Statistical Machine Learning
Homework 1

Instructor: Prof. Jun Zhu

March 8, 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 Kernel Methods

Kernel methods lift data into high-dimensional spaces. The following problems kernelize some algorithms.

We say a mapping $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exist a space $\mathcal{F}$ with an inner product $\langle \cdot, \cdot \rangle$, and a feature map $\phi : \mathcal{X} \to \mathcal{F}$ such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. For example,

- $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel when we choose $\mathcal{X} = \mathbb{R}^d$, $\mathcal{F} = \mathbb{R}^d$, and $\phi(\mathbf{x}) = \mathbf{x}$.
- $k(\mathbf{x}, \mathbf{y}) = 1 + \mathbf{x}^\top \mathbf{y}$ is a kernel when we choose $\mathcal{X} = \mathbb{R}^d$, $\mathcal{F} = \mathbb{R}^{d+1}$, and $\phi(\mathbf{x}) = (1, \mathbf{x})$. This kernel lifts the data in $\mathbb{R}^d$ into $\mathbb{R}^{d+1}$.

Problem 1 (1pts)**.** Please use the above definition of kernels to solve this problem.

1. Prove that $k(x, y) = (1 + xy)^n$ is a kernel on $\mathcal{X} = \mathbb{R}$.
2. Prove that $k(x, y) = xy - 1$ is not a kernel on $\mathcal{X} = \mathbb{R}$. *[Hint: Prove by contradiction.]*
3. (Bonus, 0.5pts). Prove that $k(x, y) = \min(x, y)$ is a kernel on $\mathcal{X} = [0, 1]$.

Problem 2 (Kernel SVM for Classification, 1pts)**.**

This problem is a direct extension of the linear SVM introduced in this course, so it is **optional**. If you are not willing to write down the answer, Problem 6 will be 4.5pts.

Given a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^m$ be a feature map. Consider the following primal SVM problem:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^m, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, \\ & y_i \mathbf{w}^\top \phi(\mathbf{x}_i) \geq 1 - \xi_i. \end{aligned} \tag{1.1}$$

1. Write down the Lagrangian function of (1.1).

2. Derive the dual problem of (1.1) using the kernel $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y})$ instead of the feature map ϕ. The feature map ϕ is not allowed to appear in the result.

3. Let $\hat{\mathbf{w}}, \hat{\boldsymbol{\xi}}$ be the solution of (1.1). Express the prediction function $f(\mathbf{x}) = \text{sign}(\hat{\mathbf{w}}^\top \phi(\mathbf{x}))$ using the kernel and the solutions of the dual problem. The feature map ϕ is not allowed to appear in the result.

Problem 3 (Kernel Ridge Regression, 1pts)**.** Given a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^m$ be a feature map. Consider the following regression problem:

$$\min_{\mathbf{w} \in \mathbb{R}^m} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^N (\mathbf{w}^\top \phi(\mathbf{x}_i) - y_i)^2. \tag{1.2}$$

1. Derive the solution $\hat{\mathbf{w}}$ of (1.2).

2. Express the prediction function $f(\mathbf{x}) = \hat{\mathbf{w}}^\top \phi(\mathbf{x})$ using the kernel $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y})$. The feature map ϕ is not allowed to appear in the result.

Problem 4 (3pts)**.** You need to implement the kernel ridge regression in this problem.

1. Randomly generate $N = 30$ samples $x_1, x_2, \ldots, x_N$ from the uniform distribution in the interval $[0, 1]$. Randomly generate noise $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ from the standard Gaussian distribution. Set $y_i = -5 + 20x_i - 16x_i^2 + \frac{2}{5}\epsilon_i$. We will use $\{(x_i, y_i)\}_{i=1}^N$ as our dataset.

2. Use the kernel ridge regression derived in the previous problem to fit the dataset.

   (a) Try to fit the dataset using different λ.

   (b) Try to fit the dataset using different kernels, e.g., $(1 + xy)^n$ with $n = 1, 2, 9$, and $\min(x, y)$, and $\exp\left(-(x - y)^2\right)$.

   (c) You need to plot the dataset $\{(x_i, y_i)\}$ and the target function $x \mapsto -5 + 20x - 16x^2$ and the prediction function $x \mapsto \hat{\mathbf{w}}^\top \phi(x)$.

2 Exponential Families

In these problems, we consider the exponential family

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^\top T(\mathbf{x}) - A(\boldsymbol{\eta})\right), \tag{2.1}$$

where $T(\mathbf{x})$ is a sufficient statistic, $A(\boldsymbol{\eta}) = \log \int h(\mathbf{x}) e^{\boldsymbol{\eta}^\top T(\mathbf{x})} \mathrm{d}\mathbf{x}$ is the partition function.

Problem 5 (0.5pts)**.** Verify that

$$\frac{\partial}{\partial \eta_i} A(\boldsymbol{\eta}) = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}[T_i(\mathbf{x})],$$
$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\boldsymbol{\eta}) = \text{Cov}_{p(\mathbf{x}|\boldsymbol{\eta})}[T_i(\mathbf{x}) T_j(\mathbf{x})],$$

where Cov is the covariance and $T_j(\mathbf{x})$ is the j-th component of $T(\mathbf{x})$.

Problem 6 (3.5pts)**.** We consider the maximum likelihood estimation of the multivariate Gaussian distribution and its convergence properties. Recall that the density function of the d-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

Given i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ from $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown parameters.

1. Find the maximum likelihood estimators (MLE) $\hat{\boldsymbol{\mu}}_{\text{ML}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{ML}}$.

   [Hint: Note that the multivariate Gaussian distribution belongs to the exponential family and use some results in slides. You can also use the following fact: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.]

2. Compute $\mathbb{E}[\hat{\boldsymbol{\mu}}_{\mathrm{ML}}]$ and $\mathbb{E}[\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}]$, where both expectations are taken with respect to $p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Are these estimators unbiased[1]?

3. Show that

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu}\|^2\right] = \frac{\mathrm{Tr}\,\boldsymbol{\Sigma}}{N}, \tag{2.2}$$

where $\mathrm{Tr}\,\boldsymbol{\Sigma}$ is the trace of the matrix $\boldsymbol{\Sigma}$.

4. Now, we consider whether the maximum likelihood estimator $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ is optimal. In the following, we assume that $\boldsymbol{\Sigma} = \sigma^2 \in \mathbb{R}$ is known and $d = 1$. So, the only unknown parameter is $\boldsymbol{\mu}$. It is easy to verify that the MLE of $\boldsymbol{\mu}$ is the same as before. For simplicity, we suppress the parameter $\boldsymbol{\Sigma}$ and use $p(\mathbf{x}|\boldsymbol{\mu})$ to represent the density $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and use $p(\mathbf{X}|\boldsymbol{\mu})$ to represent the joint density $p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N|\boldsymbol{\mu})$, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$. Let $\hat{\boldsymbol{\mu}}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ be any unbiased estimator of $\boldsymbol{\mu}$, i.e., $\mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$. Assume that the variance of $\hat{\boldsymbol{\mu}}$ is finite, i.e., $\mathrm{Var}_{p(\mathbf{X}|\boldsymbol{\mu})}[\hat{\boldsymbol{\mu}}] < \infty$, and we can interchange the integral and the differential as follows:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}[\hat{\boldsymbol{\mu}}(\mathbf{X})] = \int \frac{\partial}{\partial \boldsymbol{\mu}} [\hat{\boldsymbol{\mu}}(\mathbf{X}) p(\mathbf{X}|\boldsymbol{\mu})] \mathrm{d}\mathbf{X}. \tag{2.3}$$

   (a) Let X and Y be two random variables, prove the Cauchy-Schwarz inequality

$$[\mathrm{Cov}(X, Y)]^2 \leq (\mathrm{Var}X)(\mathrm{Var}Y). \tag{2.4}$$

   (b) Prove that

$$\mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}\left[\frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu})\right] = 0. \tag{2.5}$$

   Note that this equation is widely used in the variational inference literature.

   (c) Prove that

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}[\hat{\boldsymbol{\mu}}(\mathbf{X})] = \mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}\left[\hat{\boldsymbol{\mu}}(\mathbf{X}) \frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu})\right] = \mathrm{Cov}\left(\hat{\boldsymbol{\mu}}(\mathbf{X}), \frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu})\right). \tag{2.6}$$

   The first equation and its variants have been used in many machine learning papers in recent years.

   (d) Choose $X = \hat{\boldsymbol{\mu}}(\mathbf{X})$ and $Y = \frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu})$ in the Cauchy-Schwarz inequality. Conclude that

$$\mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}\left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\right] = \mathrm{Var}_{p(\mathbf{X}|\boldsymbol{\mu})}[\hat{\boldsymbol{\mu}}(\mathbf{X})] \geq \frac{\left(\frac{\partial}{\partial \boldsymbol{\mu}} \mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}[\hat{\boldsymbol{\mu}}(\mathbf{X})]\right)^2}{\mathbb{E}_{p(\mathbf{X}|\boldsymbol{\mu})}\left[\left(\frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu})\right)^2\right]} = \frac{\sigma^2}{N}. \tag{2.7}$$

   Thus, when the variance σ^2 is known, the maximum likelihood estimator $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ has the minimal variance among all unbiased estimators that satisfy the conditions in this problem.

[1] An estimator $\hat{\mu}$ is unbiased if $\mathbb{E}\hat{\mu} = \mu$.