i=1}^{n}(\mathbf{T}_{ij} - \bar{\mathbf{T}}_j)(\mathbf{U}_{ij} - \bar{\mathbf{U}}_j)$.

Estimation

Géométriquement, la régression PLS consiste à calculer une projection des $\mathbf{X}$ sur un hyperplan qui est à la fois une bonne estimation de $\mathbf{X}$ et dont les projections sont de bons prédicteurs des $\mathbf{Y}$. En vue de la définition d'une stratégie d'optimisation, le problème peut être vu sous la forme plus classique $\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}} + \hat{\mathbf{B}_0}$. Nous donnons Alg. 1 l'algorithme PLS1 qui permet de résoudre le problème pour $m = 1$, c'est à dire $\mathbf{Y}$ est un vecteur colonne. Dans ce cas là, $\hat{\mathbf{B}}$ est un vecteur de taille p dont l'interprétation est similaire aux vecteurs $\hat{\beta}$ du modèle linéaire multiple mais avec un modèle sous-jacent différent.

Alg. 1 Fonction $PLS1(\mathbf{X}, \mathbf{y}, l)$

$\mathbf{X}^{(0)} \leftarrow \mathbf{X}$
$\mathbf{w}^{(0)} \leftarrow \mathbf{X}^{'}\mathbf{y}/|\mathbf{X}^{'}\mathbf{y}|_2$.
for $k = 0, \ldots, l-1$ **do**
  $\mathbf{t}^{(k)} \leftarrow \mathbf{X}^{(k)}\mathbf{w}^{(k)}$
  $t_k \leftarrow {\mathbf{t}^{(k)}}^{'}\mathbf{t}^{(k)}$
  $\mathbf{t}^{(k)} \leftarrow \mathbf{t}^{(k)}/t_k$
  $\mathbf{p}^{(k)} \leftarrow {\mathbf{X}^{(k)}}^{'}\mathbf{t}^{(k)}$
  $q_k \leftarrow \mathbf{y}^{'}\mathbf{t}^{(k)}$
  if $q_k = 0$ **then**
    $l \leftarrow k$ et sort de la boucle **for** (toute la variabilité est capturée).
  end if
  if $k < (l-1)$ **then**
    $\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} - t_k\mathbf{t}^{(k)}{\mathbf{p}^{(k)}}^{'}$
    $\mathbf{w}^{(k+1)} \leftarrow {\mathbf{X}^{(k+1)}}^{'}\mathbf{y}/|{\mathbf{X}^{(k+1)}}^{'}\mathbf{y}|_2$
  end if
end for
$\mathbf{W}$ est la matrice composée des colonnes $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(l-1)}$.
$\mathbf{P}$ est la matrice composée des colonnes $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(l-1)}$.
$\mathbf{q}$ est le vecteur composé des scalaires $q_0, q_1, \ldots, q_{l-1}$.
$\mathbf{B} \leftarrow \mathbf{W}(\mathbf{P}^{'}\mathbf{W})^{-1}\mathbf{q}$
$\mathbf{B_0} \leftarrow q_0 - {\mathbf{P}^{(0)}}^{'}\mathbf{B}$
return $\mathbf{B}, \mathbf{B_0}$

Sparse PLS

On a vu Chapitre 3 l'intérêt pratique des méthodes de régularisation, telles que LASSO, qui sélectionnent des modèles parcimonieux (sparse). En plus de bien contraindre le problème de régression, ces modèles sont en effet simples à interpréter, même pour des non-spécialistes de l'analyse de données.

Ce principe peut aussi s'appliquer dans le cas de la PLS, afin de trouver à la fois des changements de bases qui mettent en lien les $\mathbf{X}$ et $\mathbf{Y}$ de manière optimale, et qui permettent de regrouper des blocs de variables ayant une influence similaire lorsque $\mathbf{X}$ et $\mathbf{Y}$ sont mis en lien. La méthode de la *sparse PLS* est alors extrêmement puissante d'un point de vue pratique.

En posant par exemple une pénalisation L_1 sur les éléments de la base $\mathbf{P}$ avec une pondération λ_P, Alg. 1 sera légèrement modifié en rajoutant la ligne suivante entre les lignes 7 et 8 : $\mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \lambda_P sign(\mathbf{p}^{(k)})$, où $sign(\mathbf{p}^{(k)})$ est un vecteur colone contenant des $\{-1, 0, 1\}$ en fonction de si chaque élément de $\mathbf{p}^{(k)}$) est respectivement négatif, nul ou positif. Notons que tout comme pour la régularisation LASSO en régression, un élément de $\mathbf{p}^{(k)}$ sera mis à zéro si son signe change pendant cette opération. Seuls les éléments de $\mathbf{p}^{(k)}$ ayant une réelle influence (en fonction de la *pression* de λ_P) auront alors une valeur non nulle et seront sélectionnés.