Mehryar Mohri
Foundations of Machine Learning 2022
Courant Institute of Mathematical Sciences
Homework assignment 2
March 3, 2022
Due: March 15, 2022 11:55 PM EDT

A VC dimension

1. We denote by $\mathcal{B}_n$ the set of all closed balls in $\mathbb{R}^n$. That is, $\mathcal{B}_n$ is the class of all subsets of the form $\left\{x \in \mathbb{R}^n : \|x - x_0\|^2 \le r^2\right\}$ for some $x_0 \in \mathbb{R}^n$ and $r \ge 0$.

   (a) Show that there exists a set of $n+1$ points in $\mathbb{R}^d$ that can be shattered by $\mathcal{B}_n$. Conclude that $\text{VCdim}(\mathcal{B}_n) \ge n+1$.

   (b) Show that the VC dimension of $\mathcal{B}_n$ is at most equal to the VC dimension of hyperplanes in $\mathbb{R}^{n+1}$. Conclude that $\text{VCdim}(\mathcal{B}_n) \le n+2$.

   (c) Show that $\text{VCdim}(\mathcal{B}_2) = 3$.

B Maximum Margin Multiple Kernel

1. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, \ldots, c\}$ a set of $c \ge 2$ classes. Let $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ be a sample of size m. Assume that $p \ge 1$ positive semi-definite (PSD) base kernels over $\mathcal{X} \times \mathcal{X}$ are given. Consider a hypothesise set based on a kernel $K_{\boldsymbol{\mu}}$ of the form $K_{\boldsymbol{\mu}} = \sum_{k=1}^{p} \mu_k K_k$ where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^\top$ is chosen from $\Delta_q = \left\{\mu : \mu \ge 0, \|\mu\|_q = 1\right\}$ with $q \ge 1$. The multi-class maximum margin multiple kernel (M^{3}K) algorithm [CMR13] is based on the following optimization:

$$
\begin{aligned}
&\min_{\boldsymbol{\mu} \in \widehat{M}_q, \boldsymbol{w}, \boldsymbol{\xi}} \ \frac{1}{2} \sum_{y=1}^{c} \sum_{k=1}^{p} \frac{\|\boldsymbol{w}_{y,k}\|^2}{\mu_k} + C \sum_{i=1}^{m} \xi_i \\
&\text{subject to: } \forall i \in [1, m], \ \xi_i \ge 0, \ \forall y \ne y_i, \\
&\qquad \xi_i \ge 1 - (\boldsymbol{w}_{y_i} \cdot \Phi(x_i) - \boldsymbol{w}_y \cdot \Phi(x_i)),
\end{aligned}
\tag{1}
$$

where $\cdot$ is defined as $\mathcal{A}_{m \times n} \cdot \mathcal{B}_{m \times n} = \sum_{i,j} \mathcal{A}(i,j)\mathcal{B}(i,j)$ for any two matrices $\mathcal{A}$ and $\mathcal{B}$ with the same dimension $m \times n$, $\widehat{M}_q \subset \Delta_q$ is a data-dependent set, $C \ge 0$ is a regularization parameter, $\boldsymbol{w}_y = (\boldsymbol{w}_{y,1}, \ldots, \boldsymbol{w}_{y,p})^\top$ is the associated hypothesis for any class $y \in \mathcal{Y}$, $\Phi(x) = \left(\Phi_{K_1}(x), \ldots, \Phi_{K_p}(x)\right)^\top$ and Φ_K denotes a feature mapping associated to the kernel K.

   (a) Read the Chapter 9.1 - 9.3.1 in the textbook and the paper [CMR13] to understand better the multi-class maximum margin multiple kernel (M^{3}K) algorithm. Write down the explicit expression of $\widehat{M}_q$ and briefly explain each term appearing in that expression.

   (b) Show how to derive the dual optimization of M^{3}K (1):

$$
\begin{aligned}
&\min_{\boldsymbol{\mu} \in \widehat{M}_q} \max_{\boldsymbol{\alpha} \in \mathbb{R}^{m \times c}} \ \sum_{i=1}^{m} \boldsymbol{\alpha}_i \cdot \mathbf{e}_{y_i} - \frac{C}{2} \sum_{i,j=1}^{m} (\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_j) \sum_{k=1}^{p} \mu_k K_k(x_i, x_j) \\
&\text{subject to: } \forall i \in [1, m], \ \boldsymbol{\alpha}_i \le \mathbf{e}_{y_i} \wedge \boldsymbol{\alpha}_i \cdot \mathbf{1} = 0,
\end{aligned}
\tag{2}
$$

   where $\boldsymbol{\alpha} \in \mathbb{R}^{m \times c}$ is a matrix, $\boldsymbol{\alpha}_i$ is its ith row, and $\mathbf{e}_l$ is the lth unit vector in $\mathbb{R}^c$, $l \in [1, c]$. Prove the equivalence of primal (1) and dual (2).

(Note: you should write down every necessary step and rigorously say why the theorems apply.)

C SVMs hand-on

(Note: please share a GitHub link to your open source code in the submission. Any submissions that do not have the code link will obtain a zero point. The graders will check the main lines to ensure that what was done was conceptually correct. The grade will be based on both the code and the answer.)

1. Download and install the `libsvm` software library from:

   `https://www.csie.ntu.edu.tw/~cjlin/libsvm`

   and briefly consult the documentation to become more familiar with the tools.

2. Download the `Abalone` data set:

   `http://archive.ics.uci.edu/ml/datasets/Abalone`

   Use the `libsvm` scaling tool to scale the features of all the data. Use the first 3133 examples for training, the last 1044 for testing. The scaling parameters should be computed only on the training data and then applied to the test data.

3. Consider the binary classification that consists of distinguishing classes 1 through 9 from the rest. Use SVMs combined with polynomial kernels to tackle this binary classification problem.

   To do that, randomly split the training data into five equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 2, 3, 4, 5$, plot the average cross-validation error plus or minus one standard deviation as a function of C (let other parameters of polynomial kernels in `libsvm` be equal to their default values), varying C in powers of 3, starting from a small value $C = 3^{-k}$ to $C = 3^k$, for some value of k. k should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with `libsvm` as the value of C increases.

4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the five-fold cross-validation error and the test errors for the hypotheses obtained as a function of d. Plot the average number of support vectors obtained as a function of d. How many of the support vectors lie on the margin hyperplanes?

5. Fix (C, d) to be (C^*, d^*). Plot the training and test errors as a function of the training sample.

6. Sparse SVM. One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose that instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the L_1 norm of the vector $\boldsymbol{\alpha}$ that defines the weight vector $\boldsymbol{w}$. This gives the following optimization problem for a kernel function K:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} \frac{1}{2} \sum_{i=1}^{m} |\alpha_i| + C \sum_{i=1}^{m} \xi_i \\ &\text{subject to } y_i \left(\sum_{j=1}^{m} \alpha_j y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \right) \geq 1 - \xi_i, i \in [1, m] \\ &\xi_i, \alpha_i \geq 0, i \in [1, m]. \end{aligned} \tag{3}$$

   (a) Derive the equivalent dual optimization problem of (3) in terms of the feature mapping Φ associated to the kernel K and write the proof clearly.

   (b) Derive the equivalent hinge loss minimization problem of (3) and write the proof clearly. Compare it with an instance of the equivalent hinge loss minimization problem of SVM shown in class.

   (c) Apply Stochastic Gradient Descent to solve the optimization problem. Plot the five-fold cross-validation training and test errors for the hypotheses obtained based on the solution $\boldsymbol{\alpha}$ as a function of d, for the best value of C measured on the validation set.

References

[CMR13] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. "Multi-class classification with maximum margin multiple kernel". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 46–54.