Statistical Machine Learning
Homework 3

Instructor: Prof. Jun Zhu

April 19, 2021

Requirements:

- We recommend that you typeset your homework using appropriate software such as LaTeX. If you submit your handwritten version, please make sure it is cleanly written up and legible. The TAs will not invest undue effort to decrypt bad handwritings.
- We have programming tasks in each homework. Please submit the source code together with your homework. Please include experiment results using figures or tables in your homework, instead of asking TAs to run your code.
- Please finish your homework independently. In addition, you should write in your homework the set of people with whom you collaborated.

1 Dimension Reduction, PCA (3pt)

Problem 1 (PCA, 3pt)**.** Implement the PCA algorithm and run it on the whole MNIST training set[1].

1. Choose d to preserve 1%, 5%, 20%, 50%, 80%, 95%, 99% information respectively. Fix two images of different digits and show the resulting images for each d.
2. Visualize the top 100 eigenvectors to see how they look.
3. Re-run your PCA implementation without centering the dataset (subtracting the sample mean) and compare the results, using the same two images you choose before.

Please plot the images in the report and submit your code alongside.

2 Learning Theory (4pt)

Problem 2 (General setting of learning, 1pt)**.** Let Z be an abstract set representing the data space, H be an abstract set representing the hypothesis space and $\ell : H \times Z \to [0, \infty)$ be a loss function, which measures how a hypothesis $h \in H$ performs on a data sample $z \in Z$. Based on this setting, we can define the generalization error of a hypothesis $h \in H$ on a distribution D of Z:

$$R(h, D) \triangleq \mathbb{E}_{z \sim D} \left[\ell(h, z) \right],$$

which measures the averaged performance of h when the underlying data generating process satisfies D. We can also define the empirical error of a hypothesis $h \in H$ on a set of data samples $S = (z_1, z_2, \cdots, z_m) \in Z^m$:

$$\hat{R}(h, S) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i),$$

[1] http://yann.lecun.com/exdb/mnist/

which measures the averaged performance of h on S. Many special settings can be viewed as special cases of this general setting. For example, the binary classification is a special case by letting

$$\begin{aligned} &Z = X \times \{0, 1\}, \\ &H = \{f : f \text{ is a mapping from } X \text{ to } \{0, 1\}\}, \\ &\ell(h, (x, y)) = 1_{h(x) \neq y}. \end{aligned}$$

Please show that the regression with squared error can be also viewed as a special case of the general setting. *[Hint: You can assume that the input space is X and the output space is $Y \subset \mathbb{R}$, then write down the corresponding Z, H and ℓ.]*

Problem 3 (Generalization error with random labels, 1pt)**.** Considering the setting for the binary classification introduced in Problem 2, please prove that if the data distribution D satisfies that the data label is randomly labelled as 0 or 1 with equal probability, then for all hypothesis h, the generalization error $R(h, D)$ is always $\frac{1}{2}$.

Problem 4 (Bound the generalization error of ERM, 2pt)**.** Consider the general setting of learning introduced in Problem 2. The empirical error minimization (ERM) algorithm minimizes the empirical error on a set of samples $S \in Z^m$ by

$$h_S^{ERM} \in \arg\min_{h \in H} \ell(h, S),$$

where h_S^{ERM} is the hypothesis output by the ERM algorithm.

1. Let D be a data distribution and $R^* \triangleq \inf_{h \in H} R(h, D)$ be the infimum of the generalization error on D, prove

$$R(h_S^{ERM}, D) - R^* \leq 2 \sup_{h \in H} |R(h, D) - \hat{R}(h, S)|.$$

2. Now we can give a generalization error bound of ERM. Suppose H is a finite set and the loss function ℓ is upper bounded by M, please show that

$$\begin{aligned} &\forall \epsilon > 0, P_{S \sim D^m}\left(R(h_S^{ERM}, D) - R^* > \epsilon\right) \leq 2|H| \exp\left(-\frac{m\epsilon^2}{2M^2}\right), \\ &\forall \delta \in (0, 1), P_{S \sim D^m}\left(R(h_S^{ERM}, D) \leq R^* + \sqrt{\frac{2M^2(\ln 2|H| + \ln \delta^{-1})}{m}}\right) \geq 1 - \delta. \end{aligned}$$

   [Hint: Hoeffding's inequality.]

3 Deep Generative Models (3pt)

Problem 5 (Gaussian VAE vs Bernoulli VAE, 3pt)**.** Consider two types of VAEs. The first one defines $p(\boldsymbol{x}|\boldsymbol{z})$ as a Bernoulli distribution:

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{x}|\boldsymbol{z} \sim \text{Bernoulli}(\mu_\theta(\boldsymbol{z})).$$

The second one defines $p(\boldsymbol{x}|\boldsymbol{z})$ as a Gaussian distribution:

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}(\mu_\theta(\boldsymbol{z}), \sigma^2\mathbf{I}).$$

In the problem we fix the dimension of $\boldsymbol{z}$ as 40, uses the MNIST dataset[1] and chooses $\mu_\theta(\boldsymbol{z})$ as an MLP.

1. Implement a Bernoulli VAE and plot samples generated by the Bernoulli VAE. (Hint: the data should be binarized before training or testing.)

2. Implement a Gaussian VAE, report your selected σ and plot samples generated by the Gaussian VAE. You can try it under different σ.

3. Compare the sample quality between Bernoulli VAE and Gaussian VAE and analyze the reason. (There is not a standard answer, just write whatever comes to mind.)