CIS5200: Machine Learning Spring 2025

Homework 2

Release Date: February 14, 2025 *Due Date: February 28, 2025*

- HW2 will count for 10% of the grade. This grade will be split between the written (30 points) and programming (20 points) parts.
- 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 more familiar with LaTeX, if you are still not comfortable.
- You will submit your solution for the written part of HW2 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 Solutions 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 (35 points)

Problem 1: Gradient Descent (25 points)

Consider a training dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where for all $i \in [m]$, $x_i \in \mathbb{R}^d$, $\|x_i\|_2 \leq 1$ and $y_i \in \{-1, 1\}$. Suppose we want to run regularized logistic regression, that is, solve the following optimization problem: for regularization term $t(w)$,

$$\min_w \underbrace{\frac{1}{m}\sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i w^\top x_i\right)\right)}_{\hat{R}(w)} + t(w)$$

Recall: For showing that a twice differentiable function f is μ-strongly convex, it suffices to show that the hessian satisfies: $\nabla^2 f \succeq \mu I$. Similarly to show hat a twice differentiable function f is L-smooth, it suffices to show that the hessian satisfies: $LI \succeq \nabla^2 f$. Here I is the identity matrix of the appropriate dimension.

Unregularized Case: Let us first analyze the unregularized case where $t(w) = 0$ for all $w \in \mathbb{R}^d$.

1.1 (3 points) Is the unregularized objective $\hat{R}(w)$ strongly convex? Explain your answer.

1.2 (3 points) Show that $\hat{R}(w)$ is 1-smooth.

1.3 (4 points) What is the largest learning rate that you can choose such that when we run gradient descent on $\hat{R}(w)$, the objective value is non-increasing at each iteration? Explain your answer.

Hint: The answer is not $1/L$ for a L-smooth function.

1.4 (2 points) What is the convergence rate of gradient descent on the unregularized problem? In other words, suppose we want to achieve $\hat{R}(w_{T+1}) - \hat{R}(w_*) \leq \epsilon$ where w_* is the global minimizer of $\hat{R}$, express the number of iterations T that we need to run gradient descent for.

Note: You do not need to reprove the convergence guarantee, just use the guarantee from the lecture notes to provide the rate.

Unregularized Case: Consider the following variation of the ℓ_2 norm regularizer called the weighted ℓ_2 norm regularizer: for $\lambda_1, \ldots, \lambda_d \geq 0$,

$$t(w) = \sum_{j=1}^{d} \lambda_j w_j^2.$$

1.5 (5 points) Show that the regularized objective $\hat{R}(w) + t(w)$ is μ-strongly convex and L-smooth for $\mu = 2 \min_{j \in [d]} \lambda_j$ and $L = 1 + 2 \max_{j \in [d]} \lambda_j$.

1.6 (4 points) If a function is μ-strongly convex and L-smooth, after T iterations of gradient descent we have:

$$\|w_{T+1} - w_*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|w_1 - w_*\|_2^2.$$

Using the above, what is the convergence rate of gradient descent on the regularized problem? In other words, suppose we want to achieve $\|w_{T+1} - w_*\|_2 \leq \epsilon$ where w_* is the global minimizer of the regularized objective, express the number of iterations T that we need to run GD.

Note: You do not need to prove the given convergence guarantee, just use the guarantee from the lecture notes to provide the rate.

1.7 (4 points) For simplicity, let us assume that $\lambda_1 = \lambda_2 = \ldots = \lambda_d = \lambda$ for some $\lambda \geq 0$. Use 1.6 to show that gradient descent on the regularized logistic regression problem converges faster as λ increases. Does this mean we should always use the largest possible λ?

Problem 2: 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 $x \in \mathbb{R}^d$, the label y is generated randomly as

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

for some fixed $w \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is some normally distributed noise with mean 0 and variance $\sigma^2 > 0$.

2.1 (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.

2.2 (2 points) Show that the risk of the predictor $f(x) = \mathbb{E}[y|x]$, that is,

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

2.3 (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, $\log \hat{L}(w, \sigma)$.

Hint: Use your expression for $p(y|x)$ from part 2.1.

2.4 (2 points) Show that the maximizer of $\log \hat{L}(w, \sigma)$ is the same as the minimizer of the empirical risk with squared loss, $\hat{R}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2$.

Hint: Take the derivative of your result from 2.3 and set it equal to zero.

2 Programming Questions (21 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". As with the previous homeworks, 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 2 to Gradescope are included in the Colab notebook. You may find this PyTorch linear algebra reference and this general PyTorch reference to be helpful in perusing the documentation and finding useful functions for your implementation.