CIS5200: Machine Learning Spring 2023

Homework 1

Release Date: January 23, 2023 *Due Date: February 3, 2023*

- HW1 will count for 10% of the grade. This grade will be split between the written (40 points) and programming (30 points) part.
- All written homework solutions are required to be formatted using LaTeX. Please use the template here. Do not modify the template. This is a good resource to get yourself started, if you are still not comfortable with using LaTeX.
- You will submit your solution for the written part of HW1 as a single PDF file via Gradescope. The deadline is **11:59 PM ET**. Contact TAs on Ed if you face any issues uploading your homeworks.
- Collaboration is permitted and encouraged for this homework, though each student must understand, write, and hand in their own submission. In particular, it is acceptable for students to discuss problems with each other; it is not acceptable for students to look at another student's written answers when writing their own. It is also not acceptable to publicly post your (partial) solution on Ed, but you are encouraged to ask public questions on Ed. If you choose to collaborate, you must indicate on each homework with whom you collaborated.

Please refer to the notes and slides posted on the website if you need to recall the material discussed in the lectures.

1 Written Questions (40 points)

Problem 1: Margin Perceptron (15 points)

Recall the Perceptron algorithm we saw in the lecture. The Perceptron algorithm terminates once it classifies all points correctly. It does not guarantee that the hyperplane that that it finds has large margin (γ) despite our assumption that the true hyperplane w_* has margin γ where $\gamma = \min_{i \in \{1,\dots,m\}} \left| w_*^\top x_i \right|$.

In this problem, we will consider the following simple modification to the Perceptron algorithm:

Algorithm 1: Margin Perceptron

Initialize $w_1 = 0 \in \mathbb{R}^d$
for $t = 1, 2, \dots$ **do**
  if $\exists i \in \{1, \dots, m\}$ *s.t.* $y_i \neq \text{sign}\left(w_t^\top x_i\right)$ *or* $|w_t^\top x_i| \leq 1$ **then** update $w_{t+1} = w_t + y_i x_i$
  else output w_t
end

We will show that Margin Perceptron stops after $9/\gamma^2$ steps and returns a hyperplane w such that

$$\min_{i\in\{1,\ldots,m\}} \frac{|w^\top x_i|}{\|w\|_2} \geq \gamma/3.$$

Note that the margin is the distance of the closest point to the hyperplane, and since $\|w\|_2$ is not necessarily norm 1, this quantity is given by $\min_{i\in\{1,\ldots,m\}} \frac{|w^\top x_i|}{\|w\|_2}$.

As in the lecture, we will assume that $\|x_i\|_2 \leq 1$ for all $i \in \{1,\ldots,m\}$ and $\|w_*\|_2 = 1$.

1.1 (2 points) Show that after every round t, we have

$$w_*^\top w_{t+1} \geq w_*^\top w_t + \gamma.$$

1.2 (4 points) Show that after every round t, we have

$$\|w_{t+1}\|_2^2 \leq \|w_t\|_2^2 + 3.$$

1.3 (3 points) Using the above two parts, show that after T rounds,

$$\gamma T \leq \|w_{T+1}\|_2 \leq \sqrt{3T}.$$

Hint: Use the Cauchy-Schwarz Inequality: $a^\top b \leq \|a\|\|b\|$.

1.4 (1 point) Use 1.3, to conclude that $T \leq 9/\gamma^2$.

1.5 (4 points) Show that the output hyperplane w satisfies

$$\min_i \frac{|w^\top x_i|}{\|w\|_2} \geq \frac{\gamma}{3}.$$

Hint: You will need to use the results in 1.2 and 1.3 plus the stopping condition of the algorithm.

1.6 (1 point) Why is it desirable to learn a predictor that has margin?

Problem 2: Bayes Optimal Classifier (15 points)

Let $\eta(x)$ denote the conditional probability of the label being 1 given a point x under the distribution $\mathcal{D}$. That is

$$\eta(x) = \Pr[y = 1|x].$$

Recall that the true risk, under the 0/1 loss, for any classifier f is

$$R(f) = \mathop{\mathbb{E}}_{x,y} \left[\mathbb{1}[f(x) \neq y]\right].$$

The *Bayes optimal classfier* w.r.t. $\mathcal{D}$ is the classifier f_* that achieves the minimum risk among all possible classifiers. In this problem, we will work out what the Bayes optimal classifier is.

2.1 (3 points) Show that

$$R(f) = \mathbb{E}_x \left[\eta(x)\mathbb{1}[f(x) = -1] + (1 - \eta(x))\mathbb{1}[f(x) = 1]\right].$$

Hint: Use the fact that $\mathbb{E}_{x,y}[\cdot] = \mathbb{E}_x \mathbb{E}_{y|x}[\cdot]$.

2.2 (3 points) Use the above to show that the minimum risk possible is

$$R(h_*) = \min_f R(f) = \mathbb{E}_x \left[\min(\eta(x), 1 - \eta(x))\right]$$

Hint: For a fixed x*, think about what the minimum loss is using 2.1.*

2.3 (2 points) Show that the Bayes optimal classifier that achieves the above loss is

$$f_*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2, \\ -1 & \text{otherwise..} \end{cases}$$

2.4 (1 point) Derive the Bayes optimal classifier under the logistic model

$$\eta(x) = \frac{1}{1 + \exp(-w^\top x)}.$$

2.5 (6 points) Now suppose we modify the loss function from 0/1 to the following cost-based loss function

$$\ell_c(\hat{y}, y) = \begin{cases} c & \text{if } y = 1, \hat{y} = -1 \\ 1 - c & \text{if } y = -1, \hat{y} = 1 \\ 0 & \text{if } y = \hat{y}. \end{cases}$$

Here the loss penalizes false negative with cost c and false positive with cost $1 - c$, penalizing different types of mistakes differently.[1]

Note that the true risk under this loss is

$$R_c(f) = \mathop{\mathbb{E}}_{x,y} \left[\ell_c(f(x), y)\right].$$

Find the Bayes optimal classifier in this setting.

Hint: Follow the same ideas you used to solve 2.1-2.3 using ℓ_c *instead of 0/1 loss.*

[1] Let us see why this is a useful loss function. Consider the case of medical diagnosis, high false negative rate means that we are predicting that patients do not have the disease when they actually do. Such a prediction could lead to the patient not getting the care they need. In such a setting, you would want c to be closer to 1.

Problem 3: MLE for Linear Regression (10 points)

Similar to the derivation of the logistic loss in the lecture using maximum (conditional) likelihood estimation, here we will derive the squared loss for linear regression.

Assume that for given x, the label y is generated randomly as

$$y = w^\top x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is some normally distributed noise with mean 0 and variance $\sigma^2 > 0$.

3.1 (3 points) Show that the above model implies that the conditional density of y given x is

$$P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - w^\top x)^2}{2\sigma^2}\right).$$

Hint: Use the density function of the normal distribution.

3.2 (2 points) Show that the risk of the predictor $f(x) = \mathbb{E}[y|x]$ is σ^2, that is,

$$R(f) = \mathop{\mathbb{E}}_{x,y}\left[(y - f(x))^2\right] = \sigma^2.$$

3.3 (3 points) Recall that the conditional likelihood for the given data $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ is

$$\hat{L}(w, \sigma) = P(y_1, \ldots, y_m | x_1, \ldots, x_m) = \prod_{i=1}^{m} P(y_i|x_i).$$

Compute the log conditional likelihood, that is, $\log \hat{L}(w, \sigma)$.

3.4 (2 points) Show that the maximizer of $\log \hat{L}(w, \sigma)$ is the minimizer of the empirical risk with squared loss, $\hat{R}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2$.

2 Programming Questions (30 points)

Use the link here to access the Google Colaboratory (Colab) file for this homework. Be sure to make a copy by going to "File", and "Save a copy in Drive". This assignment uses the PennGrader system for students to receive immediate feedback. As noted on the notebook, please be sure to change the student ID from the default '99999999' to your 8-digit PennID.

Instructions for how to submit the programming component of HW 1 to Gradescope are included in the Colab notebook. You may find this PyTorch reference to be helpful - if you get stuck, it may be helpful to review some of the PyTorch documentation and functions.