Advanced Topics in Machine Learning 2017-2018

Yevgeny Seldin Christian Igel Tobias Sommer Thune Julian Zimmert

Home Assignment 1

Deadline: Tuesday, 19 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.*

Assignment structure: The assignment contains 4 mandatory questions and 2 optional questions. All questions are designed to help you understand the material, however, the optional questions are not for submission and you can get help with the optional questions at the TA session if needed.

1 Subgradients (10 points)

This exercise is about gaining a better understanding of the concept of subgradients.

1. Construct a function $f : \mathbb{R} \to \mathbb{R}$ with differential set $[-2, 2]$ at 0.

2. Prove that the function has the desired property.

2 Convexity of logistic regression (10 points)

Logistic regression is a fundamental, widely used machine learning method (Abu-Mostafa et al., 2012, Shalev-Shwartz and Ben-David, 2014).

Let's prove that the objective function used in logistic regression is convex. This implies that a local optimum of the logistic regression objective function is also a global one and – as we will see later in this course – that optimizing the logistic regression objective function via gradient descent converges to a global optimal solution.

In this exercise, you are supposed to "flesh out" the proof presented by Shalev-Shwartz and Ben-David (2014). Remember from your machine learning lecture that the error (objective) function for logistic regression can be written as

$$\frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + e^{-y_n \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n}\right) \tag{1}$$

for training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} \subseteq (\mathbb{R}^n \times \{-1, 1\})^N$ and model parameters $\boldsymbol{w} \in \mathbb{R}^n$.

1. Prove that the derivative of
$$f(x) = \ln(1 + \exp(x))$$
is
$$f'(x) = \frac{1}{1 + \exp(-x)} \quad .$$

2. Prove that the second derivative $f''(x)$ is positive (i.e., $f'(x)$ is a monotonically increasing function). This shows that $f(x)$ is convex (the "Hessian matrix", a scalar in this case, is "positive definite").

These two statements show that the function defined by Eq. (1) is the sum of functions that are compositions of convex functions with linear functions – and hence, as we remember from last week's assignment, Eq. (1) is a convex function.

3 PAC-Bayes vs. Occam (20 points)

The change of measure inequality that is at the basis of PAC-Bayesian analysis can be seen as a replacement of the union bound, which is at the basis of Occam's razor. In this question we compare the tightness of the two approaches.

1. Prove the following theorem.

   Theorem 1. *Let S be an i.i.d. sample of n points, let ℓ be the zero-one loss, let $\mathcal{H}$ be countable, and let $\pi(h)$ be such that it is independent of the sample S and satisfies $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$. Let $\delta \in (0,1)$. Then with probability greater than $1-\delta$, for all distributions ρ over $\mathcal{H}$ simultaneously:*

   $$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \middle\| \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\sum_{h\in\mathcal{H}} \rho(h) \ln \frac{1}{\pi(h)} + \ln \frac{n+1}{\delta}}{n}. \tag{2}$$

   You can use the result from Home Assignment 1, where you have shown that under the conditions of Theorem 1, with probability greater than $1-\delta$ for all $h \in \mathcal{H}$ simultaneously

   $$\mathrm{kl}(\hat{L}(h,S)\|L(h)) \leq \frac{\ln \frac{n+1}{\pi(h)\delta}}{n}.$$

2. Recall that by PAC-Bayes-kl inequality, under the conditions of Theorem 1 we have that with probability greater than $1-\delta$, for all distributions ρ over $\mathcal{H}$ simultaneously:

   $$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \middle\| \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{n+1}{\delta}}{n}. \tag{3}$$

   Show that the PAC-Bayes-kl inequality (3) is always at least as tight as the Occam's razor bound with kl in equation (2). Hint: the entropy is always non-negative, $\mathrm{H}(\rho) \geq 0$.

4 Nonnegativity of KL (20 points)

In the handouts from Cover and Thomas (2006) you have a couple of proofs of nonnegativity of the 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?

5 Early stopping and overfitting (40 points)

Early stopping is a widely used technique to avoid overfitting in models trained by iterative methods, such as gradient descent. For example, it is used to avoid overfitting in training neural networks. A possible way of implementing early stopping is to set aside a validation set S^{val} and return the model that minimizes the validation loss (for example, see `https://www.quora.com/How-does-one-employ-early-stopping-in-TensorFlow`).

Let $h_1, h_2, h_3, \dots$ be a sequence of models obtained after $1, 2, 3, \dots$ epochs of training a neural network (you do not have to know the details of the training procedure to answer the question). Let $\hat{L}(h_1, S^{\text{val}})$, $\hat{L}(h_2, S^{\text{val}})$, $\hat{L}(h_3, S^{\text{val}}), \dots$ be the corresponding sequence of validation errors on a fixed validation set S^{val}.

1. Let h_{t^*} be the neural network returned after training with early stopping. In which of the following cases $\hat{L}(h_{t^*}, S^{\text{val}})$ is an unbiased estimate of $L(h_{t^*})$ and in which cases it is not. Please, explain your answer.
   (a) Early stopping always stops after 100 iterations and returns $h_{t^*} = h_{100}$.
   (b) Early stopping runs as long as the validation error is decreasing and returns the last model before the increase in validation error. In other words, let i^* be the first time when $\hat{L}(h_{i^*}, S^{\text{val}}) > \hat{L}(h_{i^*-1}, S^{\text{val}})$, then $t^* = i^* - 1$.
   (c) Early stopping keeps track of the best model observed so far and keeps running as long as no improvement is observed for a significant number of rounds (this is the procedure proposed at the link cited above). At the end it returns the best model observed ever during training.
2. Derive a high-probability bound (a bound that holds with probability greater than $1 - \delta$) on $L(h_{t^*})$ in terms of $\hat{L}(h_{t^*}, S^{\text{val}})$, δ, and the sample size n for the three cases above. (Hint: note that in the last two cases you do not know in advance how many training epochs the algorithm will perform. The bound may depend on the index t^*.)
3. How would you use the bound to redefine the last two early stopping procedures?
4. The last approach suggests to "run the procedure as long as no improvement in the validation loss is observed for a significant number of rounds", but does not say explicitly what number of rounds is "significant". How would you use the bound to know when you can definitely stop (i.e., when the bound will definitely stop improving)?

Remark: There are several possible ways to solve Points 2, 3, and 4. Any reasonable solution that shows your understanding of the problem is acceptable. You do not necessarily have to provide the best possible bound.

Good luck!
Yevgeny, Christian, Tobias & Julian

References

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data.* AMLbook, 2012.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications and Signal Processing, 2nd edition, 2006.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.