# 3 juin 2005 – Durée 3h

## Modalités

Les documents de cours et travaux dirigés ainsi que l'usage de calculatrices sont autorisés. **Aucun** livre ne l'est[1]. Il est recommandé de bien séparer les questions en faisant notamment apparaître leur numéro.

Les durées figurant auprès des titres des exercices sont des durées indicatives quant à la réalisation des exercices concernés. Le respect de ces durées pour la réalisation des exercices n'est bien entendu pas exigé.

## 1 Cours – 30 mn

1. Qu'est-ce que le phénomène de sur-apprentissage ? Donner une cause possible de l'apparition de ce phénomène (en vous référant par exemple aux perceptrons multi-couches) et une manière d'y remédier.
2. Expliquer succinctement ce qu'est la validation croisée et ce à quoi elle peut être utile.
3. Qu'est-ce que la représentation TD-IDF ? Quelles sont les idées importantes de ce codage ? (Il n'est pas nécessaire de rappeler les équations.)

## 2 Arbres de décision – 55 min

Dans cet exercice on utilisera le **critère d'entropie** pour la construction des arbres de décision. On rappelle qu'étant donné une distribution de probabilité $p_1, \ldots, p_n$ définie sur $n$ modalités, l'entropie de cette distribution est

$$Ent(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

où $\log_2$ est le logarithme en base 2 (qui est donc défini par $\log_2(x) = \ln(x)/\ln(2)$).

Exemple : si l'on demande à 10 personnes de choisir leur couleur préférée parmi les $n = 3$ couleurs `rouge`, `bleu` et `vert` et que l'on obtient la répartition des réponses selon le tableau suivant

| `rouge` | `bleu` | `vert` |
|---|---|---|
| 5 | 3 | 2 |

alors l'entropie de la distribution des couleurs préférées est

$$\begin{aligned} Ent(\frac{5}{10}, \frac{3}{10}, \frac{2}{10}) &= -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} \\ &= 1.48 \end{aligned}$$

(On remarquera que dans le cours et en travaux dirigés nous avons travaillé avec $n = 2$ puisque les problèmes de classification considérés étaient binaires, i.e. impliquaient deux classes.)

1. On se propose de construire un arbre de décision à partir du tableau 1 (page 2). Ce tableau répertorie les résultats de différents sportifs (pratiquant des sports éventuellement différents) décrits selon l'importance plus ou moins grande de l'épreuve à laquelle ils ont participé, de leur statut de favori ou non dans cette épreuve et de leur nationalité.

   (a) Construire un arbre de décision correspondant à ce jeu de données permettant d'estimer le résultat `victoire` ou `defaite` d'un sportif (rappel : utiliser le critère d'entropie). Fournir et expliquer les détails des calculs.

[1] Excepté les dictionnaires de traduction

| **Épreuve** | **Statut** | **Nationalité** | **Résultat** |
|---|---|---|---|
| majeure | favori | francais | defaite |
| majeure | outsider | francais | defaite |
| intermediaire | outsider | francais | victoire |
| intermediaire | favori | francais | defaite |
| mineure | favori | francais | defaite |
| majeure | favori | autre | victoire |
| majeure | outsider | autre | defaite |
| mineure | favori | autre | victoire |
| intermediaire | outsider | autre | defaite |
| intermediaire | favori | autre | victoire |

TAB. 1 – Résultats des sportifs

(b) Expliquer ce qu'est une feuille pure. Dans quelle situation est-il impossible de construire un arbre n'ayant que des feuilles pures ? Proposer une instance qui, ajoutée à l'ensemble des données du tableau 1 (page 2) conduit à cette situation.

2. Supposons qu'une feuille d'un arbre de décision ne soit pas pure ; comment classifier une instance dont la description conduit à cette feuille ? Proposer une méthode (très simple) pour quantifier la confiance qu'on peut avoir dans la classification d'une instance qui se fait à partir d'une feuille impure (indice : cette confiance prendra une valeur entre 0 et 1) ?
3. **Une question sur votre capacité de généralisation !** Dans cet exercice il est demandé de construire un arbre de décision pour un problème de classification à trois classes. Plus précisément, le but est de construire un arbre de décision à partir des données du tableau 2 (page 3) permettant de déterminer la nature de lentille de contact (`aucune`, `souple`, `dure`) qu'un individu peut porter en fonction de caractéristiques physiologiques (âge, type d'anomalie de la vue, présence d'astigmatie, production de larmes). Construire l'arbre de décision associé à cet échantillon d'apprentissage.

# 3 Noyaux – 20mn

Dans cet exercice, on désigne par $\langle \cdot, \cdot \rangle_d$ l'application

$$\langle \cdot, \cdot \rangle_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$
$$(\mathbf{u}, \mathbf{v}) \mapsto \langle \mathbf{u}, \mathbf{v} \rangle_d = \sum_{i=1}^{d} u_i v_i$$

1. Soit l'application $\phi$ définie par

$$\phi : \mathbb{R}^2 \to \mathbb{R}^6$$
$$\mathbf{u} \mapsto \phi(\mathbf{u}) = \left[3u_1, -3u_2, 2u_1u_2, \sqrt{2}u_1^2, \sqrt{2}u_2^2, \sqrt{2}\right]^\top$$

simplifier l'écriture du noyau $k$ défini par $k(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_6$ en faisant en particulier apparaître le produit scalaire $\langle \mathbf{u}, \mathbf{v} \rangle_2$

2. Dans cette question, on suppose que $d$ est un entier naturel fixé. Soit l'application $k_{\cos}$ définie par

$$k_{\cos} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$
$$(\mathbf{u}, \mathbf{v}) \mapsto k_{\cos}(\mathbf{u}, \mathbf{v}) = \cos(\mathbf{u}, \mathbf{v})$$

où $\cos(\mathbf{u}, \mathbf{v})$ désigne le cosinus de l'angle entre les deux vecteurs $\mathbf{u}$ et $\mathbf{v}$. Trouver une application $\phi_{\cos}$ de $\mathbb{R}^d$ dans $\mathbb{R}^d$ telle que

$$k_{\cos}(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_d$$

| Age | Type anomalie | Astigmatie | Production larmes | Prescription |
|---|---|---|---|---|
| jeune | myope | non | réduite | aucune |
| jeune | myope | non | normale | souple |
| jeune | myope | oui | réduite | aucune |
| jeune | myope | oui | normale | dure |
| jeune | hypermétrope | non | réduite | aucune |
| jeune | hypermétrope | non | normale | souple |
| jeune | hypermétrope | oui | réduite | aucune |
| jeune | hypermétrope | oui | normale | dure |
| adulte | myope | non | réduite | aucune |
| adulte | myope | non | normale | souple |
| adulte | myope | oui | réduite | aucune |
| adulte | myope | oui | normale | dure |
| adulte | hypermétrope | non | réduite | aucune |
| adulte | hypermétrope | non | normale | souple |
| adulte | hypermétrope | oui | réduite | aucune |
| adulte | hypermétrope | oui | normale | aucune |
| senior | myope | non | réduite | aucune |
| senior | myope | non | normale | aucune |
| senior | myope | oui | réduite | aucune |
| senior | myope | oui | normale | dure |
| senior | hypermétrope | non | réduite | aucune |
| senior | hypermétrope | non | normale | souple |
| senior | hypermétrope | oui | réduite | aucune |
| senior | hypermétrope | oui | normale | aucune |

TAB. 2 – Prescription de lentilles

3. Soit l'application $k_q$ définie par

$$k_q : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$
$$(\mathbf{u}, \mathbf{v}) = ([u_1\ u_2]^\top, [v_1\ v_2]^\top) \mapsto k_q(\mathbf{u}, \mathbf{v}) = u_1^2 v_1^2 + 4u_2 v_2 + 3u_1 v_1 u_2 v_2$$

Trouver une application $\phi$ de $\mathbb{R}^2$ dans $\mathbb{R}^3$ telle que

$$k_q(\mathbf{u}, \mathbf{v})) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_3 .$$

# 4 Réseaux RBF – 55mn

Les réseaux *Radial Basis Function* (RBF) sont des réseaux de neurones qui sont particulièrement adaptés aux problèmes de régression (note : ils sont également très utilisés pour la classification), c'est-à-dire aux problèmes où il faut associer une valeur réelle $y$ à un vecteur $\mathbf{x} \in \mathbb{R}^d$.

Quelques exemples d'utilisation de ce type de réseaux : estimation de l'espérance de vie d'un patient en fonctions de certaines caractéristiques physiologiques (poids, taille, âge, taux de globules blancs dans le sang, etc.), prévision de la température d'une ville un jour donné en fonction de la pression de l'air, de l'hygrométrie (taux d'humidité), de l'ensoleillement mesurés les jours précédents. En somme, l'espace des $y$ peut être n'importe quel intervalle de $\mathbb{R}$ et non pas simplement un ensemble discret comme c'est le cas pour des problèmes de classification.

Dans cet exercice on se propose d'établir les équations d'apprentissage d'un réseau RBF à partir d'un ensemble d'apprentissage $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ avec $\mathbf{x}_i \in \mathbb{R}^d$ et $y \in \mathbb{R}$. Etant

donné un vecteur $\mathbf{x} \in \mathbb{R}^d$, un réseau RBF calcule la valeur $f(\mathbf{x})$ à associer à $\mathbf{x}$ selon[2] :

$$f(\mathbf{x}; \mathbf{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma_1, \ldots, \sigma_K) = \sum_{k=1}^{K} w_k g_k(\mathbf{x}; \boldsymbol{\mu}_k, \sigma_k),$$
$$\text{avec } g_k(\mathbf{x}; \boldsymbol{\mu}_k, \sigma_k) = \exp\left(-\frac{1}{2\sigma_k^2}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right),$$

$K$ un entier positif fixé *a priori* et $\mathbf{w} = [w_1 \cdots w_k]^\top \in \mathbb{R}^K$. Les paramètres à estimer à partir de $\mathcal{S}$ sont $\mathbf{w}$ (appelé *vecteur de poids*), $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ (appelés *centres*) et $\sigma_1, \ldots, \sigma_K$ (appelés *largeurs*).

1. Supposons qu'il existe une distribution $p(X, Y)$ sur $\mathbb{R}^d \times \mathbb{R}$ à partir de laquelle l'échantillon i.i.d $\mathcal{S}$ a été généré.
   (a) Quelle est la signification de i.i.d (réponse très courte) ?
   (b) En s'aidant de $p(X, Y) = p(Y|X)p(X)$ donner l'expression générale de la probabilité $p((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ d'observer les exemples de $\mathcal{S}$.
2. Pour l'apprentissage d'un réseau RBF, on fait l'hypothèse que la densité $p(Y = y|X = \mathbf{x})$ définissant la probabilité d'associer la valeur $y$ au vecteur $\mathbf{x}$ est une loi normale de moyenne $f(\mathbf{x})$ et d'écart-type $s$ que l'on suppose connu. On a donc $Y \sim \mathcal{N}(f(\mathbf{x}), s^2)$.
   (a) Expliciter $p(Y = y|X = \mathbf{x})$ (en fonction de $\mathbf{w}$, $K$, $s^2$ et des $g_k$ – et de $y$ et $\mathbf{x}$ bien évidemment) sous cette hypothèse.
   (b) Montrer que la vraisemblance $\mathcal{V}$ de l'échantillon $\mathcal{S}$ par rapport au modèle RBF est donnée par

$$\mathcal{V}((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)) = \frac{1}{Z} \exp\left(-\frac{1}{2s^2}\sum_{i=1}^{\ell}(f(\mathbf{x}_i) - y_i)^2\right) \times \prod_{i=1}^{\ell} p(X_i = \mathbf{x}_i)$$

   où $Z$ est une constante dont on donnera l'expression.
   (c) En supposant que l'on peut négliger le terme $\prod_{i=1}^{\ell} p(X_i = \mathbf{x}_i)$ dans l'expression de $\mathcal{V}$, donner l'expression de la log-vraisemblance $\mathcal{L}$ de l'échantillon $\mathcal{S}$ par rapport au modèle RBF.
   (d) Expliquer pourquoi le problème d'optimisation à résoudre pour l'apprentissage RBF peut s'écrire de la manière suivante :

$$\min_{\mathbf{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma_1, \ldots, \sigma_K} \frac{1}{2}\sum_{i=1}^{\ell}(f(\mathbf{x}_i) - y_i)^2.$$

   Quelle est la valeur minimale que peut prendre la fonction à minimiser ?
3. En fait, le problème d'optimisation que l'on doit résoudre pour l'apprentissage RBF est plus précisément

$$\min_{\mathbf{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma_1, \ldots, \sigma_K} L(\mathcal{S}; \mathbf{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma_1, \ldots, \sigma_K) = \frac{1}{2}\sum_{i=1}^{\ell}(f(\mathbf{x}_i) - y_i)^2 + \frac{\lambda}{2\ell}\sum_{k=1}^{K} w_k^2$$

   où $\lambda$ est un réel positif.
   (a) Quelle est la particularité de la solution $f$ lorsque $\lambda$ est infiniment grand, c'est-à-dire lorsque $\lambda \to +\infty$ (indice : $f$ est constante) ? Justifier.
   (b) Quelle est l'utilité du terme $\frac{\lambda}{2\ell}\sum_{k=1}^{K} w_k^2$ ? Quel est le risque encouru lorsque l'on ne prend pas garde à l'inclure dans le problème d'optimisation ?

[2] Par souci de clarté et concision, la dépendance de $f$ par rapport à $\mathbf{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma_1, \ldots, \sigma_K$ est parfois omise et on écrira $f(\mathbf{x})$ à la place de $f(\mathbf{x}; \mathbf{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \sigma_1, \ldots, \sigma_K)$

4. (On suppose à partir de maintenant que $\lambda$ est fixé.) Il n'est pas possible de résoudre simplement le problème de minimisation précédent et on peut en trouver une solution par le moyen d'une descente de gradient. On s'intéresse à évaluer les gradients de $L$ par rapport aux différents paramètres à estimer.
   (a) Donner l'expression de $\nabla_{\boldsymbol{\mu}_p} L$, gradient de $L$ par rapport au centre $\boldsymbol{\mu}_p$
   (b) Donner l'expression de $\frac{\partial L}{\partial \sigma_p}$, dérivée partielle de $L$ par rapport à $\sigma_p$.
   (c) Montrer que le gradient $\nabla_{\mathbf{w}} L$ de $L$ par rapport à $\mathbf{w}$ vérifie bien

$$\nabla_{\mathbf{w}} L = \left( G^\top G + \frac{2\lambda}{\ell} \right) \mathbf{w} - G^\top \mathbf{y}$$

   où

$$G = \begin{bmatrix} g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & \cdots & g_K(\mathbf{x}_1) \\ g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & \cdots & g_K(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ g_1(\mathbf{x}_\ell) & g_2(\mathbf{x}_\ell) & \cdots & g_K(\mathbf{x}_\ell) \end{bmatrix},$$

   et où $G^\top$ est la transposée de $G$ et $\mathbf{y} = [y_1 \cdots y_\ell]^\top$
   (d) En déduire une condition nécessaire et suffisante pour que $\nabla_{\mathbf{w}} L = \mathbf{0}$.
5. (Question bonus.) Donner l'algorithme d'apprentissage par RBF.

# 5 Relaxation Lagrangienne (à traiter à la fin) – 20mn

On va utiliser la méthode de relaxation Lagrangienne pour résoudre le problème d'optimisation suivant :

$$\min_{q_1,\ldots,q_n} F(q_1, \ldots, q_n) = \sum_{i=1}^{n} q_i \ln q_i$$
$$\text{sous la contrainte} \sum_{i=1}^{n} q_i = 1$$

1. Ecrire le Lagrangien $L$ de ce problème d'optimisation.
2. Donner une condition nécessaire à l'obtention d'un minimum de ce problème à partir de $L$.
3. En déduire la valeur de $q_1, \ldots, q_n$.