CSC 665: Homework 3

Chicheng Zhang

November 18, 2019

Please complete the following set of problems. You must do the exercises completely on your own (no collaboration allowed). The exam is due **on Nov 14, 12:30pm, on Gradescope**. You are free to cite existing theorems from the textbooks and course notes.

Problem 1

Consider the homogeneous, soft-margin SVM optimization problem:

$$\begin{aligned} \underset{w,\xi}{\text{minimize}} \quad & \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i && && (1) \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle) \geq 1 - \xi_i, && \forall i \in \{1, \ldots, n\}, && (2) \\ & \xi_i \geq 0, && \forall i \in \{1, \ldots, n\}. \end{aligned}$$

1. Introducing dual variables $\alpha_i \geq 0$ for each constraint i, $i \in \{1, \ldots, n\}$ and $\beta_i \geq 0$ for each constraint i, $i \in \{1, \ldots, n\}$, compute the Lagrangian function $L(w, \xi, \alpha, \beta)$.
2. Derive the dual optimization problem.
3. Use the KKT condition to interpret: which of the training examples are "support vectors" that contribute to the SVM solution?

Solution

1. Define

$$L(w, \xi, \alpha, \beta) \triangleq \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\langle w, x_i \rangle)) + \sum_{i=1}^{n} \beta_i(-\xi_i)$$

It can be readily seen that the original optimization problem is equivalent to

$$\min_{w,\xi} \max_{\alpha \geq 0, \beta \geq 0} L(w, \xi, \alpha, \beta).$$

2. The dual problem is $\max_{\alpha \geq 0, \beta \geq 0} D(\alpha, \beta)$, where

$$\begin{aligned}
D(\alpha, \beta) &= \min_{w, \xi} L(w, \xi, \alpha, \beta) \\
&= \min_{w, \xi} \frac{\lambda}{2} \|w\|^2 - \left\langle w, \sum_{i=1}^n \alpha_i y_i x_i \right\rangle + \sum_{i=1}^n \xi_i (1 - \alpha_i - \beta_i) + \sum_{i=1}^n \alpha_i \\
&= \min_{w} \frac{\lambda}{2} \|w\|^2 - \left\langle w, \sum_{i=1}^n \alpha_i y_i x_i \right\rangle + \sum_{i=1}^n \min_{\xi_i} \xi_i (1 - \alpha_i - \beta_i) + \sum_{i=1}^n \alpha_i \\
&= \begin{cases} -\frac{1}{2\lambda} \| \sum_{i=1}^n \alpha_i y_i x_i \|^2 + \sum_{i=1}^n \alpha_i, & \forall i, \alpha_i + \beta_i = 1 \\ -\infty, & \exists i, \alpha_i + \beta_i \neq 1 \end{cases}
\end{aligned}$$

Therefore, we can further simplify the dual optimization problem: the above problem is equivalent to

$$\max_{\alpha \geq 0, \beta \geq 0, \alpha_i + \beta_i = 1, \forall i} -\frac{1}{2\lambda} \| \sum_{i=1}^n \alpha_i y_i x_i \|^2 + \sum_{i=1}^n \alpha_i$$

Note that β does not appear in the objective, therefore the only purpose of variables β is to create additional constraints on α. The set of feasible α is: $\left\{\alpha : \alpha_i \in [0, 1] \forall i\right\}$. Therefore, the dual problem can also be written as:

$$\max_{\alpha : \alpha_i \in [0,1], \forall i} -\frac{1}{2\lambda} \| \sum_{i=1}^n \alpha_i y_i x_i \|^2 + \sum_{i=1}^n \alpha_i.$$

3. Suppose $(w^\star, \xi^\star)$ and $(\alpha^\star, \beta^\star)$ are optimal solutions of the primal and dual problems respectively. The stationarity condition in the KKT condition with respect to w states that

$$\nabla_w L(w^\star, \xi^\star, \alpha^\star, \beta^\star) = 0,$$

Simplifying, we get that

$$w^\star = \frac{1}{\lambda} \sum_{i=1}^n \alpha_i^\star y_i x_i.$$

Those examples (x_i, y_i)'s such that $\alpha_i^\star > 0$ contributes to the solution of SVM.

The complementary slackness condition in the KKT condition with respect to α states that

$$\alpha_i^\star (1 - \xi_i^\star - y_i \langle w^\star, x_i \rangle) = 0.$$

Therefore, if $\alpha_i^\star > 0$, then $y_i \langle w^\star, x_i \rangle = 1 - \xi_i^\star \leq 1$. This implies that the "support vectors" that have margin value less than or equal to one.

Additional discussions (optional). There are two types of support vectors:

1. $\xi_i^\star = 0$. In this case, $y_i \langle w^\star, x_i \rangle = 1$, i.e. these examples have margin equal to one. They have general contribution coefficients $\alpha_i^\star \in [0, 1]$.

2. $\xi_i^\star > 0$. In this case, $y_i \langle w^\star, x_i \rangle < 1$. By the complementary slackness with respect to β, we get $\beta_i^\star \xi_i^\star = 0$, which implies that $\beta_i^\star = 0$. Furthermore, by the stationary condition with respect to ξ, we have $1 - \alpha_i^\star - \beta_i^\star = 0$, implying that $\alpha_i^\star = 1$. This shows that all examples with margins strictly less than 1 contribute maximally (with coefficient 1) to the SVM solution.

Problem 2

Suppose we have k finite hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$, and m training examples drawn iid from D. In addition we are given the promise that there exists $i_0 \in \{1, \ldots, k\}$ such that $\min_{h \in \mathcal{H}_{i_0}} \mathrm{err}(h, D) = 0$ (but we don't know the identity of i_0); Can we design an algorithm that produces classifiers with generalization error $O(\frac{\ln |\mathcal{H}_{i_0}|}{m})$ with high probability? Why or why not?

Solution

Yes. Let S be the set of training examples, and let $\hat{i} = \arg\min \left\{ |\mathcal{H}_i| : \min_{h \in \mathcal{H}_i} \mathrm{err}(h, S) = 0 \right\}$; and let $\hat{h} = \arg\min_{h \in \mathcal{H}_{\hat{i}}} \mathrm{err}(h, S)$. We show that with probability $1 - \delta$,

$$\mathrm{err}(\hat{h}, D) \le \frac{\ln |\mathcal{H}_{i_0}| + \ln \frac{k}{\delta}}{m}.$$

First, note that by the realizability assumption on $\mathcal{H}_{i_0}$ with respect to D, $\min_{h \in \mathcal{H}_{i_0}} \mathrm{err}(h, S) = 0$. By the optimality of $\hat{i}$, we have that

$$|\mathcal{H}_{\hat{i}}| \le |\mathcal{H}_{i_0}|.$$

Second, define E_i as the event that for all classifier h in $\mathcal{H}_i$, $\mathrm{err}(h, S) = 0 \Rightarrow \mathrm{err}(h, D) \le \frac{\ln |\mathcal{H}_i| + \ln \frac{k}{\delta}}{m}$. Define $E = \cap_{i=1}^{k} E_i$ In the lectures on realizable PAC learning, we have shown that $\mathbb{P}(E_i) \ge 1 - \frac{\delta}{k}$. Therefore, by union bound, $\mathbb{P}(E) \ge 1 - \delta$. Suppose for the rest of the proof that event E happens.

Observe that, on event E, event $E_{\hat{i}}$ also happens, so

$$\mathrm{err}(\hat{h}, D) \le \frac{\ln |\mathcal{H}_{\hat{i}}| + \ln \frac{k}{\delta}}{m} \le \frac{\ln |\mathcal{H}_{i_0}| + \ln \frac{k}{\delta}}{m}.$$

Remark. In fact, there is a refined structural risk minimization procedure that achieves a similar type of guarantee, while retaining the SRM guarantee in the non-realizable case. It computes

$$(\hat{i}, \hat{h}) = \arg \min_{i \in \{1, \ldots, k\}, h \in \mathcal{H}_i} \mathrm{err}(h, S) + \sqrt{\mathrm{err}(h, S) \epsilon(i, m)} + \epsilon(i, m),$$

where $\epsilon(i, m) = \Theta(\frac{\ln |\mathcal{H}_i| + \ln \frac{k}{\delta}}{m})$.

Problem 3

Show that for AdaBoost, at iteration t, the updated distribution D_{t+1} satisfies that

$$\sum_{i=1}^{m} D_{t+1}(i) \mathbf{1}(h_t(x_i) \ne y_i) = \frac{1}{2}.$$

Why is this a reasonable update?

Solution

Recall that $D_{t+1}(i) = D_t(i) e^{-\alpha_t y_i h_t(x_i)} / Z_t$, where

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}.$$

In additon, as discussed in class, $Z_t = 2\sqrt{(1 - \epsilon_t)\epsilon_t}$.

This implies that for i such that it gets misclassified by h_t, we have $y_i h_t(x_i) = -1$, therefore,

$$D_{t+1}(i) = D_t(i) \cdot \frac{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{2\sqrt{(1-\epsilon_t)\epsilon_t}} = D_t(i) \cdot \frac{1}{2\epsilon_t}.$$

Therefore,

$$\sum_{i=1}^{m} D_{t+1}(i)\mathbf{1}(h_t(x_i) \neq y_i) = (\sum_{i=1}^{m} D_t(i)\mathbf{1}(h_t(x_i) \neq y_i)) \cdot \frac{1}{2\epsilon_t} = \epsilon_t \cdot \frac{1}{2\epsilon_t} = \frac{1}{2}.$$

This update is reasonable, as it "forces" the weak learner to generate weak classifiers that are diverse: for example, if at iteration t, classifier h_t has been generated; then at iteration $t+1$, the weak learner would generate h_{t+1} that make predictions different from h_t at a reasonably large region, if the weak learning condition holds.

Problem 4

In this exercise, we conduct experiment with AdaBoost with a simple benchmark dataset named *ringnorm*.

1. Generate 100 training and 100 test examples from the following distribution D supported on $\mathbb{R}^{10} \times \{\pm 1\}$: $\mathbb{P}_D(Y=-1) = \mathbb{P}_D(Y=+1) = \frac{1}{2}$, $X|_{Y=+1} \sim \mathrm{N}((0,\ldots,0), 4I)$; $X|_{Y=-1} \sim \mathrm{N}((\frac{2}{\sqrt{20}},\ldots,\frac{2}{\sqrt{20}}), I)$.

2. Define base hypothesis class $\mathcal{H} = \left\{\sigma \cdot (2I(x_i \leq t) - 1), \sigma \in \{\pm 1\}, i \in \{1,\ldots,d\}, t \in \mathbb{R}\right\}$ as the set of bi-directional decision stumps. Let the weak learner $\mathcal{B}$ be: given a weighted dataset, return the classifier $h \in \mathcal{H}$ that has the smallest weighted error. Implement AdaBoost with $\mathcal{B}$, and run it for 300 iterations. At time t, suppose the following cumulative voting classifier

$$H_t(x) = \mathrm{sign}(f_t(x)), \quad f_s(x) = \sum_{s=1}^{t} \alpha_s h_s(x)$$

   is produced.

   Plot AdaBoost's learning curves: the training error of H_t, the test error of H_t and the training exponential loss of f_t as a function of iteration t. What do you see?

3. Given a voting classiifer f_t, define its normalization as

$$\bar{f}_t(x) = \frac{f_t(x)}{\sum_{s=1}^{t} \alpha_s} = \frac{\sum_{s=1}^{t} \alpha_s h_s(x)}{\sum_{s=1}^{t} \alpha_s}. \quad (3)$$

   Now, given an example (x,y), define its normalized margin at timestep t as $y\bar{f}_t(x)$. At iterations $t = 10, 30, 50, 100, 300$, show histograms of normalized margins of training examples. Do you see any tendency at t increases?

Solution

1. Depends on the language you use. In Python, you can use `np.random.randn` to generate examples from Gaussian distributions.

2. See Figure 1 below.

   An important observation is that, the empirical exponential loss is always nonincreasing. This can also be seen from AdaBoost's theoretical analysis that the empirical exponential loss at time t is the product of all normalization factors $\prod_{s=1}^{t} Z_s$, where all Z_s's are at most 1.

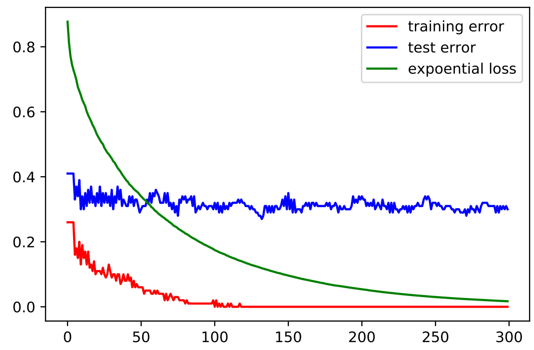

Figure 1: Training exponential loss, training error and test error vs. the number of iterations in AdaBoost.

Moreover, at every round t, the empirical error is upper bounded by the empirical exponential loss. This is due to the simple fact that $\mathbf{1}(z \leq 0) \leq \exp(-z)$. Both training error and empirical exponential loss tend to zero as learning proceeds.

For test error, it does not decrease as steeply as training error; it converges to a number around 0.3. (If you get final error around numbers such as 0.2, etc, that is also reasonable; I have seen this in other runs of my experiments.)

3. See Figure 2 below. At number of iterations t increases, examples of negative margins get "pushed" to have positive margins. On the other hand, the mode of the normalized margins gets shifted to the left, and as a consequence, more examples have small but positive margins.

Problem 5 (No need to submit)

Show that AdaBoost produces large-margin voting classifiers under the γ-weak learning assumption. If at every iteration t, $\epsilon_t \leq \frac{1}{2} - \gamma$, then after T rounds, the *margin error* of the output classifier will also decrease exponentially in T. Specifically, show:

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(y_i \bar{f}_T(x_i) \leq \frac{\gamma}{2}) \leq \exp\Big\{-\Omega(T\gamma^2)\Big\}.$$

where $\bar{f}_T$ is the normalized voting classifier defined as per Equation (3).

Solution

We first bound the margin error using exponential loss:

$$\mathbf{1}(y_i \bar{f}_T(x_i) \leq \frac{\gamma}{2}) = \mathbf{1}\left(y_i \sum_{t=1}^{T} \alpha_s h_s(x_i) \leq \frac{\gamma}{2} \cdot \sum_{t=1}^{T} \alpha_t\right) \leq \exp\left(\frac{\gamma}{2} \cdot \sum_{t=1}^{T} \alpha_t\right) \cdot \exp\left(-\sum_{t=1}^{T} \alpha_t y_i h_t(x_i)\right)$$

Therefore,

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(y_i\bar{f}_T(x_i)\leq\frac{\gamma}{2}) &\leq \exp\left(\frac{\gamma}{2}\cdot\sum_{t=1}^{T}\alpha_s\right)\cdot\left(\sum_{i=1}^{m}\exp\left(-\sum_{t=1}^{T}\alpha_s y_i h_s(x_i)\right)\right)\\
&= \exp\left(\gamma\cdot\sum_{t=1}^{T}\alpha_t\right)\cdot\prod_{t=1}^{T}Z_t\\
&= \prod_{t=1}^{T}\big(\exp(\gamma\alpha_t)\cdot Z_t\big)
\end{aligned}$$

where the equality is by the standard exponential loss analysis of AdaBoost.

Let $\gamma_t = \frac{1}{2}-\epsilon_t$. Let us look at $N_t = \big(\exp(\gamma\alpha_t)\cdot Z_t\big)$ in more detail. Note that $\gamma\leq\gamma_t$, which implies that $\exp(\gamma\alpha_t)\leq(1-\epsilon_t)^{\frac{\gamma_t}{4}}\epsilon_t^{-\frac{\gamma_t}{4}}$.

Therefore, N_t can be upper bounded as:

$$N_t\leq(1-\epsilon_t)^{\frac{\gamma_t}{4}}\epsilon_t^{-\frac{\gamma_t}{4}}\cdot 2\cdot\epsilon_t^{\frac{1}{2}}(1-\epsilon_t)^{\frac{1}{2}} = 2\cdot(\frac{1}{2}-\gamma_t)^{\frac{1}{2}-\frac{\gamma_t}{4}}\cdot(\frac{1}{2}+\gamma_t)^{\frac{1}{2}+\frac{\gamma_t}{4}}$$

Consequently,

$$\ln N_t = \ln 2 + \left(\frac{1}{2}-\frac{\gamma_t}{4}\right)\ln\left(\frac{1}{2}-\gamma_t\right)+\left(\frac{1}{2}+\frac{\gamma_t}{4}\right)\ln\left(\frac{1}{2}+\gamma_t\right).$$

Let $F(\theta)=\ln 2+\left(\frac{1}{2}-\frac{\theta}{4}\right)\ln\big(\frac{1}{2}-\theta\big)+\left(\frac{1}{2}+\frac{\theta}{4}\right)\ln\big(\frac{1}{2}+\theta\big)$. It can be checked that $F(0)=0$, $F'(0)=0$, and

$$F''(\theta) = -\frac{0.25\theta-0.5}{(0.5+\theta)^2}+\frac{0.5}{0.5-\theta}+\frac{0.5}{0.5+\theta}+\frac{0.25\theta-0.5}{(0.5-\theta)^2},$$

which is at most -2 for all $\theta\in[0,0.5)$ (see e.g. This plot by WolframAlpha).

By Taylor's theorem, this implies that $F(\theta)=\frac{F'(\xi)}{2}\theta^2$ for some ξ in $(0,\theta)$, which is at most $-\theta^2$. Therefore, $N_t\leq e^{-\theta^2}$.

In summary,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(y_i\bar{f}_T(x_i)\leq\frac{\gamma}{2})\leq\prod_{t=1}^{T}N_t\leq\exp\left\{-\sum_{t=1}^{T}\gamma_t^2\right\}\leq\exp\Big(-T\gamma^2\Big).$$

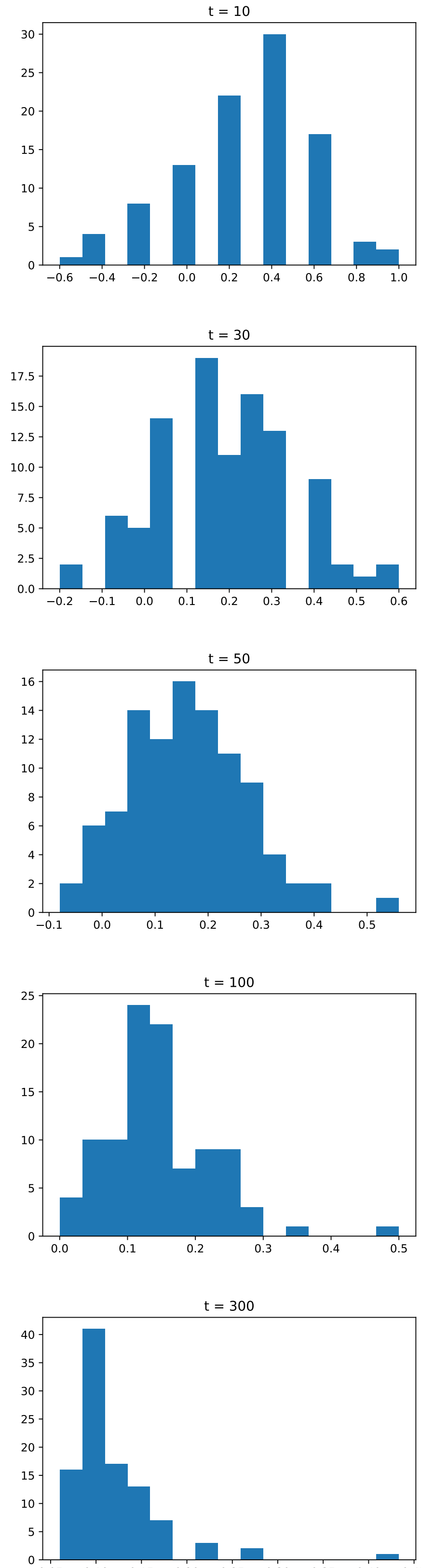

Figure 2: Histograms of margins at different iterations.