Machine Learning October, 2022

Problem Set 1

Instructor: Prof. Jie Tang, Prof. Jun Zhu *Student Name*

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 Collaborators and Sources

Please list your collaborators and sources here.

2 Maximum Likelihood Estimators (3pts)

Problem 1 (3pts)**.** 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) $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}$.
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

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_{\mathrm{ML}}-\boldsymbol{\mu}\|^2\right] = \frac{\operatorname{Tr}\boldsymbol{\Sigma}}{N}, \tag{2.1}$$

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

3 Kernel SVM (4pts)

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,

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

- $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ 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 2 (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.]*

Problem 3 (Kernel SVM for Classification, 3pts)**.**
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:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^m, \boldsymbol{\xi}\in\mathbb{R}^N} \quad & \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, \\ & y_i \mathbf{w}^\top \phi(\mathbf{x}_i) \geq 1 - \xi_i. \end{aligned} \tag{3.1}$$

1. Write down the Lagrangian function of (3.1).
2. Derive the dual problem of (3.1) 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 (3.1). Express the prediction function $f(\mathbf{x}) = \operatorname{sign}(\hat{\mathbf{w}}^\top \phi(\mathbf{x}))$ using the kernel and the solutions of the dual problem. The feature map ϕ is not allowed to appear in the result.

4 Boosting: from Weak to Strong (3pts)

Boosting takes a weak learning algorithm - any learning algorithm that gives a classifier that is slightly better than random - and transforms it into a strong classifier, which does much better than random. In this problem, we show that using a binary classification problem on $\mathbb{R}^1$ as example, we can get a strong classifier which achieves zero error on training dataset by boosting simple thresholding-based decision stumps.

Here we first list some definitions and existing results on boosting for reference. Given a training dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$, we call a vector $p = \{p^{(i)}\}_{i=1}^m$ a distribution on the examples if $p^{(i)} \geq 0, \forall i$ and $\sum_{i=1}^m p^{(i)} = 1$.

Definition 1. *A weak learner with margin $\gamma > 0$ is that, if for any distribution p on the m training examples there exists one weak classifier ϕ_j such that*

$$\sum_{i=1}^m p^{(i)} \mathbf{1}\left\{y^{(i)} \neq \phi_j\left(x^{(i)}\right)\right\} \leq \frac{1}{2} - \gamma, \tag{4.1}$$

where $\mathbf{1}\left\{y^{(i)} \neq \phi_j\left(x^{(i)}\right)\right\}$ is 1 if the expression inside the bracket is true and 0 otherwise.

Definition 2. *The thresholding-based decision stumps are functions on $x \in \mathbb{R}^1$, indexed by a threshold s and sign $+/-$, such that*

$$\phi_{s,+}(x) = \begin{cases} 1 & \text{if } x \geq s \\ -1 & \text{if } x < s \end{cases} \tag{4.2}$$

and

$$\phi_{s,-}(x) = \begin{cases} -1 & \text{if } x \geq s \\ 1 & \text{if } x < s \end{cases} \tag{4.3}$$

We have $\phi_{s,+}(x) = -\phi_{s,-}(x)$.

Theorem 1. *(Convergence of Boosting) If in each iteration, the boosting procedure (similar to the algorithm shown in the lecture notes) can generate a weak classifier with margin* γ*, then*

$$J_t \leq \sqrt{1-4\gamma^2} J_{t-1}, \tag{4.4}$$

where J_t *is denoted as the error rate on training dataset after the t -th iteration. Obviously, if* $J_t < 1/m$*, we have zero training error.*

We consider the following binary classification problem on $\mathbb{R}^1$ with training dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^1, y \in \{1, -1\}$. We additionally assume that $x^{(i)}$ are distinct and

$$x^{(1)} > x^{(2)} > \ldots > x^{(m)} \tag{4.5}$$

We would like guarantee that for any distribution p on the training set, there is some $\gamma > 0$ and a threshold s such that,

$$\sum_{i=1}^m p_i \mathbf{1}\left\{y^{(i)} \neq \phi_{s,+}\left(x^{(i)}\right)\right\} \leq \frac{1}{2} - \gamma \tag{4.6}$$

or

$$\sum_{i=1}^m p_i \mathbf{1}\left\{y^{(i)} \neq \phi_{s,-}\left(x^{(i)}\right)\right\} \leq \frac{1}{2} - \gamma \tag{4.7}$$

That it, in each iteration we get a weak classifier with margin γ. Hence zero training error can be achieved by boosting.

Problem 4 (3pts). Please use the above definition to solve this problem

1. Prove that for each threshold s, there is some $m_0(s) \in \{0, 1, \ldots, m\}$ such that

$$\sum_{i=1}^m p_i \mathbf{1}\left\{\phi_{s,+}\left(x^{(i)}\right) \neq y^{(i)}\right\} = \frac{1}{2} - \frac{1}{2}\left(\sum_{i=1}^{m_0(s)} y^{(i)} p_i - \sum_{i=m_0(s)+1}^{m} y^{(i)} p_i\right) \tag{4.8}$$

   and

$$\sum_{i=1}^m p_i \mathbf{1}\left\{\phi_{s,-}\left(x^{(i)}\right) \neq y^{(i)}\right\} = \frac{1}{2} - \frac{1}{2}\left(\sum_{i=m_0(s)+1}^{m} y^{(i)} p_i - \sum_{i=1}^{m_0(s)} y^{(i)} p_i\right) \tag{4.9}$$

   Treat sums over empty sets of indices as zero, so that $\sum_{i=1}^0 a_i = 0$ for any a_i, and similarly $\sum_{i=m+1}^m a_i = 0$. *[Hint: Note that* $\mathbf{1}\{y=-1\} = \frac{1-y}{2}$*,* $\mathbf{1}\{y=1\} = \frac{1+y}{2}$*]*

2. Define that for each $m_0 \in \{0, 1, \ldots, m\}$ we have

$$f(m_0) = \sum_{i=1}^{m_0} y^{(i)} p_i - \sum_{i=m_0+1}^{m} y^{(i)} p_i \tag{4.10}$$

   Prove that given $\gamma = \frac{1}{2m}$ we have

$$\max_{m_0} |f(m_0)| \geq 2\gamma \tag{4.11}$$

   [Hint: Prove that $|f(m_0) - f(m_0+1)| \geq \frac{2}{m}$*]*

3. Based on the above answer, how large margin γ can *thresholded decision stumps* guarantee on any training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$? (i.e., how "good" a weak classifier can we get in each boosting iteration?) Give an upper bound on the number of thresholded decision stumps required to achieve zero error on a given training set.