Advanced Topics in Machine Learning 2017-2018

Yevgeny Seldin Christian Igel Tobias Sommer Thune Julian Zimmert

Home Assignment 4

Deadline: Tuesday, 3 October 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 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 $\gamma \in \{\gamma_1, \dots, \gamma_l\}$ and $C \in \{C_1, \dots, C_m\}$. 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?

2 Least Squares SVMs (15 points)

In Least Squares SVMS (LS-SVMs, Suykens & Vandewalle, 1999) the hinge loss in the L_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.

3 Regularization by relative entropy and the Gibbs distribution (15 points)

In this question we will show that regularization by relative entropy leads to solutions in a form of the Gibbs distribution. Lets assume that we have a finite hypothesis class $\mathcal{H}$ of size m and we want to minimize

$$\mathcal{F}(\rho) = \alpha \mathbb{E}_\rho \left[\hat{L}(h,S)\right] + \mathrm{KL}(\rho\|\pi) = \alpha \sum_{h=1}^m \rho(h)\hat{L}(h,S) + \sum_{h=1}^m \rho(h) \ln \frac{\rho(h)}{\pi(h)}$$

with respect to the distribution ρ. This objective is closely related to the objective of PAC-Bayes-λ inequality when λ is fixed and this sort of minimization problem appears in many other places in machine learning. Lets slightly simplify and formalize the problem. Let $\rho = (\rho_1, \dots, \rho_m)$ be the posterior distribution, $\pi = (\pi_1, \dots, \pi_m)$ the prior distribution, and $L = (L_1, \dots, L_m)$ the vector of losses. You should solve

$$\begin{aligned} \min_{\rho_1,\dots,\rho_m} \quad & \alpha \sum_{h=1}^m \rho_h L_h + \sum_{h=1}^m \rho_h \ln \frac{\rho_h}{\pi_h} \\ s.t. \quad & \sum_{h=1}^m \rho_h = 1 \\ & \forall h : \rho_h \geq 0 \end{aligned} \tag{1}$$

and show that the solution is $\rho_h = \frac{\pi_h e^{-\alpha L_h}}{\sum_{h'=1}^m \pi_{h'} e^{-\alpha L_{h'}}}$. Distribution of this form is known as the Gibbs distribution.

Guidelines:

1. Instead of solving minimization problem (1), solve the following minimization problem

$$\begin{aligned} \min_{\rho_1,\dots,\rho_m} \quad & \alpha \sum_{h=1}^m \rho_h L_h + \sum_{h=1}^m \rho_h \ln \frac{\rho_h}{\pi_h} \\ s.t. \quad & \sum_{h=1}^m \rho_h = 1, \end{aligned} \tag{2}$$

   i.e., drop the last constraint in (1).

2. Use the method of Lagrange multipliers to show that the solution of the above problem has a form of $\rho_h = \pi_h e^{-\alpha L_h + \texttt{something}}$, where `something` is something involving the Lagrange multiplier.

3. Show that $\rho_h \geq 0$ for all h. (This is trivial. But it gives us that the solutions of (1) and (2) are identical.)

4. Finally, $e^{\texttt{something}}$ should be such that the constraint $\sum_{h=1}^m \rho_h = 1$ is satisfied. So you can easily get the solution. You even do not have to compute the Lagrange multiplier explicitly.

4 Follow The Leader (FTL) algorithm for i.i.d. full information games (30 points)

Follow the leader (FTL) is a playing strategy that on round t plays the action that was most successful up to round t ("the leader"). Derive a bound on pseudo regret of FTL in i.i.d. full information games with K possible actions and outcomes bounded in the $[0,1]$ interval (you can work with rewards or losses, as you like). You can use the following guidelines:

1. You are allowed to solve the problem for $K = 2$. (The guidelines are not limited to $K = 2$.)

2. Let $\mu(a)$ be expected reward of action a and let $\hat{\mu}_t(a)$ be empirical estimate of the reward of action a on round t (the average of rewards observed so far). Let a^* be an optimal action (there may be more than one optimal action, but then things only get better [convince yourself that this is true], so we can assume that there is a single a^*). Let $\Delta(a) = \mu(a^*) - \mu(a)$. FTL plays $a \neq a^*$ on rounds t for which $\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)$. So let us analyze how often this may happen.

3. Construct Hoeffding's confidence bounds on the deviation of $\hat{\mu}_t(a)$ from $\mu(a)$. Use the following form of the inequality $\mathbb{P}\left\{\hat{\mu}_t(a) - \mu(a) \geq \sqrt{\frac{\ln \frac{1}{\delta_t}}{2t}}\right\} \leq \delta_t$ (check yourself which side you need for which action).

4. Note that when the width of confidence intervals for a and a^* is smaller than $\frac{1}{2}\Delta(a)$ we can distinguish between a and a^* unless one of the confidence intervals fails, which happens with probability at most δ_t.

5. Note that δ_t is a free parameter that we can set any way we like. So lets set it in such a way that the width of the confidence interval will be $\frac{1}{2}\Delta(a)$.

6. Note that δ_t cannot be larger than 1, so you should get an initial period in the game, where confidence intervals are not under control. What is the length of this period?

7. Now if we put everything together - we have an initial period during which we have no control over the estimates and after the initial period we can get a bound on the expected number of rounds when confidence intervals fail. Combine the two results to get a bound on the expected number of times $\mathbb{E}[N_T(a)]$ that a is played and use it to get a bound on $\bar{R}_T$. Overall, you should get a bound of a form $\bar{R}_T \leq \sum_{a:\Delta(a)>0}\left(\frac{c_1}{\Delta(a)} + \frac{c_2}{1-\exp(\Delta(a)^2/2)}\Delta(a)\right)$, where c_1 and c_2 are constants. *Note that in the full information i.i.d. setting the regret does not grow with time!!!* (Since the bound is independent of T.)

   Hint: at some point in the proof you will need to sum up a geometric series. A geometric series is a series of a form $\sum_{t=0}^{\infty} r^t$ and for $r < 1$ we have $\sum_{t=0}^{\infty} r^t = \frac{1}{1-r}$. In your case r will, actually, be an exponent $r = e^c$ for some constant c.

5 Decoupling exploration and exploitation in i.i.d. multiarmed bandits (25 points)

Assume an i.i.d. multiarmed bandit game, where the observations are not coupled to the actions. Specifically, we assume that on each round of the game the player is allowed to observe the reward of a single arm, but it does not have to be the same arm that was played on that round (and if it's not the same arm, the player does not observe his own reward, but he observes the reward of an alternative option).

Derive a playing strategy and a regret bound for this game. (You should solve this problem for a general K and you should get that the regret does not grow with time.)

Remark: note that in this setting the exploration is "free", because we do not have to play suboptimal actions in order to test their quality. And if we contrast this with the standard multiarmed bandit setting we observe that the regret stops growing with time instead of growing logarithmically with time. Actually, the result that you should get is much closer to the regret bound in Question 1 than to the regret bound for multiarmed bandits. Thus, it is not the fact that we have just a single observation that makes i.i.d. multiarmed bandits harder than full information games, but the fact that this single observation is linked to the action. (In adversarial multiarmed bandits the effect of decoupling is more involved (Avner et al., 2012, Seldin et al., 2014).)

Good luck!
Yevgeny, Christian, Tobias & Julian

References

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.