Machine Learning: from Theory to Practice
Exam

E. Le Pennec and F. d'Alché-Buc

November 2015

1 Course matter

1. Describe the principle of Kernel PCA as a minimization problem in a RKHS $\mathcal{H}_k$ associated to a PSD kernel k
2. What kind of kernel k would you choose for sequence data ? Propose a definition
3. Describe the methodological steps to achieve to solve a supervised learning problem in general and specifically using a kernel method and a RKHS view of the problem ?
4. Explain how the bootstrap resampling scheme helps to stablilize the tree methods.
5. Explain the underlying principle of the penalization schemes.

2 Theoretical matter

Course

1. Give the representer theorem called "Minimal norm interpolant" in a RKHS $\mathcal{H}_k$ associated to a PSD kernel k
   - Prove this theorem

Leave-one-out error.

The goal of this exercise is to show that the leave-one-out error which is generally very costly to compute can be computed using the training algorithm only once in the case of Kernel Ridge Regression. Let k a PSD kernel and $\mathcal{H}_k$ the associated RKHS. Let $\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ the training sample of size n where $x_i \in \mathcal{X}$, a non empty set and $y \in \mathbb{R}$. Let $\mathcal{S}_i$ denote the sample of size $n-1$ obtained from $\mathcal{S}$ by removing (x_i, y_i). For any training sample $\mathcal{A}$, $h_{\mathcal{A}}$ denotes a function obtained by training on $\mathcal{A}$. By definition, for the squared loss, the leave-one-out error with respect to $\mathcal{S}$ for Kernel Ridge Regression (KRR) is defined as:

$$\hat{R}_{LOO}(KRR) = \frac{1}{n}\sum_{i=1}^{n}(h_{\mathcal{S}_i}(x_i) - y_i)^2, \tag{1}$$

where $h_{\mathcal{S}_i}$ has been obtained by KRR on $\mathcal{S}_i$.

1. Recall the loss function to be minimized in $\mathcal{H}_k$ for Kernel Ridge Regression.

2. Let $\mathcal{S}_i' = \{(x_1, y_1), \ldots, (x_i, h_{\mathcal{S}_i}(x_i)), \ldots, (x_n, y_n)\}$. Show that $h_{\mathcal{S}_i} = h_{\mathcal{S}_i'}$.

3. Define $\mathbf{y}_i = \mathbf{y} - y_i \mathbf{e}_i + h_{\mathcal{S}_i}(x_i)\mathbf{e}_i$, that is the vector of labels with the i^{th} component replaced by $h_{\mathcal{S}_i}(x_i)$ with $\mathbf{e}_i$, i^{th} canonical orthonormal basis vector. Prove that for KRR $h_{\mathcal{S}_i}(x_i) = \mathbf{y}_i^T (K + \lambda\ I)^{-1} K \mathbf{e}_i$. NB: K is the $n \times n$ Gram matrix associated with kernel k for KRR and I is the identity matrix of size $n \times n$. We advise to use a matrix notation with a matrix H to define such as $H_i \mathbf{y}_i$ for the output vector $[h_{\mathcal{S}_i}(x_1)), \ldots, h_{\mathcal{S}_i}(x_n)]$.

4. Prove that the leave-one-out error admits the following simple expression in terms of $h_{\mathcal{S}}$:
$$\hat{R}_{LOO}(KRR) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{h_{\mathcal{S}}(x_i) - y_i}{\mathbf{e}_i^T (K + \lambda\ I)^{-1} K \mathbf{e}_i} \right]^2.$$

3 A Few Useful Things to Know about Machine Learning - P. Domingos

1. Explain how the term *representation*, *evaluation* and *optimization* are related to the following definition of Learning:

   > A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

2. Explain why the *no free lunch* theorem means that we have to make assumptions on our data.

3. Explain why a nearest-neighbor type method is doomed in high dimension if the data are not concentrated in a subspace of low dimension.

4. Do you agree with Domingo's point of view on the theoretical guarantees that can be obtained?

5. Why is that the fact *correlation does not imply causation* is not an issue when doing prediction? How do this becomes a limitation if one wants to do prescription?

4 Practical matter

1. What is overfitting ?

2. Is it always possible to achieve a null training error ?

3. What is a V-fold cross validation scheme and how to use it?

4. In a classification task with SVM, propose some efficient ways to find a *good* subset of variables taken in a very large set.