Advanced Topics in Machine Learning 2015-2016

Yevgeny Seldin Christian Igel Brian Brost

Home Assignment 3

Deadline: Sunday, 27 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 (The growth function - 17 points)**.**

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 2 (VC-dimension - 17 points)**.** You do not have to be very formal in this question, some "hand-waving" is allowed. If you formally (mathematically) prove points 1, 5, and 6 you will get some extra points.

1. Let $\mathcal{H}$ be the class of circles in $\mathbb{R}^2$ (each $h \in \mathcal{H}$ is defined by the center of the circle $c \in \mathbb{R}^2$ and its radius r; all points inside the circle are labeled positively and outside negatively). What is the VC-dimension $d_{VC}(\mathcal{H})$?
2. Let $\mathcal{H}$ be the class of all convex sets in $\mathbb{R}^2$, so that points inside the sets are positive and outside negative (see Example 2.2.3 in the `VC dimension.pdf` handout under `Lecture 3`). What is the VC-dimension $d_{VC}(\mathcal{H})$?
3. Provide an example of a learning problem (a distribution on $\mathcal{X} \times \mathcal{Y}$) in which learning with convex sets is likely to lead to severe overfitting (large distance between $\hat{L}(h, S)$ and $L(h)$), whereas learning with circles will provide a reliable estimate of $L(h)$ (small distance between $\hat{L}(h, S)$ and $L(h)$). (Hint: consult the illustration for Example 2.2.3.)

4. Provide an example of a learning problem in which you expect that the distance between $L(h)$ and $\hat{L}(h,S)$ to be similar for learning with circles and learning with convex sets. The goal of this question is to demonstrate that VC-dimension is a worst-case characterization (with respect to the data distribution $p(X,Y)$) of the capacity of a hypothesis class. For benign data distributions the reality might be better than the VC bound.

5. What is the VC-dimension of the class of homogeneous separating hyperplanes in $\mathbb{R}^2$? (A homogeneous hyperplane is the one that passes through the origin.)

6. What is the VC-dimension of the class of general separating hyperplanes in $\mathbb{R}^2$?

7. Let $\mathcal{H}$ be a hypotheses class with $d_{VC}(\mathcal{H}) = 10$. What should be the sample size n (according to generalization bound) so that $L(h) \leq \hat{L}(h,S) + 0.05$ for all $h \in \mathcal{H}$ with probability at least 95%?

Question 3 (Margin-based VC bound for SVMs - 17 points)**.** The margin-based generalization bound for SVMs that we derived in class (and in the lecture notes) assumed that the data lies in the unit ball ($\|\mathbf{x}\| \leq 1$ for all $\mathbf{x} \in \mathcal{X}$). Derive a generalization bound for the case where data lies in a ball of radius R ($\|\mathbf{x}\| \leq R$ for all $\mathbf{x} \in \mathcal{X}$).

Question 4 (Occam's razor bound and the lower VC bound - 17 points)**.** Recall the hypothesis classes from Question 3 on Occam's razor in Home Assignment 1.

1. What is the VC-dimension of $\mathcal{H}_d$?

2. What is the VC-dimension of $\mathcal{H}$?

3. In Question 3.2 in Home Assignment 1 we derived a high-probability bound on $L(h) - \hat{L}(h,S)$ for learning with $\mathcal{H}$. Why there is no contradiction between that result, the VC lower bound, and point 2. of this question?

Question 5 (Occam's razor bound with kl inequality and PAC-Bayesian bound - 17 points)**.**

1. Use kl inequality in order to derive a tighter Occam's razor bound.

2. Use the above result and Jensen's inequality in order to show that for any distribution π that is independent of the sample, with probability at least $1-\delta$ for all distributions ρ:

$$\mathrm{kl}(\hat{L}(\rho,S) \| L(\rho)) \leq \frac{\ln\left(\frac{n+1}{\delta}\right) + \sum_{h\in\mathcal{H}} \rho(h) \ln \frac{1}{\pi(h)}}{n}.$$

Remark: Compare the result with PAC-Bayesian bound and observe that $\mathrm{KL}(\rho\|\pi) = \sum_{h\in\mathcal{H}} \rho(h) \ln \frac{\rho(h)}{\pi(h)} = \sum_{h\in\mathcal{H}} \rho(h) \ln \frac{1}{\pi(h)} - \mathrm{H}(\rho) \leq \sum_{h\in\mathcal{H}} \rho(h) \ln \frac{1}{\pi(h)}$*, since the entropy* $\mathrm{H}(\rho) \geq 0$*. Thus, the change of measure inequality that is at the basis of PAC-Bayesian inequality is more powerful that the union bound that is at the basis of Occam's razor bound.*

Question 6 (Nonnegativity of KL-divergence - 15 points)**.** In the handouts you have a couple of proofs of nonnegativity of KL-divergence. In this question you will derive yet another proof of this fact.

1. Prove that $\ln x \leq x - 1$ for all $0 < x < \infty$.

2. Use the above inequality to prove that $\mathrm{KL}(p\|q) \geq 0$. You can assume that p and q are discrete distributions. (Hint: show that $-\mathrm{KL}(p\|q) \leq 0$. Note that you should separately treat the case when there is a point x for which $q(x) = 0$ and $p(x) > 0$ and you should also treat x-es for which $p(x) = 0$ separately.)

3. What is the condition for equality $\mathrm{KL}(p\|q) = 0$ in your proof?

Good luck!
Yevgeny, Christian, & Brian