Foundations of ML & AI

Theodoros Evgeniou - Nicolas Vayatis

Exercise Set No 3

Exercice 1 (Application of concentration inequality)

Let $\mathcal{H}$ be a class of $\{-1,1\}$-valued functions over $\mathbb{R}^d$. Let $D_n = ((X_1,Y_1),\ldots,(X_n,Y_n))$ an IID sample of classification data in $\mathbb{R}^d \times \{-1,1\}$.

We recall that the empirical Rademacher complexity of $\mathcal{H}$ wrt to the sample $X_1^n = \{X_1,\ldots,X_n\}$ is defined as :

$$\widehat{R}_n(\mathcal{H},X) = \mathbb{E}_\varepsilon\left(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(X_i)\,\middle|\, X_1^n\right) \tag{1}$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are *IID* Rademacher random variables, and they also are independent of D_n.

Set $\delta > 0$.

1. Use McDiarmid's inequality to show that, with probability at least $1-\delta$:

$$\mathbb{E}_X(\widehat{R}_n(\mathcal{H},X)) \leq \widehat{R}_n(\mathcal{H},X) + \sqrt{\frac{2\log(1/\delta)}{n}}$$

Hint : show that $\widehat{R}_n(\mathcal{H},X)$ satisfies the bounded difference condition. The inequality : $\sup(U+V) \leq \sup(U) + \sup(V)$ will be used to this end.

2. Set $\mathcal{F} = \{(x,y) \mapsto \mathbb{I}\{y \neq h(x)\} \ : \ h \in \mathcal{H}\}$. Show how $\widehat{R}_n\big(\mathcal{F},(X,Y)\big)$ and $\widehat{R}_n(\mathcal{H},X)$ are related.
Hint : use the fact that $\mathbb{I}\{y \neq h(x)\} = (1-yh(x))/2$ and also that ε and $-Y\varepsilon$ have the same distribution (why ?).

3. We admit the following bound : with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{(X,Y)}(\mathbb{I}\{Y\neq h(X)\}) - \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i \neq h(X_i)\}\right) \leq$$

$$\leq 2\mathbb{E}_{(X,Y)}(\widehat{R}_n\big(\mathcal{F},(X,Y)\big)) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Then show that the following inequality holds, with probability at least $1-\delta$:

$$\forall h \in \mathcal{H},\ \mathbb{E}_{(X,Y)}(\mathbb{I}\{Y\neq h(X)\}) \leq \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i \neq h(X_i)\} + c_1\widehat{R}_n(\mathcal{H},X) + c_2\sqrt{\frac{\log(2/\delta)}{2n}}$$

where the constants c_1 and c_2 will be provided explicitly.
Hint : use the results of the first two questions together with the union bound.

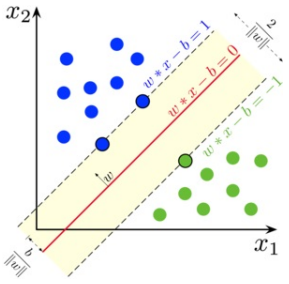
Exercice 2 (Paper discussion)

For at least two of the following papers, provide the following elements : (i) formulation of the objective and the regularization considered (with a justification), (ii) scheme of the proposed algorithm, (iii) theoretical guarantees (if any).

`https://web.stanford.edu/~hastie/Papers/spc_jcgs.pdf`
`https://icml.cc/2012/papers/674.pdf`
`https://www.di.ens.fr/~fbach/STS394.pdf`
`https://papers.nips.cc/paper/3143-multi-task-feature-learning.pdf`
`https://ttic.uchicago.edu/~argyriou/papers/mtl_feat.pdf`
`https://pdfs.semanticscholar.org/bf65/dc3164be43919088695d7f43ee2e51d4b614.pdf`

Exercice 3 (Optimal margin classifiers)

Part A. We consider the binary classification problem in the case of linearly separable data in $\mathbb{R}^d$. The goal is to find a separating hyperplane : $\Delta(w,b) = \{x \ : \ w^T x + b = 0\}$ with $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, which maximizes the "thickness" of the separation between the two classes (-1/+1).



Denote the data sample by $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $Y_i \in \{-1, +1\}$. The optimization problem to solve is the minimization over (w, b) of the following function :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (Y_i (w^T X_i + b) - 1)$$

where $\alpha_1, \ldots, \alpha_n \geq 0$ are the so-called Lagrange multipliers.

1. Compute the gradient of $\mathcal{L}$ wrt w, as well as its derivative wrt b
2. Formuate the dual optimization problem determined by

$$\max_{\alpha \geq 0} \mathcal{L}^*(\alpha) = \max_{\alpha \geq 0} \min_{w,b} \mathcal{L}(w, b, \alpha)$$

   with the constraints that have emerged from the previous question.
3. Denote by $\alpha_1^*, \ldots, \alpha_n^*$ the solutions to the dual optimization problem. What is the form of the equation defining the separating hyperplane w ?

Part B. Now assume that the data are not exactly linearly separable. We allow some of them to be on the 'wrong' side of the frontier and this is monitored thanks to slack variables ξ_i associated to each point (X_i, Y_i) in the sample. This leads to the modified primal Lagrangian formulation as follows :

$$\mathcal{L}(w, b, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (Y_i (w^T X_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

where $\beta_1, \ldots, \beta_n \geq 0$ is an additional set of Lagrange multipliers and C is regularization parameter considered to be fixed (it correspond to the maximal amount of "stiffness" of the hyperplane due to misclassified data).

1. Compute the gradient of $\mathcal{L}$ wrt w, as well as its derivative wrt b
2. Formulate the dual optimization problem determined by

$$\max_{\alpha \geq 0, \beta \geq 0} \mathcal{L}^*(\alpha, \beta) = \max_{\alpha \geq 0, \beta \geq 0} \min_{w,b} \mathcal{L}(w, b, \alpha, \beta)$$

   with the constraints that have emerged from the previous question.
3. Denote by $\alpha_1^*, \ldots, \alpha_n^*$ as well as $\beta_1^*, \ldots, \beta_n^*$ the solutions to the dual optimization problem. What has changed compared to Part A ?
4. Provide a categorization of data points wrt the solution of the problem according to the values of α_i^* (whether it is equal to zero, C or strictly between 0 and C).