Master de Chemoinformatique/Master In Silico Drug Design M1S1
Examen de mathématiques pour la chimie*

MARCOU Gilles

Décembre 2011

Résumé

Documents autorisés. Durée : 2h. La copie de l'étudiant sera constituée d'un support papier traditionnel et d'un ou plusieurs fichiers de calcul pour Maple. ***Toute réponse doit être motivée ou sera considérée comme nulle.***

Exercice 1

Cette exercice illustre la régularisation de Tikhonov pour la régression multilinéaire. Dans la suite, on suivra les conventions suivantes :

- le nombre d'exemples connus pour construire un modèle est noté N_{tr} ;
- le nombre d'exemples inconnus sur lesquels le modèle est appliqué est noté N_{te} ;
- le nombre de variables explicatives est noté d ;
- la matrice contenant les valeurs des variables explicatives pour chacun des exemples connus est noté X_{tr} (N_{tr} lignes et d colonnes) ;
- la matrice contenant les valeurs des variables explicatives pour chacun des exemples inconnus est noté X_{te} (N_{te} lignes et d colonnes) ;
- le vecteur contenant la valeur de chacun des exemples connus de la variable expliquée est noté Y_{tr} (N_{tr} éléments) ;
- le vecteur contenant la valeur de chacun des exemples inconnus -utilisés à des fins de vérification- de la variable expliquée est noté Y_{te} (N_{te} éléments) ;
- le vecteur de dimension d qui contient les valeurs des paramètres du modèle linéaire est noté W (d éléments).

La régularisation de Tikhonov consiste à minimiser l'erreur du modèle de régression linéaire avec une contrainte sur l'amplitude des éléments de la matrice W. Plus formellement il s'agit de minimiser $\|X_{tr}.W - Y_{tr}\| + b\|W\|$ où b est un scalaire réel appelé paramètre de régularisation et les barres ($\|\cdot\|$) désignent la norme Euclidienne.

Les données sont stockées dans les fichiers `test_d.dat`, `train_d.dat`, `test_p.dat` et `train_p.dat.dat`.

Question 1.1

Définissez ce qu'est une norme.

---

*Enseignants: G. Marcou, *Université Louis Pasteur, Institut de Chimie, 4, rue Blaise Pascal, 67000 Strasbourg*

Question 1.2

Définissez ce qu'est un espace vectoriel Euclidien

Dans un premier temps, les dimensions du problème sont fixées à de petits nombres : $N_{tr} = 4$ et $d = 2$.

Question 1.3

Construisez les matrices X_{tr}, Y_{tr} et W initalisées avec les symboles x, y et w respectivement. Construisez une liste, notée `symb`, de tous les symboles utilisés dans votre vecteur W.

Question 1.4

Construisez l'expression à minimiser $\|X_{tr}.W - Y_{tr}\| + b\|W\|$.

Pour minimiser cette expression, il suffit d'annuler le gradient par rapport aux symboles de W, ceux contenus dans la liste `symb`.

Question 1.5

Définissez ce qu'est un gradient.

Question 1.6

Calculez le gradient. Utilisez la commande **Del** du paquet **VectorCalculus**.

Question 1.7

Vérifiez que l'expression obtenue est égale au vecteur dont l'expression est

$$2(X_{tr}^t(X_{tr}W - Y_{tr}) + bW) \tag{1}$$

avec X_{tr}^t désignant la transposée de X_{tr}. La différence entre ce vecteur et le gradient doit être nulle.

Annuler le gradient se traduit en une équation pour le vecteur W dont la solution est :

$$W = (X_{tr}^t X_{tr} + b)^{-1}(X_{tr}^t Y_{tr}) \tag{2}$$

Cette technique est maintenant utilisée pour modéliser la solubilité acqueuse de 818 composés sur la base de 48 descripteurs moléculaires et appliquée au calcul de la solubilité acqueuse de 817 autres composés. Les données doivent être chargées dans Maple.

Question 1.8

Utilisez l'expression de W annulant le gradient pour estimer numériquement ce vecteur à partir des données expérimentales. Choisissez $b = 6$.

Le modèle est ensuite utilisé sur le jeu de données test en calculant le produit matriciel de X_{te} avec W.

Question 1.9

Estimez la solubilité acqueuse pour chacune des 817 molécules du jeu de données test : $Y_{pred} = X_{te}.W$. Calculez l'erreur quadratique moyenne : $\mathrm{RMSE} = \sqrt{\|Y_{pred} - Y_{te}\|^2/N_{te}}$ où les barres $\|\cdot\|$ désigne la norme Euclidienne. Tracez sur un graphique Y_{pred} en fonction de Y_{te}.

Question 1.10

Ecrire une procédure pour calculer un modèle à partir des données (X_{tr},Y_{tr}), le testant sur les données (X_{te},Y_{te}) et calculant la RMSE prenant en paramètre une valeur pour b.

Question 1.11

Faire varier b entre 1 et 20 par pas de 1 et tracer l'évolution de l'erreur quadratique moyenne avec b.

Question 1.12

Livrez vos conclusions.

Exercice 2

Cet exercice concerne la *résolution cinétique classique*. Ce phénomène exploite la différence de constante cinétique entre les deux énantiomères d'un mélange racémique. Un exemple de réaction ayant cette caractéristique est l'addition de (+)-menthole à un mélange racémique d'acide phénylglycolique.

Dans un premier temps, on considère deux réactions :
$A_R + M \rightarrow B_R$ avec une constante de réaction k_R.
$A_S + M \rightarrow B_S$ avec une constante de réaction k_S.

Question 2.1

Ecrire et résoudre symboliquement les équations cinétiques des réactions précédentes en les supposant d'ordre 1, en supposant que M est en excès ($M(t) = M_0$) et en fixant les conditions initiales : $A_R(0) = A_S(0) = A_0$.

Question 2.2

Tracez sur un même graphique l'évolution des concentrations des espèce $A_R(t)$ et $A_C(t)$ entre 0 et 5 secondes. Donnez les valeurs numériques suivantes aux paramètres : $k_R = 2$, $k_S = 1$ et $A_0 = 1$. L'utilisation de la commande **subs** est recommandée pour conserver une expression symbolique des solutions.

Vous observerez qu'au cours de la réaction l'excès d'un énantiomère sur l'autre passe par un maximum : c'est la résolution cinétique classique. En général deux autres quantitiés sont préférées pour suivre la réaction.

Le taux de conversion (**conversion rate, c**) est le nombre de moles du mélange racémique consommées ramené au nombre de moles initiales

$$c(t) = \frac{A_R(0) + AS(0) - (A_R(t) + A_S(t))}{A_R(0) + A_S(0)} \tag{3}$$

L'excès énantiomérique (*enantiomeric excess, ee*) est la valeur absolue de la différence entre les concentrations des deux énantiomères, ramené à la somme de leurs concentrations

$$ee(t) = \frac{|A_R(t) - A_S(t)|}{A_R(t) + A_S(t)} \tag{4}$$

Question 2.3

Tracez sur un même graphique le taux de conversion et l'excès énantiomérique entre 0 et 5 secondes avec les valeurs numériques $k_R = 2$,$k_S = 1$ et $A_0 = 1$ et tracez un autre graphique en utilisant les valeurs numériques $k_R = 4$, $k_S = 1$ et $A_0 = 1$.

Question 2.4

Si, $|A_R(t) - A_S(t)| = A_S(t) - A_R(t)$, vérifiez que

$$\frac{\ln((1 - c(t)) * (1 - ee(t)))}{\ln((1 - c(t)) * (1 + ee(t)))} = \frac{k_S}{k_R} \tag{5}$$

Il sera nécessaire de supposer que k_R, k_S et t sont des réels à l'aide de la commande **assume**. Attention, il est nécessaire ici de remplacer la valeur absolue dans l'expression de $ee(t)$ par $|A_R(t) - A_S(t)| = A_S(t) - A_R(t)$ puisque dans les applications numériques, c'est effectivement le cas. Adaptez les expressions déjà préparées pour vous.

Question 2.5

Tracez sur un même graphique les courbes représentant $ee(t)$ en fonction de $c(t)$ avec les paramètres $k_S = 1$, $A_0 = 1$ et $k_R = 1.5, 2, 4, 8, 16$ et 32. Il s'agit de courbes paramétriques, utilisez donc l'aide sur les termes **?plot,parametric**. Les axes des graphiques couvrir l'étendue 0 à 1.

Barème

- Question 1.1 : 1 point
- Question 1.2 : 1 point
- Question 1.3 : 1 point
- Question 1.4 : 1 point
- Question 1.5 : 1 point
- Question 1.6 : 1 point
- Question 1.7 : 1 point
- Question 1.8 : 1 point
- Question 1.9 : 2 points
- Question 1.10 : 2 points
- Question 1.11 : 1.5 points
- Question 1.12 : 1 point
- Question 2.1 : 1.5 points
- Question 2.2 : 2 points
- Question 2.3 : 1 point
- Question 2.4 : 1 point
- Question 2.5 : 2 points