Advanced Topics in Machine Learning 2017-2018

Yevgeny Seldin   Christian Igel   Tobias Sommer Thune   Julian Zimmert

Home Assignment 3

Deadline: Tuesday, 26 September 2017, 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 the 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 the calculations and comment your solutions.*

1 SVM and regularization (15 points)

Consider the primal optimization probelm of the 1-norm soft margin SVM:

$$\begin{aligned} \text{minimize}_{\boldsymbol{\xi},\boldsymbol{w},b} \quad & \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{\ell}\xi_i \\ \text{subject to} \quad & y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle + b) \geq 1 - \xi_i \ , \ \ i = 1,\dots,\ell \\ & \xi_i \geq 0 \ , \ \ i = 1,\dots,\ell \end{aligned}$$

Proof that for fixed $\boldsymbol{w}$ the optimal slack variables are given by

$$\xi_i = \max(0, 1 - y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i)\rangle + b)) = L_{\text{hinge}}(y_i, f(x_i)) \ .$$

2 Karush-Kuhn-Tucker (KKT) theorem (15 points)

This exercise is about gaining a better understanding of the KKT theorem. It is easy to see directly the solutions. However, you should still formally check all condition of the KKT theorem – that is the point of this exercise.

Consider the three optimization problems:

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{w}\| \qquad \boldsymbol{w} \in \mathbb{R}^2 \\ \text{subject to} \quad & w_1 + w_2 + 1 \geq 0 \end{aligned} \tag{I}$$

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{w}\| \qquad \boldsymbol{w} \in \mathbb{R}^2 \\ \text{subject to} \quad & w_1 + w_2 + 1 \leq 0 \end{aligned} \tag{II}$$

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{w}\| \qquad \boldsymbol{w} \in \mathbb{R}^2 \\ \text{subject to} \quad & 2w_1 + 2w_2 + 2 = 0 \end{aligned} \tag{III}$$

Give the solutions to all three problem. Verify that the solutions are optimal by checking all conditions of the KKT theorem as given in the lecture. Document theses checks.

3 SVM model selection (15 points)

Consider a soft-margin SVM with radial Gaussian kernel. There are typically two hyperparameters, one called γ for the kernel bandwidth (or σ^2 parameterizing the "variance" of the kernel) and on called C for the regularization trade-off *as defined in the lecture.*

The normal way to adjust γ and C is by grid-search. One varies $\gamma \in \{\gamma_1, \ldots, \gamma_l\}$ and $C \in \{C_1, \ldots, C_m\}$. For each combination of C and γ, the performance of the model is estimated using cross-validation.

If you vary C for a given γ in a particular order, you can add a termination criterion that allows you to skip some of the C values without changing the end result. This can save a lot of computation time depending on the grid.

In which order should you vary C? What is the stopping condition? Why is it guaranteed that the solution does not change even if you skip certain values of C for the given γ?

4 PAC-Bayesian Aggregation (55 points)

In this question you are asked to reproduce an experiment from Thiemann et al. (2017, Section 6, Figure 2). We quote the relevant part of the description of the experiment from the paper below. "Our prediction strategy" refers to alternating minimization of the bound in Theorem 3.18 in Yevgeny's lecture notes, which is equivalent to the bound in Theorem 6 in the paper. You are only required to reproduce the experiment for the first dataset, Ionosphere, which you can download from the UCI repository (Asuncion and Newman, 2007). You are allowed to use any programming language you like and any SVM solver you chose. Please, document carefully what you do and clearly annotate your graphs, including legend and axis labels. Some additional hints and comments are provided after the quote.

We compared the prediction accuracy and run time of our prediction strategy with a baseline of RBF kernel SVMs tuned by cross-validation. For the baseline we used 5-fold cross-validation for selecting the soft-margin parameter, C, and the bandwidth parameter γ of the kernel $k(X_i, X_j) = \exp(-\gamma\|X_i - X_j\|^2)$. The value of C was selected from a grid, such that $\log_{10} C \in \{-3, -2, \ldots, 3\}$. The values for the grid of γ-s were selected using the heuristic proposed in Jaakkola et al. (1999). Specifically, for $i \in \{1, \ldots, n\}$ we defined $G(X_i) = \min_{(X_j,Y_j)\in S \wedge Y_i \neq Y_j} \|X_i - X_j\|$. We then defined a seed γ_J by $\gamma_J = \frac{1}{2\cdot median(G)^2}$. Finally, we took a geometrically spaced grid around γ_J, so that $\gamma \in \{\gamma_J 10^{-4}, \gamma_J 10^{-2}, \ldots, \gamma_J 10^{4}\}$.

For our approach we selected m subsets of r points uniformly at random from the training set S. We then trained an RBF kernel SVM for each subset. The kernel bandwidth parameter γ was randomly selected for each subset from the same grid as used in the baseline. In all our experiments very small values of r, typically up to $d + 1$ with d being the input dimension, were sufficient for successfully competing with the prediction accuracy of the baseline and provided the most significant computational improvement. For such small values of r it was easy to achieve perfect separation of the training points and, therefore, selection of C was unnecessary. The performance of each weak classifier was validated on $n - r$ points not used in its training. The weighting of classifiers ρ was then computed through alternating minimization of the bound in Theorem 6.

In most of PAC-Bayesian literature it is common to replace randomized prediction with ρ-weighted majority vote. From a theoretical point of view the error of ρ-weighted majority vote is bounded by at most twice the error of the corresponding randomized classifier, however, in practice it usually performs better than randomized prediction (Germain et al., 2009). In our main experiments we have followed the standard choice of using the ρ-weighted majority vote.

In the second experiment [the one you are required to reproduce] we provide a closer look at the effect of increasing the number m of weak SVMs when their training set sizes r are kept fixed. We picked $r = d + 1$ and ran our training procedure with 20 values of m in $[1, n]$. In Figure 2 we present the prediction accuracy of the resulting weighted majority vote vs. prediction accuracy of the baseline for four datasets. We also show the running time of our procedure vs. the baseline. The running time of the baseline includes cross-validation and training of the final SVM on the whole training set, while the running time of PAC-Bayesian aggregation includes training of m weak SVMs, their validation, and the computation of ρ. In addition, we report the value of PAC-Bayes-kl bound from Theorem 2 on the expected loss of the randomized classifier defined by ρ. The kl divergence was inverted numerically to obtain a bound on the expected loss $\mathbb{E}_\rho [L(h)]$. The bound was adapted to our construction by replacing n *with $n - r$ and $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ with $\mathbb{E}_\rho\left[\hat{L}^{\mathrm{val}}(h,S)\right]$. We note that since the bound holds for any posterior distribution, it also holds for the distribution found by minimization of the bound in Theorem 6. However, since Theorem 6 is a relaxation of PAC-Bayes-kl bound, using PAC-Bayes-kl for the final error estimate is slightly tighter. The bound on the loss of ρ-weighted majority vote is at most a factor of 2 larger than than the bound for the randomized classifier. In calculation of the bound we used $\delta = 0.05$. We conclude from the figure that relatively small values of m are sufficient for matching or almost matching the prediction accuracy of the baseline, while the run time is reduced dramatically. We also note that the bound is exceptionally tight.*

Additional hints and comments

Hint: direct computation of the update rule for ρ,

$$\rho(h) = \frac{\pi(h)e^{-\lambda(n-r)\hat{L}^{\mathrm{val}}(h,S)}}{\sum_{h'}\pi(h')e^{-\lambda(n-r)\hat{L}^{\mathrm{val}}(h',S)}},$$

is numerically unstable, since for large $n-r$ it leads to division of zero by zero. A way to fix the problem is to normalize by $e^{-\lambda(n-r)\hat{L}^{\mathrm{val}}_{\min}}$, where $\hat{L}^{\mathrm{val}}_{\min} = \min_{h''}\hat{L}^{\mathrm{val}}(h'',S)$. This leads to

$$\rho(h) = \frac{\pi(h)e^{-\lambda(n-r)\hat{L}^{\mathrm{val}}(h,S)}}{\sum_{h'}\pi(h')e^{-\lambda(n-r)\hat{L}^{\mathrm{val}}(h',S)}} = \frac{\pi(h)e^{-\lambda(n-r)\left(\hat{L}^{\mathrm{val}(h,S)}-\hat{L}^{\mathrm{val}}_{\min}\right)}}{\sum_{h'}\pi(h')e^{-\lambda(n-r)\left(\hat{L}^{\mathrm{val}}(h',S)-\hat{L}^{\mathrm{val}}_{\min}\right)}}.$$

Calculation of the latter expression for $\rho(h)$ does not lead to numerical instability problems.

Comments:

1. Theorem 6 in the paper uses a slightly tighter version of PAC-Bayes-kl inequality compared to Theorem 3.18 in the lecture notes. Specifically, the $\ln(n+1)$ term is replaced by $\ln(2\sqrt{n})$, which leads to $\ln((n-r)+1)$ in Theorem 3.18 to be replaced by $\ln(2\sqrt{n-r})$ in Theorem 6 in the paper. We do not mind which one you select to work with, but remember to document it in your report.

2. Ideally, you would repeat the experiment several times, say 10, and report the average + some form of deviation, e.g. standard deviation or quantiles, over the repetitions. We made a sin by not doing it in the paper and you are allowed to repeat the sin in order to save time. Please, do not do that in real papers.

Good luck!
Yevgeny, Christian, Tobias & Julian

References

Arthur Asuncion and David J. Newman. UCI machine learning repository, 2007. `http://archive.ics.uci.edu/ml/index.php`.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

Tommi Jaakkola, Mark Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. In *In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 1999.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2017. `http://arxiv.org/abs/1608.05610`.