CSE 326/426 (Fall 2022) Homework 5

Due on 11:55 pm, Friday, Sep 30, 2022

Grading: All questions have the same points (10 each). Two questions from the first four will be randomly selected for grading for all students. The last question is required for graduates.

Submitting: Only electronic submissions in PDF format on Coursesite are accepted. Name your file as

<Your LIN>_HW5.pdf

A few ways to create a single PDF file. i) Use Microsoft Word and insert your writing or any figure taken, scanned, or plotted. Then choose "File→Print" in the main menu and you will find an option for outputting the file to a PDF file. ii) Use Latex to write your solution and include any figures. iii) Use the online Google Doc as an alternative of Word. It has sufficient features to combine multiple images and texts. Exporting to PDF files is similar as with Word.

Please DO NOT compress the PDF file as that will slow down the processing of your submission and the grading.

Questions:

1. Let the Lagrangian function of the primal SVM be

$$L(\mathbf{w}, \boldsymbol{\lambda}) = f(\mathbf{w}) - \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) = f(\mathbf{w}) - \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{w}). \tag{1}$$

   with the multipliers $\boldsymbol{\lambda} \succeq \mathbf{0}$ and inequality constraints $\mathbf{g}(\mathbf{w}) \succeq \mathbf{0}$. Prove the weak duality property:

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\lambda}) \leq \min_{\mathbf{w}} \max_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{w}, \boldsymbol{\lambda}). \tag{2}$$

   [*Hints: First prove $L(\mathbf{w}, \boldsymbol{\lambda}') \leq \max_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{w}, \boldsymbol{\lambda})$ for any $\mathbf{w}$. Second, prove $\min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\lambda}') \leq \min_{\mathbf{w}} \max_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{w}, \boldsymbol{\lambda})$ for any $\boldsymbol{\lambda}'$ on the left hand side. Lastly, maximize both sides over $\boldsymbol{\lambda}'$ to obtain the final inequality. Note that for any function $f(x, y)$, $\max_x f(x, y)$ is no longer a function of x but a function of y. The same for minimization.*]

2. Let $\mathcal{D}$ be the uniform distribution on the unit circle $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 = 1\}$ so that the probability of any segment of the circle is proportional to the ratio of the segment length to 2π. Let the labeling function f be $f(\mathbf{x}) = 1$ if $|x_2| > \sqrt{2}/2$ and $f(\mathbf{x}) = 0$ otherwise. Given two training examples at $\mathbf{x}^{(1)} = [0, 1]^\top$ and $\mathbf{x}^{(2)} = [1, 0]^\top$, find the maximum margin linear classifier $\hat{h}(\mathbf{x}; \mathbf{w})$ from the hypothesis class $\mathcal{H} = \{h : h(\mathbf{x}; \mathbf{w}) = w_1 x_1 + w_2 x_2\}$ that can correctly classify these two training examples. Find the two geometric margins.

3. Calculate the generalization error $L_{\mathcal{D},f}(\hat{h})$ with respect to the distribution $\{\mathcal{D}\}$ and the labeling function f in the above question. Make sure you find the regions where $\hat{h}$ mis-classifies the data and calculate the probability of these regions under distribution $\mathcal{D}$.

4. (Kernelizing linear regression) Given the MSE loss function of linear regression

$$L(\mathbf{w}; \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m}) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \left\langle \mathbf{w}, \mathbf{x}^{(i)} \right\rangle \right)^2, \tag{3}$$

Let $\mathbf{w} = \sum_{j=1}^{m} \alpha_j \mathbf{x}^{(j)}$. Re-write the loss function in terms of linear kernel $\left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle = K_{ij}$.

[*Hints: refer to the derivation of the dual problem of SVM using the KKT condition* $\mathbf{w} = \sum_{j=1}^{m} \alpha_j \mathbf{x}^{(j)}$ *for some dual variables* $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]$. *Note that* $K_{ij} = K_{ji}$.]

5. (Graduate only) Continue working on Question 3, fixing f and the $\hat{h}$ you have identified, find a data distribution $\mathcal{D}$ so that the generalization error is zero. Make sure to explicitly define the probability of the regions on the circle where the classifier $\hat{h}$ makes correct prediction. Define the regions using set notation (e.g., $\{\mathbf{x} : \|\mathbf{x}\| = 1, x_1 > 1/3\}$, $\{\mathbf{x} : \|\mathbf{x}\| = 1, x_2 < 0.5\}$, etc.)