Machine Learning 2019-2020

Home Assignment 3

Yevgeny Seldin **Christian Igel**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **10 December 2019 22:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your source code in the PDF file.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF.
- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in the speed grader. Zipped pdf submissions will not be graded.
- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- We strongly recommend that your code is written in python 3. If you prefer to use another mainstream language (e.g. R, Matlab, C/C++), you can do so at your own risk and responsibility.
- Your code should include a README text file describing how to compile and run your program, as well as a list of all relevant libraries needed for compiling or using your code.
- Handwritten solutions will not be accepted, please use the provided latex template to write your report.

1 The Role of Independence (5 points)

Design an example of identically distributed, but *dependent* Bernoulli random variables $X_1, \dots, X_n$ (i.e., $X_i \in \{0,1\}$), such that

$$\mathbb{P}\left(\left|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \frac{1}{2}\right) = 1,$$

where $\mu = \mathbb{E}\left[X_i\right]$.

Note that in this case $\frac{1}{n}\sum_{i=1}^{n} X_i$ does not converge to μ as n goes to infinity. The example shows that independence is crucial for convergence of empirical means to the expected values.

2 How to Split a Sample into Training and Test Set (35 points)

In this question you will analyze one possible approach to the question of how to split a dataset S into training and test sets, $S^{\texttt{train}}$ and $S^{\texttt{test}}$. As we have already discussed, overly small test sets lead to unreliable loss estimates, whereas overly large test sets leave too little data for training, thus producing poor prediction models. The optimal trade-off depends on the data and the prediction model. So can we let the data speak for itself? We will give it a try.

1. To warm up: assume that you have a fixed split of S into $S^{\texttt{train}}$ and $S^{\texttt{test}}$, where the size of $S^{\texttt{test}}$ is $n^{\texttt{test}}$. You train a model $\hat{h}^*_{S^{\texttt{train}}}$ on $S^{\texttt{train}}$ using whatever procedure you want. Then you compute the test loss $\hat{L}(\hat{h}^*_{S^{\texttt{train}}}, S^{\texttt{test}})$. Derive a bound on $L(\hat{h}^*_{S^{\texttt{train}}})$ in terms of $\hat{L}(\hat{h}^*_{S^{\texttt{train}}}, S^{\texttt{test}})$ and $n^{\texttt{test}}$ that holds with probability at least $1-\delta$.

2. Now we want to find a good balance between the sizes of $S^{\texttt{train}}$ and $S^{\texttt{test}}$. We consider m possible splits $\{(S^{\texttt{train}}_1, S^{\texttt{test}}_1), \dots, (S^{\texttt{train}}_m, S^{\texttt{test}}_m)\}$, where the sizes of the test sets are $n_1, \dots, n_m$, correspondingly. For example, it could be $(10\%, 90\%), (20\%, 80\%), \dots, (90\%, 10\%)$ splits or anything else with a reasonable coverage of the possible options. We train m prediction models $\hat{h}^*_1, \dots, \hat{h}^*_m$, where $\hat{h}^*_i$ is trained on $S^{\texttt{train}}_i$. We calculate the test loss of the i-th model on the i-th test set $\hat{L}(\hat{h}^*_i, S^{\texttt{test}}_i)$. Derive a bound on $L(\hat{h}^*_i)$ in terms of $\hat{L}(\hat{h}^*_i, S^{\texttt{test}}_i)$ and n_i that holds for all $\hat{h}^*_i$ simultaneously with probability at least $1-\delta$.

   Comment: No theorem from the lecture notes applies directly to this setting, because they all have a fixed sample size n, whereas here the sample sizes vary, $n_1, \dots, n_m$. You have to provide a complete derivation.

3. We expect that most students will treat all the splits in the previous point equally. Note, however, that models trained on more data are a-priori expected to perform better. Propose a way to give them an advantage by using non-uniform treatment [a "prior"] that will give preference to classifiers trained on more samples and repeat the analysis. You have to propose one explicit prior and do the analysis with that prior.

   Be careful! *The "prior" has to be selected before you start working with the data. You cannot pick one prior and after processing the data decide that it was not a good prior and select another one, because this will lead to overfitting.*

4. Curious how this would perform in practice? Wait for the next assignment.

3 Occam's Razor (25 points)

We want to design an application for bilingual users. The application should detect the language in which the person is typing based on the first few letters typed. In other words, we want to design a classifier that takes a short string (that may be less than a full word) as input and predicts one of two languages, say, Danish or English. For simplicity we will assume that the alphabet is restricted to a set Σ of 26 letters of the Latin alphabet plus the white space symbol (so in total $|\Sigma| = 27$). Let Σ^d be the space of strings of length d. Let $\mathcal{H}_d$ be the space of functions from Σ^d to $\{0,1\}$, where Σ^d is the input string and $\{0,1\}$ is the prediction (Danish or English). Let $\mathcal{H} = \bigcup_{d=0}^{\infty} \mathcal{H}_d$ be the union of $\mathcal{H}_d$-s.

1. Derive a high-probability bound[1] for $L(h)$ that holds for all $h \in \mathcal{H}_d$.
2. Derive a high-probability bound for $L(h)$ that holds for all $h \in \mathcal{H}$.
3. Explain the trade-off between picking short strings (small d) and long strings (large d). Which terms in the bound favor small d (i.e., they increase with d) and which terms in the bound favor large d (i.e., they decrease with d)?
4. We have presented a lower bound, where we constructed an example of a problem with a large hypothesis class (of size 2^{2n}), where the empirical loss of the empirically best hypothesis was always zero, but the expected loss of the empirically best hypothesis was at least $1/4$. The hypothesis class $\mathcal{H}$ in this question is obviously infinite. Explain why there is no contradiction between the bound in Point 2 and the lower bound.

[1] A bound that holds with probability at least $1 - \delta$.

Optional, not for submission You are very welcome to experiment with the influence of the string length d on the performance. You can find a lot of texts in different languages at `http://www.gutenberg.org/catalog/`. Do you observe the effect of initial improvement followed by overfitting as you increase d?

4 Kernels (35 points)

The first question should improve the understanding of the geometry of the kernel-induced feature space. You can directly use the result to implement a kernel nearest-neighbor algorithm.

The second question should make you more familiar with the basic definition of the important concept of positive definiteness.

The third question is important to understand the real dimensionality of learning problems using a linear kernel – one reason why linear kernels are often treated differently in efficient implementations.

4.1 Distance in feature space

Given a kernel k on input space $\mathcal{X}$ defining RKHS $\mathcal{H}$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ denote the corresponding feature map (think of $\Phi(x) = k(x, .)$). Let $x, z \in \mathcal{X}$. Show that the distance of $\Phi(x)$ and $\Phi(z)$ in $\mathcal{H}$ is given by

$$\|\Phi(x) - \Phi(z)\| = \sqrt{k(x,x) - 2k(x,z) + k(z,z)}$$

(if distance is measured by the canonical metric induced by k).

4.2 Sum of kernels

Let $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive-definite kernels.

Prove that $k(x,z) = k_1(x,z) + k_2(x,z)$ is also positive-definite.

4.3 Rank of Gram matrix

Let the input space be $\mathcal{X} = \mathbb{R}^d$. Assume a linear kernel, $k(x,z) = x^{\mathrm{T}} z$ for $x, z \in \mathbb{R}^d$ (i.e., the feature map Φ is the identity) and m input patterns $x_1, \ldots, x_m \in \mathbb{R}^d$.

Prove an upper bound on the rank of the Gram matrix from the m input patterns in terms of d and m.