Advanced Topics in Machine Learning 2015-2016

Yevgeny Seldin Christian Igel Brian Brost

Home Assignment 2

Deadline: Sunday, 20 September, 2015, 23:59

The assignments must be answered individually - each student must write and submit his/her own solution. We encourage you to work on the assignments on your own, but we do not prevent you from discussing the questions in small groups. If you do so, you are requested to list your group partners in your individual submission.

Submission format: *Please, upload your answers in a single* `.pdf` *file and additional* `.zip` *file with all the code that you used to solve the assignment. (The* `.pdf` *should* ***not*** *be part of the* `.zip` *file.)*

IMPORTANT: *We are interested in how you solve the problems, not in the final answers. Please, write down all your calculations.*

Question 1 (Tighter generalization bound for consistent hypotheses (33 points))**.**

1. n balls are drawn from a bin with $2n$ balls uniformly at random *without replacement* (meaning that once a ball is taken out it does not return back to the bin). It is known that at least ε-fraction of the balls are red (for $0 < \varepsilon \leq \frac{1}{2}$) and the rest are green. (In other words, there are at least $2n\varepsilon$ red balls and at most $2n(1-\varepsilon)$ green balls.) Show that $\mathbb{P}\{n \text{ green balls are pulled out in a raw}\} \leq e^{-n\varepsilon}$. You can use the inequality $1 + x \leq e^x$.

2. Let $\mathcal{H}$ be an infinite hypothesis class. Prove that for all $h \in \mathcal{H}$ that satisfy $\hat{L}(h,S) = 0$ we have with probability greater than $1 - \delta$:

$$L(h) \leq \underbrace{\hat{L}(h,S)}_{=0} + O\left(\frac{\ln\left(\frac{m_{\mathcal{H}}(2n)}{\delta}\right)}{n}\right) = O\left(\frac{\ln\left(\frac{m_{\mathcal{H}}(2n)}{\delta}\right)}{n}\right).$$

   Please, derive a complete result with all the coefficients, not an $O\left(\frac{\ln\left(\frac{m_{\mathcal{H}}(2n)}{\delta}\right)}{n}\right)$-type bound (meaning that you should compute the constant in front of and inside the logarithm). *(Hint: Traverse the proof of the general result that we did in class and replace Hoeffding's inequality with the result from point 1. where necessary.)*

Remark: If you compare the result with the bound for the general case, $L(h) \leq \hat{L}(h,S) + \sqrt{\frac{8\ln\left(\frac{2m_{\mathcal{H}}(2n)}{\delta}\right)}{n}}$, *you can observe that in the special case when* $\hat{L}(h,S) = 0$ *the bound is much tighter. This is a special case of a more general phenomenon of faster concentration of* $L(h)$ *as* $\hat{L}(h,S)$ *approaches zero. You will observe this phenomenon more closely in Question 4.*

Question 2 (SVM training time in practice (33 points))**.** SVM implementations solve the SVM optimization problem up to a certain accuracy. When using SMO and the stopping condition discussed in the lecture, this accuracy is controlled by the threshold parameter ε (see slides). Let us study the effect of adjusting ε in practice.

Install an SVM solver, for example LIBSVM (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`) or Shark (`http://image.diku.dk/shark`). Download a large and difficult enough benchmark data set (i.e., the training should not be instantaneous and require a significant amount of SMO steps in relation to the number of training data points), split into training and test set. Train a non-linear SVM with different values of ε (in Shark you can use something like `trainer.setMinAccuracy(epsilon);` recall from StatML how to do model selection, i.e., find good values for the regularization and the kernel parameter(s)). Measure the training times and the accuracies on the test set. What do you observe?

Question 3 (Asymmetry of kl-divergence (8 points))**.** For $p, q \in [0,1]$ let $\mathrm{kl}(p\|q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$ be the kl-divergence between two Bernoulli distributions with biases p and q. Prove that kl is asymmetric in its arguments by providing an example of p and q for which $\mathrm{kl}(p\|q) \neq \mathrm{kl}(q\|p)$.

Question 4 (Comparison of Hoeffding's Inequality with kl inequality (26 points))**.** Let $p = \mathbb{P}\{X = 1\}$ be bias of a coin. You observed a sample $X_1, \dots, X_n$ of n coin flips. Let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias of the coin. Plot Hoeffding's upper bound on p and kl upper bound on p as a function of $\hat{p}$ (for $\hat{p} \in [0,1]$) in one figure for $n = 1000$ and confidence parameter $\delta = 0.01$ and compare the two bounds. (Clip the bound at 1, because otherwise it is anyway meaningless.) Repeat the experiment for $n = 5000$ and $n = 10000$ (i.e., plot 3 figures for the 3 different values of n). What conclusions can you make regarding the relative quality of the bounds? Repeat the same for the lower bound on p.

We are giving you a MATLAB function for inverting the kl-divergence with respect to its second argument. The function computes the "upper inverse" $p^+ = \mathrm{kl}^{-1^+}(\hat{p}, \varepsilon) = \max\{p : \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$. The inversion is computed via binary search. You are not obliged to use the function, you can write your own if you like. For the "lower inverse" $p^- = \mathrm{kl}^{-1^-}(\hat{p}, \varepsilon) = \min\{p : \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$ you can either adapt the "upper inverse" function (and we leave it to you to think how to do this) or write your own function. Whatever way you chose you should explain in your main `.pdf` submission file how you computed the upper and the lower bound. The plots should also be part of the main `.pdf` submission. Please, attach all code that you used for solving the assignment in a separate `.zip` file.

Question 5 (The growth function, postponed to Home Assignment 3)**.**

1. Let $\mathcal{H}$ be a finite hypothesis set with $|\mathcal{H}| = M$ hypotheses. Prove that $m_{\mathcal{H}}(n) \leq \min\{M, 2^n\}$. What is the VC-dimension of $\mathcal{H}$?
2. Prove that $m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2$.
3. Prove by induction that
$$\sum_{i=0}^{d} \binom{n}{i} \leq n^d + 1.$$
4. Use the above result to derive a bound on $m_{\mathcal{H}}(n)$.
5. Substitute the result into the VC generalization bound (note that bounding $m_{\mathcal{H}}(2n)$ directly is tighter than going via the result in Point 2). What should be the relation between d and n in order for the bound to be meaningful?

Question 6 (The growth function, not for submission)**.** Calculate $m_{\mathcal{H}}(n)$ in the following cases and compare it with 2^n:

1. $\mathcal{H}$ is the set of positive and negative "rays" on a line (positive rays are described in Example 2.2.1 on page 43 in the handouts and negative rays are the opposite of positive rays).
2. $\mathcal{H}$ is the set of positive and negative intervals on a line (consult Example 2.2.2 in the handouts for the description of positive intervals).
3. $\mathcal{H}$ is the set of separating hyperplanes in $\mathbb{R}^2$. Calculate $m_{\mathcal{H}}(3)$, $m_{\mathcal{H}}(4)$, and $m_{\mathcal{H}}(5)$.

Good luck!
Yevgeny, Christian, & Brian