Machines à Vecteurs Supports

Benjamin Monmege
benjamin.monmege@lsv.ens-cachan.fr

5 avril 2012

Dans ce TD, on s'intéresse au problème de classification. On se donne deux *classes* $\mathcal{C}_1, \mathcal{C}_2$ formant une partition de l'espace $\mathbb{R}^D$ et des données $(x_n)_{1 \leq n \leq N}$ avec $x_n \in \mathbb{R}^D$ appartenant à l'une des classes. On définit $t_n = 1$ si $x_n \in \mathcal{C}_1$, et $t_n = -1$ si $x_n \in \mathcal{C}_2$. On cherche alors à apprendre les classes $\mathcal{C}_1, \mathcal{C}_2$ et à les utiliser pour prédire la classification de nouveaux exemples. On étudie les Machines à Vecteurs Supports (SVM).

Dans l'exercice 1, on va prouver le théorème suivant :

Théorème 1. *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable au voisinage d'un point $x^* \in \mathbb{R}^n$ et soient $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ ($i \in \{1, \dots, m\}$, $j \in \{1, \dots, \ell\}$) des fonctions affines de $\mathbb{R}^n$. Si x^* est un minimum local de f sous les contraintes $g_i(x) \leq 0$ et $h_j(x) = 0$, alors il existe des constantes μ_i et λ_j ($i \in \{1, \dots, m\}$, $j \in \{1, \dots, \ell\}$) vérifiant les 3 conditions suivantes :*

1. Stationnaire *La fonction $f + \sum_{i=1}^{m} \mu_i g_i + \sum_{j=1}^{\ell} \lambda_j h_j$ possède un gradient nul en x^*.*

2. Duale *Pour tout i, $\mu_i \geq 0$.*

3. Relâchement supplémentaire *Pour tout i, $\mu_i g_i(x^*) = 0$.*

Si, de plus, la fonction f est convexe, s'il existe x^* tel que pour tout i et j, $g_i(x^*) \leq 0$ et $h_j(x^*) = 0$, alors l'existence de ces constantes vérifiant les 3 conditions est suffisante pour prouver l'optimalité de x^*.

Exercice 1 (Multiplicateurs de Lagrange)**.** On cherche à prouver le Théorème 1.

1. Prouver le théorème dans le cas $m = 0 \wedge \ell = 1$. Montrer que trouver l'optimum x^* et la constante λ_1 équivaut à résoudre un problème d'optimisation de la fonction $L(x, \lambda_1) = f + \lambda_1 h_1$.
2. Prouver le théorème dans le cas $\ell = 0 \wedge m = 1$. Montrer que trouver l'optimum x^* et la constante μ_1 équivaut à résoudre un problème d'optimisation de la fonction $M(x, \mu_1) = f + \mu_1 g_1$ sous la contrainte $\mu_1 \geq 0$.
3. Prouver le théorème dans le cas général.

Dans la suite, on suppose fixée une fonction $\varphi : \mathbb{R}^D \to \mathbb{R}^M$, envoyant les données en entrée dans l'espace des *features* $\mathbb{R}^M$.

Exercice 2 (SVM dans le cas séparable)**.** Dans cet exercice, on suppose qu'il existe $\underline{w} \in \mathbb{R}^M$ et $\underline{b} \in \mathbb{R}$ tels que la classe $\mathcal{C}_1$ est définie par l'équation $\underline{w}^T \varphi(x) + \underline{b} \geq 0$: dans l'espace des *features*, les classes sont linéairement séparables. On définit $y(x) = w^T \varphi(x) + b$ pour des paramètres w, b quelconques. On appelle $\mathcal{H}_{w,b}$ l'hyperplan défini par l'équation $y(x) = 0$.

1. Montrer que la distance euclidienne d'une donnée x à l'hyperplan $\mathcal{H}_{w,b}$ vaut $|y(x)|/\|w\|$.
2. En déduire un problème de maximisation sous contrainte permettant d'exprimer qu'on cherche à trouver des paramètres w, b maximisant la *marge* entre l'hyperplan séparateur et les données (x_n). Trouver un problème de minimisation d'une fonction quadratique sous contrainte affine équivalent.

3. En utilisant le Théorème 1, trouver un problème dual équivalent qui dépend uniquement de nouveaux paramètres $\mu_1, \ldots, \mu_N$. En déduire l'existence d'un ensemble $\mathcal{S}$ d'indices n tels que les données x_n vérifient $t_n y(x_n) = 1$: on les appelle *vecteurs supports*.
4. Trouver alors les meilleurs valeurs à donner aux paramètres w et b pour apprendre les deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$, en fonction d'une solution du problème dual. Comment prédire la classe d'un nouvel exemple ?

Exercice 3 (SVM dans le cas non séparable)**.** On cherche maintenant à étendre la méthode précédente au cas où les classes $\mathcal{C}_1$ et $\mathcal{C}_2$ ne sont pas linéairement séparables dans l'espace des *features*.

1. Montrer que dans le cas linéairement séparable, on peut se ramener à minimiser la fonction d'erreur $\sum_{n=1}^{N} E_\infty(y(x_n)t_n - 1) + \lambda\|w\|^2$ en fonction des paramètres w et b avec une fonction E_∞ à définir.
2. Dans le cas non linéairement séparable, on relâche la pénalité E_∞ pour les points mal classifiés. Pour ce faire, on introduit des variables ressort $\xi_n \geq 0$ définies par $\xi_n = 0$ si l'exemple n vérifie $t_n y(x_n) \geq 1$ (c'est-à-dire la donnée x_n est bien classifiée et se trouve *loin* de l'hyperplan séparateur), et $\xi_n = |t_n - y(x_n)|$ sinon. Montrer pourquoi il est alors intéressant de minimiser en fonction des variables w, b, ξ_n la fonction d'erreur $C\sum_{n=1}^{N} \xi_n + \frac{1}{2}\|w\|^2$ sous les contraintes $t_n y(x_n) \geq 1 - \xi_n$, et $\xi_n \geq 0$, pour $n = 1, \ldots, N$: on a ici introduit un paramètre de régularisation $C > 0$. Que se passe-t-il lorsque $C \to \infty$?
3. À l'aide du Théorème 1, trouver un problème dual équivalent qui dépend uniquement de nouveaux paramètres (au nombre de $2N$). En déduire à nouveau l'existence d'un ensemble $\mathcal{S}$ d'indices n tels que les données x_n vérifient $t_n y(x_n) = 1 - \xi_n$.
4. Trouver alors les meilleurs valeurs à donner aux paramètres w et b pour apprendre les deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$, en fonction d'une solution du problème dual.
5. En référence, à la question 1, prouver qu'on peut se ramener à minimiser la fonction d'erreur $\sum_{n=1}^{N} E_H(y(x_n)t_n) + \lambda\|w\|^2$ en fonction des paramètres w et b avec une fonction $E_H(z)$ la partie positive de $1 - z$. En particulier, remarquer que $\lambda = 1/(2C)$.

Exercice 4 (Régression logistique)**.** Dans cet exercice, on suit une approche Bayésienne pour résoudre le problème de classification.

1. On suppose ici qu'on modélise le fait que la donnée x appartienne à la classe $\mathcal{C}_k$ ($k \in \{1, 2\}$) à l'aide des fonctions de densité par classe $\mathbb{P}(\phi(x) \mid \mathcal{C}_k)$ et qu'on possède une distribution a priori $\mathbb{P}(\mathcal{C}_k)$ sur les classes. En utilisant le théorème de Bayes, montrer que la probabilité a posteriori $\mathbb{P}(\mathcal{C}_1 \mid \phi(x))$ peut être décrite à l'aide de la fonction sigmoïde σ, $\mathbb{P}(\phi(x) \mid \mathcal{C}_k)$ et $\mathbb{P}(\mathcal{C}_k)$ ($k \in \{1, 2\}$).
2. On suppose désormais que les fonctions de densité par classe suivent des lois Gaussiennes ayant la même matrice Σ de covariance. Montrer qu'alors la probabilité a posteriori $\mathbb{P}(\mathcal{C}_1 \mid \phi(x))$ s'écrit sous la forme $\sigma(w^T\phi(x) + b)$ avec des paramètres w et b à déterminer.
3. Pour l'apprentissage, on se donne un ensemble $(x_n)_{1\leq n\leq N}$ de données, qu'on note $\mathbf{x}$, avec leur classification $(t_n)_{1\leq n\leq N}$ dans $\{-1, 1\}$, noté $\mathbf{t}$. On est donc ramené à chercher les meilleurs paramètres w et b en maximisant la fonction de vraisemblance $\mathbb{P}(\mathbf{t}|w, \mathbf{x})$. En posant $y_n = w^T\phi(x_n) + b$, expliquer l'égalité $\mathbb{P}(\mathbf{t} \mid w, \mathbf{x}) = \prod_{n=1}^{N} \sigma(y_n t_n)$. En déduire un algorithme de descente de gradient utilisant la fonction d'erreur définie à partir de l'opposé de la log-vraisemblance. Après avoir ajouté un terme $\lambda\|w\|^2$ de régularisation, comparer cette fonction d'erreur avec celle obtenue dans le cas du SVM.