## Apprentissage supervisé

## Exercice 1

On a observé les données suivantes : les features sont dans  $\mathbb{R}^2$  et les labels sont dans {rouge, bleu}.

- 1. Donner les valeurs de l'erreur empirique associée à la perte 0/1 des classifieurs construits par
  - l'algorithme des 1-plus proche voisins (1-NN)
  - l'algorithme des 3-plus proche voisins (3-NN).
- 2. Où metteriez vous le premier "split" d'un arbre de décision ? (vous pouvez le dessiner sur la figure)
- 3. A partir de quelle profondeur a-t-on un arbre d'erreur empirique nulle?



## Exercice 2

On considère le problème de régression avec une régularisation ridge :

$$\underset{w \in \mathbb{R}^d, c \in \mathbb{R}}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^n \left( y_i - (x_i^\top w_i + c) \right)^2 + \frac{\lambda}{2} \|w\|_2^2 \right)$$

avec, pour  $1 \leq i \leq n$ ,  $x_i \in \mathbb{R}^d$  le vecteur de prédicteurs,  $y_i \in \mathbb{R}$  la réponse,  $w \in \mathbb{R}^d$  le vecteur de poids du modèle et  $c \in \mathbb{R}$  l'intercept. On note de plus

$$f_i(w,c) = \left(y_i - (x_i^\top w_i + c)\right)^2 + \frac{\lambda}{2} ||w||_2^2$$
 and  $f(w,c) = \sum_{i=1}^n f_i(w,c)$ .

1. On définit la matrice de design X comme

$$X = \begin{pmatrix} x_1^\top \\ \dots \\ x_n^\top \end{pmatrix}.$$

Récrire la fonction f à minimiser en fonction du vecteur de paramètre  $\theta = (w, c)^{\top}$ , du vecteur  $Y = (y_1, \dots, y_n)^{\top}$  et de la matrice  $[\mathbf{1}|X]$  où  $\mathbf{1}$  est le vecteur contenant n fois la valeur 1 et  $[\cdot|\cdot]$  est la concaténation.

- 2. Calculer le gradient de f(w,c) noté  $\nabla f(w,c)$  par rapport au paramètre  $\theta = (w,c)^{\top}$ .
- 3. Dans la suite, on suppose qu'un petit nombre d'observations  $(x_i, y_i)$  sont des données aberrantes, dans le sens où  $y_i$  est loin de la valeur attendue. Au lieu d'avoir un intercept fixe  $c \in \mathbb{R}$ , on considère à présent un intercept par observation, c'est-à-dire  $c_i \in \mathbb{R}$  pour  $i = 1, \ldots, n$ , ce qui donne lieu au problème de minimisation suivant :

$$\underset{w \in \mathbb{R}^d, c \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n f_i(w, c_i) + \gamma ||c||_1 \right\}$$

où  $f_i(w,c)$  a la même forme que précédemment,  $\gamma > 0$  est un paramètre de régularisation supplémentaire, et où  $||c||_1$  est la norme  $\ell_1$  ( $||c||_1 = \sum_{i=1}^n |c_i|$ ). Ecrire la matrice de design associée à ce nouveau problème à partir de X et  $\mathrm{Id}_n$ , la matrice identité de taille  $n \times n$ . Quelle est sa dimension ?

- 4. Expliquer en quelques mots pourquoi on utilise la pénalisation  $\ell_1$  pour le paramètre c, ainsi que l'effet du paramètre de régularisation  $\gamma > 0$ .
- 5. Ecrire l'algorithme de descente de gradient proximal pour résoudre le problème d'optimisation.

## Exercice 3

- 1. Compléter le graphique du réseau de neurone pour la classification binaire avec une couche cachée à 1 neurone et dont la fonction d'activation en sortie de la couche cachée est g (vous pouvez vous aider des slides 5 à 7 du cours 4). On notera
  - $b^H$  le vecteur de bias de la couche cachée
  - $W^H$  le vecteur des poids de la couche cachée.
- 2. Préciser les dimensions de  $b^H,\,W^H,\,b^O,\,W^O$
- 3. Pour des valeurs fixées de  $b^H,\,W^H,\,b^O,\,W^O,\,$  donnez la forme mathématique de  $\hat{y}(x).$

On considère la perte logistique

$$\ell(y, \hat{y}(x)) = -y \log(\hat{y}(x)) - (1 - y) \log(1 - \hat{y}(x))$$

où y est le label observé et  $x=(x_1,\ldots,x^d)^{\top}$ . On suppose, pour simplifier, que  $b^H=b^O=0$  et que g est la fonction sigmoïde. On veut calculer les gradients en  $W^H$  et  $W^O$  par back-propagation.

4. Calculer successivement les gradients

$$\nabla_{\hat{y}(x)}\ell(y,\hat{y}(x)), \nabla_{z}\circ\hat{y}(x), \nabla_{W}\circ z^{O}$$

en déduire l'expression de  $\nabla_{WO}\ell(y,\hat{y}(x))$ .

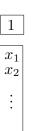
5. Vérifier votre calcul en remarquant que

$$\ell(y, \hat{y}(x)) = -y \log(\sigma(W^{O}h)) - (1 - y) \log(1 - \sigma(W^{O}h)).$$

- 6. Si  $W^{O,(k)}$  est la valeur de  $W^O$  à l'itération k, quelle sera sa valeur à l'itération k+1 pour une descente gradient de pas  $\eta$  ?
- 7. Continuer avec

$$\nabla_h W^O, \nabla_{z^H} h, \nabla_{W^H} z^H$$

en déduire  $\nabla_{W^H} \ell(y, \hat{y}(x))$ .



 $x_d$ 

