Machine Learning 2017/2018

Home Assignment 5

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

The deadline for this assignment is **2 January 2018**. 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 (Matlab / R / Python scripts / C / C++ / Java / etc.) 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.
- 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.

1 The growth function

1. Let $\mathcal{H}$ be a finite hypothesis set with $|\mathcal{H}| = M$ hypotheses. Prove that $m_{\mathcal{H}}(n) \leq \min\{M, 2^n\}$.

2. Bound the VC-dimension of $\mathcal{H}$ above (i.e., $|\mathcal{H}| = M$).

3. Prove that $m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2$.

4. You can use Point 1.4 in Home Assignment 4 to bound $m_{\mathcal{H}}(n)$, as well as $m_{\mathcal{H}}(2n)$. Show that for $n \geq 2$ bounding $m_{\mathcal{H}}(2n)$ directly leads to a tighter result than bounding it via $m_{\mathcal{H}}(n)^2$ using Point 3.

2 VC-dimension

1. Let $\mathcal{H}_+$ be the class of "positive" 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 $r \in \mathbb{R}$; all points inside the circle are labeled positively and outside negatively). Prove that $d_{VC}(\mathcal{H}_+) \geq 3$.

2. Let $\mathcal{H} = \mathcal{H}_+ \cup \mathcal{H}_-$ be the class of "positive" and "negative" circles in $\mathbb{R}^2$ (the "negative" circles are negative inside and positive outside). Prove that $d_{VC}(\mathcal{H}) \geq 4$.

3. What kind of statement would you have to prove in order to show that $d_{VC}(\mathcal{H}_+) \leq 3$? What kind of statement would you have to prove in order to show that $d_{VC}(\mathcal{H}) \leq 4$?

4* ***Optional*** Prove that $d_{VC}(\mathcal{H}_+) \leq 3$ and $d_{VC}(\mathcal{H}) \leq 4$. [Doing this formally is hard, but it will earn you extra honor.]

3 Generalization Bound for Learning Gaussian RBF Kernel Bandwidth

This question was given in the re-exam in fall 2016.

Gaussian RBF Kernels (where RBF stands for Radial Basis Function) are extremely popular in machine learning (Abu-Mostafa et al., 2012, Online Chapter 8, page 37). The kernel is defined by

$$k(\mathbf{x}, \mathbf{x}') = e^{-\gamma \|\mathbf{x} - \mathbf{x}'\|^2}$$

and has a free parameter γ. In practice, it is common to tune γ by (cross-) validation. In this question we study the effect that such tuning has on generalization properties of the classifier.

Every answer in this question must be supported either by citation or by derivation (or both). You are allowed to cite our lecture notes, slides, and the course

book (Abu-Mostafa et al., 2012). If you would like to cite other sources it is advisable to double-check with us their trustworthiness.

In this question we use the following terminology. A *high-probability bound* is a bound that holds with high probability for all hypotheses h in the relevant hypothesis space. We call a bound *trivial* if the statement can be obtained without making any calculations (for example, if a bound states that the probability of some event is bounded by 2, then it is a trivial bound, because we know without any calculations that probabilities are always bounded by 1).

The subpoints of this question progressively build on their predecessors. However, the question is built in such a way that even if you get stuck at one point you may still be able to proceed with subsequent points, so if it happens to you do not give up too early.

1. Assume that $\mathbf{x} = x \in \mathbb{R}^1$, i.e., it is one-dimensional. In this case Gaussian RBF kernel corresponds to a mapping

$$\Phi_\gamma(x) = e^{-\gamma x^2}\left(1, \sqrt{\frac{(2\gamma)^1}{1!}}x, \sqrt{\frac{(2\gamma)^2}{2!}}x^2, \sqrt{\frac{(2\gamma)^3}{3!}}x^3, \dots\right)$$

   (Abu-Mostafa et al., 2012, Online Chapter 8, page 37). Let $\mathcal{H}_\gamma$ be the hypothesis space of linear separators in the space defined by Φ_γ. What do you think is the VC-dimension of $\mathcal{H}_\gamma$ and why do you think so? What kind of statement would you have to prove in order to justify your result? Optionally: provide a formal proof.

2. Derive a bound on the norm of the feature mapping, $\|\Phi_\gamma(x)\|$.

3. Let $\mathbf{w}$ be a vector defining a homogeneous hyperplane in the space defined by Φ_γ. Let $h = \mathbf{w}$ denote the corresponding hypothesis. Derive a high-probability margin-based bound on $L(\mathbf{w})$ that depends on the norm of the vector $\mathbf{w}$. You should point out where in the derivation you are using the result of Point 2.

   Put attention that Φ_γ includes a leading coordinate "1". Therefore, the bias term can be absorbed into the first coordinate of $\mathbf{w}$ (the vector that defines a separating hyperplane) and we are talking about linear separation by *homogeneous* hyperplanes. Homogeneous hyperplanes are hyperplanes that are passing through the origin. A homogeneous hyperplane is defined by equation $\mathbf{w}^T\Phi_\gamma(x) = 0$ (without the bias term b). When working with homogeneous hyperplanes you should drop the "+1" term in Theorem 3.8 in Yevgeny's lecture notes, which comes from the bias term. I.e., you should replace $\mathcal{H}_\rho$ in Theorem 3.8 with $\mathcal{H}_\rho = \left\{\mathbf{w} : \|\mathbf{w}\| \leq \frac{1}{\rho}\right\}$ and the bound should be $d_{VC}(\mathcal{H}_\rho) \leq \lceil R^2/\rho^2 \rceil$.

4. Assume that you would like to work with K different bandwidth parameters $\gamma_1, \ldots, \gamma_K$. Let $\mathcal{H} = \bigcup_{i=1}^{K} \mathcal{H}_{\gamma_i}$. Derive a margin-based bound for learning with $\mathcal{H}$.

5. There are several ways of selecting γ. One way is to select γ that minimizes the generalization bound you have derived in the previous point. Another way is to set aside a validation set S_{val} and select γ that minimizes the validation error $\hat{L}(\mathbf{w}, S_{\text{val}})$. Assume that you have K different hypotheses $\mathbf{w}_1, \ldots, \mathbf{w}_K$ corresponding to K different values of γ and you pick the one that minimizes the validation error. Lets denote it by $\mathbf{w}^*$ and lets assume that the size of the validation set S_{val} is m. Derive a generalization bound for $\mathbf{w}^*$ in terms of its validation error.

6. In practice, when running soft-margin SVMs (Abu-Mostafa et al., 2012, Online Chapter 8, page 40) it is common to tune both the kernel bandwidth parameter γ and the regularization parameter C (Abu-Mostafa et al., 2012, Equation (8.30)). A typical approach is to take a grid of K values of γ and M values of C. Consider two ways of tuning C:

   (a) We pick the hypothesis $\mathbf{w}^*$ that minimizes a generalization bound based on the training error (similar to the one in Point 4).

   (b) We pick the hypothesis $\mathbf{w}^*$ that minimizes a generalization bound based on the validation error (similar to the one in Point 5).

   In which case(s) do we need to modify the bound, so that it will take into account selection of C and in which case(s) we don't? In other words: if you would be required to rederive the bounds in Points 4 and 5 under the assumption that you are tuning both γ and C, what bound(s) will remain the same and what bound(s) will have an additional factor M appearing somewhere inside the bound?

   Please, explain your answer. Simple yes/no answers will not be accepted.

4 Random Forests

4.1 Normalization

As discussed, normalizing each component to zero mean and variance one (measured on the training set) is a common preprocessing step, which can remove undesired biases due to different scaling. Using this normalization affects different classification methods differently.

- Is nearest neighbor classification affected by this type of normalization? If your answer is yes, give an example. If your answer is no, provide convincing arguments why not.
- Is random forrest classification affected by this normalization? If your answer is yes, give an example. If your answer is no, provide convincing arguments why not.

4.2 Random forests in practice (not for delivery)

Install – if necessary – and apply software implementing random forests. For example, both R and Python provide good random forest implementations. In Python you can use `RandomForestClassifier` and `RandomForestRegressor`. To our knowledge, the fastest random forest classifier is provided by the most recent version (i.e., you have to download from the repository) of the Shark machine learning library (Igel et al., 2008). Still lacks a lot of features, but is fast as a shark.

Apply random forests to the example problems from the previous assignments. Play with the hyperparameters.

References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning from data*. AML-book, 2012.

C. Igel, T. Glasmachers, and V. Heidrich-Meisner. Shark. *Journal of Machine Learning Research*, 9:993–996, 2008.