CS480/680: Introduction to Machine Learning

Homework 3

Due: 11:59 pm, March 18, 2025, submit on LEARN and Crowdmark.

NAME

student number

Submit your writeup in pdf and all source code in a zip file (with proper documentation). Write a script for each programming exercise so that the TA can easily run and verify your results. Make sure your code runs!
[Text in square brackets are hints that can be ignored.]

Exercise 1: Adaboost (5 pts)

In this exercise we will interpret Adaboost as minimizing the exponential loss:

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{n}\sum_{i=1}^{n} \exp\left[-y_i \sum_t \alpha_t h_t(\mathbf{x}_i)\right], \tag{1}$$

where h_t are the so-called weak learners, and the aggregated classifier

$$h_{\boldsymbol{\alpha}}(\mathbf{x}) := \sum_t \alpha_t h_t(\mathbf{x}). \tag{2}$$

Note that we assume $y_i \in \{\pm 1\}$ in this exercise.

Let us introduce the uniform distribution $\mathbf{p}_1 = \frac{1}{n}\mathbf{1}$ over our training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$, and rewrite (1) as:

$$\min_{\boldsymbol{\alpha}} \ \mathbb{E}_{\mathbf{p}_1} \exp\Big[-\mathsf{Y}\sum_t \alpha_t h_t(\mathsf{X})\Big], \tag{3}$$

where $(\mathsf{X}, \mathsf{Y}) \sim \mathbf{p}_1$, i.e., with probability $p_{i1} = \frac{1}{n}$, $\mathsf{X} = \mathbf{x}_i, \mathsf{Y} = y_i$.

1. (1 pt) Let $q > 0$ be an arbitrary function (or vector). By normalization, i.e., $q \leftarrow q/\int q$ or $q \leftarrow q/q^\top 1$ we obtain a density function (or probability mass function). Find probability density (vector) $\mathbf{p}_2$ below so that

$$\mathbb{E}_{\mathbf{p}_1} \exp\Big[-\mathsf{Y}\sum_{t=1}^{2} \alpha_t h_t(\mathsf{X})\Big] = Z_1 \cdot \mathbb{E}_{\mathbf{p}_2} \exp\Big[-\mathsf{Y}\sum_{t=2}^{2} \alpha_t h_t(\mathsf{X})\Big], \tag{4}$$

as well as the formula for Z_1 (a positive constant).

Ans:

2. (1 pt) Apply the previous exercise repeatedly with probability densities (vectors) $\mathbf{p}_t$ so that

$$\mathbb{E}_{\mathbf{p}_1} \exp\Big[-\mathsf{Y}\sum_{t=1}^{T} \alpha_t h_t(\mathsf{X})\Big] = \prod_{t=1}^{T} Z_t, \tag{5}$$

where each Z_t is a positive constant. Explain what is Z_t for each t.

Ans:

3. (1 pt) Prove the following bound on the training error:

$$\mathbb{E}_{\mathbf{p}_1} [\![\mathsf{Y} \underbrace{h_{\boldsymbol{\alpha}}(\mathsf{X})}_{\hat{\mathsf{Y}}} \leq 0]\!] = \mathbb{E}_{\mathbf{p}_1} [\![\mathsf{Y} \sum_{t=1}^{T} \alpha_t h_t(\mathsf{X}) \leq 0]\!] \leq \prod_{t=1}^{T} Z_t, \tag{6}$$

where each Z_t is given in the previous exercise. Recall that $h_{\boldsymbol{\alpha}}(\mathsf{X}) = \sum_{t=1}^{T} \alpha_t h_t(\mathsf{X})$ is the aggregated classifier.

Ans:

4. (1 pt) Assuming in the t-th iteration we have found h_t. We now aim to find its coefficient α_t by considering the following (convex) minimization problem:

$$\min_{\alpha_t} \mathbb{E}_{\mathbf{p}_t} \exp[-\mathsf{Y}\alpha_t h_t(\mathsf{X})], \tag{7}$$

Suppose there is indeed a minimizer, then it must satisfy (by setting derivative to 0):

$$0 = \mathbb{E}_{\mathbf{p}_t} \big\{ \mathsf{Y} h_t(\mathsf{X}) \cdot \exp[-\mathsf{Y}\alpha_t h_t(\mathsf{X})] \big\}. \tag{8}$$

From the above result deduce that

$$0 = \mathbb{E}_{\mathbf{p}_{t+1}} [\mathsf{Y} h_t(\mathsf{X})], \tag{9}$$

and show that for (deterministic) weak classifiers $h_t \in \{\pm 1\}$:

$$\mathbb{E}_{\mathbf{p}_{t+1}} [\![h_t(\mathsf{X}) \neq \mathsf{Y}]\!] = \tfrac{1}{2}, \tag{10}$$

namely that in the next iteration $t+1$, the previous classifier h_t has error exactly $\frac{1}{2}$.

Ans:

5. (1 pt) In (8) above we obtained a nonlinear equation of α_t. Although it is possible to find α_t through numerical root finding algorithms, we prefer to derive a closed-form solution. Assuming $h_t \in [-1, 1]$ is given, we can apply the bound (the so-called Jensen's inequality)

$$\exp(-\alpha u) \leq \frac{1+u}{2}\exp(-\alpha) + \frac{1-u}{2}\exp(\alpha) \tag{11}$$

to (7) first and then derive the optimal α_t.

[This is essentially the coefficient $\log \frac{1}{\beta_t}$ that we saw in class, up to some trivial changes.]

Ans:

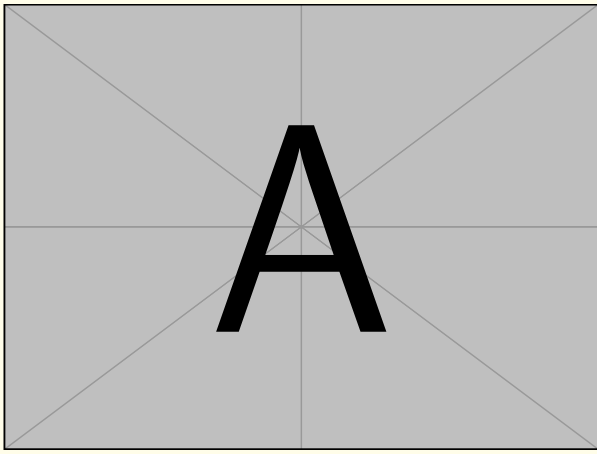
(Remark) We have not talked much about choosing weak classifiers h_t. Here is the catch: we could simply pretend we enumerate all weak classifiers (infinitely many!) in our final aggregate $h_{\boldsymbol{\alpha}}$. All we need to figure out is the weight α_t that we assign to each weak classifier h_t, and a zero α means the corresponding weak classifier is effectively discarded. The Adaboost algorithm starts with $\alpha \equiv 0$ and only changes one α into nonzero in each iteration. Thus, after T iterations, we have at most T nonzero α's. On a high level, this is very similar to kernels where the dual problem only involves n nonzero Lagrangian multiplier α's (n being the size of the training set). See, we do not need to fear infinite dimensions!

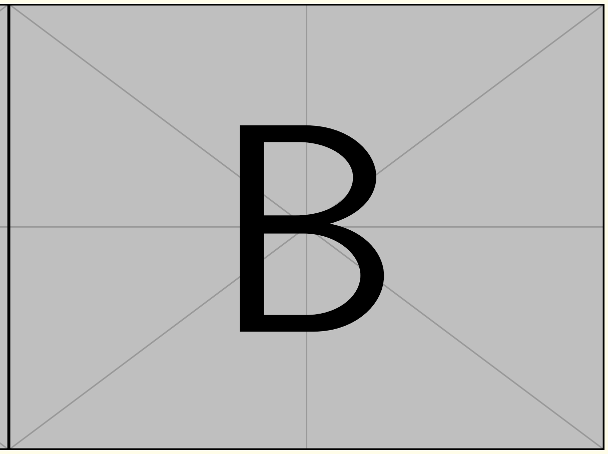
Exercise 2: Vision Transformers (10 pts)

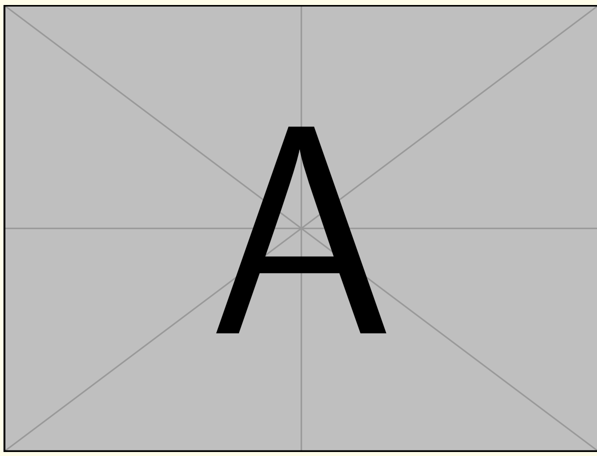
Please follow the instructions of this `ipynb` file.

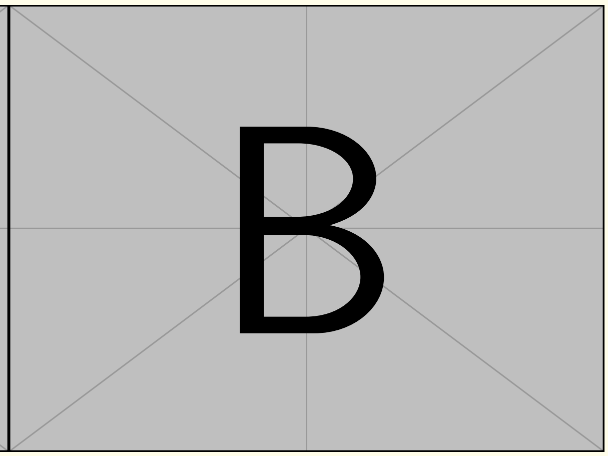
1. (1+3+2 = 6 pts) Complete the missing coding parts in the provided `ipynb` file.

2. (1 pt) Visualization of patches:

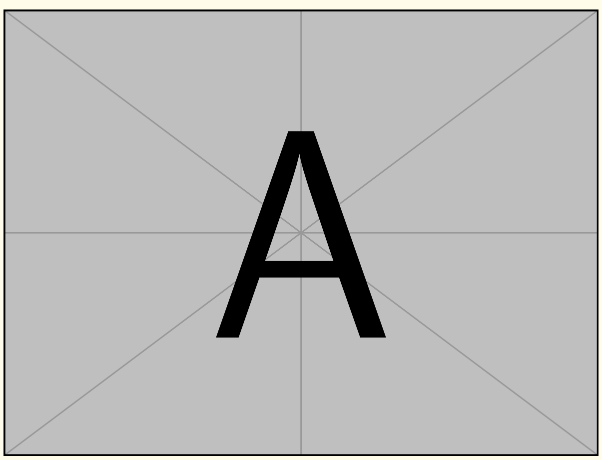


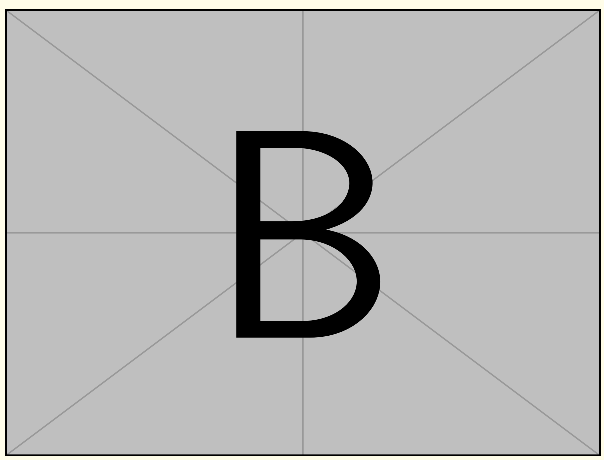






3. (1 pt) The test accuracy I obtained on MNIST is: xxx%
4. (2 pts) Training / Validation accuracy vs. epoch:





Exercise 3: Generative Adversarial Networks (5 pts)

Let us consider the game between the generator $\mathsf{q}(\mathbf{x})$ (the implicit density of $\mathsf{T}_{\boldsymbol{\theta}}(\mathsf{Z})$) and the discriminator $S(\mathbf{x})$:

$$\inf_{\mathsf{q}} \sup_{S} \int_{\mathbf{x}} S(\mathbf{x})\mathsf{p}(\mathbf{x})\mathrm{d}\mathbf{x} + \int_{\mathbf{x}} \log\left(1 - \exp(S(\mathbf{x}))\right)\mathsf{q}(\mathbf{x})\mathrm{d}\mathbf{x} + \log 4. \tag{12}$$

We remind that q is a probability density (so is p which is given) while S is any real-valued function.

1. (1 pt) Fix an arbitrary generator q and find the resulting optimal discriminator S.

   Ans:

2. (1 pt) Plug the optimal discriminator S above back to (12) and find the optimal generator q.

   Ans:

Now we swap the order of the two players:

$$\sup_{S} \inf_{\mathsf{q}} \int_{\mathbf{x}} S(\mathbf{x})\mathsf{p}(\mathbf{x})\mathrm{d}\mathbf{x} + \int_{\mathbf{x}} \log\big(1 - \exp(S(\mathbf{x}))\big)\mathsf{q}(\mathbf{x})\mathrm{d}\mathbf{x} + \log 4. \tag{13}$$

3. (1 pt) Fix an arbitrary discriminator S and find an optimal generator q.

   Ans:

4. (1 pt) Plug the optimal generator q above back to (13) and find the optimal discriminator S.

   [Hint: average $\leq$ max.]

   Ans:

5. (1 pt) Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function. We see in class that the f-divergence admits the following variational form:

$$-\mathbb{D}_f(\mathsf{p}\|\mathsf{q}) = \inf_{S:\mathbb{R}^d \to [0,1]} -\mathbb{E}_{\mathsf{X}\sim\mathsf{p}}[S(\mathsf{X})] + \mathbb{E}_{\mathsf{X}\sim\mathsf{q}}[f^*(S(\mathsf{X}))], \tag{14}$$

where for simplicity we have restricted the range of S to $[0, 1]$. Now consider the following distribution $(\mathsf{X}, \mathsf{Y}) \sim \mathcal{D}$ where $\mathsf{Y} = \pm 1$ with equal probability while

$$[\mathsf{X} \mid \mathsf{Y} = 1] \ \sim \ \mathsf{p} \text{ and } [\mathsf{X} \mid \mathsf{Y} = -1] \ \sim \ \mathsf{q}. \tag{15}$$

We claim that

$$-\tfrac{1}{2}\mathbb{D}_f(\mathsf{p}\|\mathsf{q}) = \inf_{S:\mathbb{R}^d \to [0,1]} \ \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathcal{D}}[\ell(\mathsf{Y}, S(\mathsf{X}))]. \tag{16}$$

Express the binary loss function ℓ in terms of f. Thus, given an f-divergence, we may rewrite it as a binary classification problem! Conversely, given any *proper* loss function ℓ, we may reverse the argument and induce an f-divergence from the binary loss ℓ.

Ans: