Advanced Topics in Machine Learning 2015-2016

Yevgeny Seldin Christian Igel Brian Brost

Home Assignment 1

Deadline: Sunday, 13 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 (Hoeffding's Inequality, From Home Assignment 0, 25 points)**.** An airline has collected a sample of 10000 flight reservations and figured out that in this sample 5 percent of passengers who made a reservation on a certain flight did not show up. They introduce a policy to sell 100 tickets for a flight that can hold only 99 passengers. Bound the probability that the number of people that show up for a flight will be larger than the number of seats.

Question 2 (Constrained optimization, 25 points)**.** Use the method of Lagrange multipliers presented in the lecture to solve the following optimization problem. The search space is $M = \mathbb{R}^2$. Find the closest point on the circle with radius 1 around the origin (i.e., $\{(x, y) \,|\, x^2 + y^2 = 1\}$) to the point $(-2, 2)$. Measure the distance by the standard Euclidean distance. Start by writing down the objective function to be minimized together with the single constraint (note that the Lagrangian will be simpler than in the general case).

Question 3 (Occam's Razor, 25 points)**.** We are building a classifier that takes a short text substring as input and predicts whether the following character is a white space. (You can think about text autocompletion as an application.) Let Σ be the set of 26 letters of the Latin alphabet plus the white space symbol (so in total $|\Sigma| = 27$) and 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 whether the next character is a white space. Let $\mathcal{H} = \bigcup_{d=0}^{\infty} \mathcal{H}_d$ be the union of $\mathcal{H}_d$-s.

1. Derive a high-probability bound (a bound that holds with probability $1 - \delta$) on $L(h) - \hat{L}(h, S)$ for learning with $\mathcal{H}_d$. (*Remark: $\mathcal{H}_d$ is finite, so you will get a tighter and simpler result if you work with bounds for finite hypothesis sets rather than with the VC-bound.*)

2. Derive a high-probability bound on $L(h) - \hat{L}(h, S)$ for learning with $\mathcal{H}$.

Question 4 (Learning intervals on a line, 25 points)**.** Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \{0, 1\}$. The target function $f_{a^*b^*}$ for $0 \leq a^* \leq b^* \leq 1$ is defined by $f_{a^*b^*}(x) = \begin{cases} 1, & \text{if } x \in [a^*, b^*] \\ 0, & \text{otherwise} \end{cases}$, where a^* and b^* are unknown. In other words, $f_{a^*b^*}$ labels all points in an unknown interval $[a^*, b^*]$ in $[0, 1]$ by 1 and all other points by 0. The distribution $p(X)$ over $\mathcal{X}$ is also unknown.

We are given a sample $S = (X_1, Y_1), \ldots, (X_N, Y_N)$ sampled i.i.d. according to $p(X)$ and labeled by $f_{a^*b^*}$. In order to learn $f_{a^*b^*}$ from S we use hypothesis set $\mathcal{H} = \{h_{ab} : a, b \in [0, 1] \text{ and } a \leq b\}$, where $h_{ab} = \begin{cases} 1, & \text{if } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$. (Each $h_{ab} \in \mathcal{H}$ corresponds to an interval in $[0, 1]$.) Derive a generalization

bound (a bound on $L(h_{ab})$) that holds with probability $1-\delta$ for all $h_{ab} \in \mathcal{H}$ that are consistent with the sample S. (h_{ab} is consistent with S if $\hat{L}(h_{ab}, S) = 0$.)

Remark: Note that since $f_{a^*b^*} \in \mathcal{H}$ there will always be at least one (but most likely an infinite number) $h_{ab} \in \mathcal{H}$ that is consistent with the sample. A learning algorithm can return any h_{ab} that is consistent with the sample, for example, the minimal interval that contains all positive points in the sample. A meaningful bound on $L(h_{ab})$ for all h_{ab} that are consistent with S implies that it is possible to learn $f_{a^*b^*}$ even though $\mathcal{H}$ is infinite.

Question 5 (Tighter generalization bound for consistent hypotheses - Postponed to Home Assignment 2)**.**

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 will observe that in the special case when $\hat{L}(h,S) = 0$ the bound is much tighter. Later in the course we will show that as $\hat{L}(h,S)$ approaches zero the complexity term gradually decreases from $O\left(\sqrt{\frac{\ln\left(\frac{m_{\mathcal{H}}(2n)}{\delta}\right)}{n}}\right)$ down to $O\left(\frac{\ln\left(\frac{m_{\mathcal{H}}(2n)}{\delta}\right)}{n}\right)$.

Question 6 (The growth function - Postponed to Home Assignment 2)**.**

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 7 (SVM training time in practice - Postponed to Home Assignment 2)**.** 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 8 (SVMs, from Home Assignment 0, not for submission)**.** In Least Squares SVMS (LS-SVMs, Suykens & Vandewalle, 1999) the hinge loss in the 2-norm soft margin SVM is replaced by the squared loss $L(y, y) = (y - y)^2$.

Please recall the Representer Theorem (e.g., see lecture notes Theorem 4.3 on page 39).

1. Is an approach via a Lagrangian really necessary?
2. What are the constraints on the objective variables?

Suykens, J.A.K.; Vandewalle, J. (1999) "Least squares support vector machine classifiers", *Neural Processing Letters*, 9 (3): 293-300.

Question 9 (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