CSE517A – Homework 1

M. Neumann

Jan 31 2025

- Please keep your written answers brief and to the point. Incorrect or rambling statements can hurt your score on a question.
- If your hand writing is not readable, we **cannot give you credit**. We recommend you type your solutions in LaTeX and compile a .pdf for each answer. **Start every problem on a new page!**
- This will be due **Jan 31 2025** at **11:59pm**, with an automatic 3-day extension (cf. course syllabus for more information).
- You may work in groups of at most 2 students.
- Submission instructions:
  - Start every problem on a **new page**.
  - Submissions will be exclusively accepted via **Gradescope**.
  - You will need to **match your pages** to the problems in Gradescope. Not matching the pages will result in a score penalty.
  - Find more instructions on how to get your Gradescope account and submit your work on the course webpage.

Note, that if you use *any* **resources outside the course materials** to derive (part of) your solution – this includes **AI tools and chat bots** –, you will need to cite the source in your homework submission. This also holds for **online sources** such as StackOverflow or Wikipedia. If you collaborate with anyone other than your partner, it is your responsibility to indicate this in your submission. **Course materials** (that do not need to be cited) are the lecture notes, course books, and resources linked therein or on Canvas, plus course materials of any prerequisite course officially listed as such in the course listing.

Citing the source(s) does **not** legitimate the *copying* of existing solutions to any given problem, neither does it legitimate that another person (that is not you or your partner) directly solves any (part of the) problems for you. Please, refer to the **course syllabus** for more details.

Problem 1 (*25 points*) Warm-up: Practicing Matrix Notation

Read **[FCML]** Chapter 1.3 Vector/Matrix Notation. The following subproblems follow **[FCML]** 1.7 Exercises. Note that we use the *slightly modified notation* from our lectures here.
We have $i = 1, \ldots, n$ data points and a 2-dimensional input space ($d = 2$):

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix}, X = \begin{bmatrix} x_{11} & x_{12} & \cdots x_{1n} \\ x_{21} & x_{22} & \cdots x_{2n} \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Note, in this problem we ignore the bias term w_0 for simplicity.

(a) (*5 pts*) Show that

$$\mathbf{w}^\top X X^\top \mathbf{w} = w_1^2 \left(\sum_{i=1}^{n} x_{1i}^2 \right) + 2 w_1 w_2 \left(\sum_{i=1}^{n} x_{1i} x_{2i} \right) + w_2^2 \left(\sum_{i=1}^{n} x_{2i}^2 \right)$$

Hint: it's easiest to do $XX^\top$ first.

(b) (*5 pts*) Show that $(X^\top \mathbf{w})^\top = \mathbf{w}^\top X$ by multiplying out both sides.

(c) (*5 pts*) Show that $\sum_i \mathbf{x}_i y_i = Xy$. Hint: remember that when multiplying a scalar by a vector (or matrix), we multiply each element of the vector (or matrix) by that scalar.

(d) (*5 pts*) Show that $\sum_i \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w} = XX^\top \mathbf{w}$. Hint: reuse your computations from part (a).

(e) (*5 pts*) Show that the following notations for the ordinary least squares (OLS) objective function are equivalent:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 = (X^\top \mathbf{w} - \mathbf{y})^\top (X^\top \mathbf{w} - \mathbf{y})$$

Reminder: stick with $d = 2$ for now.

Problem 2 (*20 points*) Loss Function Optimization via Gradient Descent

Derive the gradient update (with step size c) for your weight vector $\mathbf{w}$ for each of the following loss functions. Treat the input space dimension d as arbitrary. Ignore the bias term w_0 for simplicity. Note: $\|\mathbf{w}\|_2^2 = \mathbf{w}^\top \mathbf{w}$, $\|\mathbf{w}\|_1 = \sum_{\alpha=1}^{d} |w_\alpha|$, and λ and C are non-negative constants.

(a) (*5 pts*) Ridge Regression (*cf. implementation project 1*)

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

Hint: try to use matrix notation. Use the identities shown in **[FCML]** Table 1.4.

(b) (*5 pts*) Lasso Regression:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_1$$

Hint: try to use matrix notation. Use the identities shown in **[FCML]** Table 1.4.

(c) (*5 pts*) Logistic Regression ($y_i \in \{+1, -1\}$, *cf. implementation project 1*)

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \mathbf{w}^\top \mathbf{x}_i\right)\right)$$

(d) (*5 pts*) Linear Support Vector Machine ($y_i \in \{+1, -1\}$, *cf. implementation project 1*)

$$\mathcal{L}(\mathbf{w}) = C \sum_{i=1}^{n} \max\{1 - y_i \mathbf{w}^\top \mathbf{x}_i,\ 0\} + \|\mathbf{w}\|_2^2$$

Hint: This is equivalent to using the following expression (use this in *implementation project 1*):

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \max\{1 - y_i \mathbf{w}^\top \mathbf{x}_i,\ 0\} + \lambda \|\mathbf{w}\|_2^2$$

Problem 3 (*35 points*) Logistic Regression and Newton's Method

Let us re-visit Logistic Regression, however with $y_i \in \{0, 1\}$.

(a) (*7 pts*) Show that with these new labels the objective function can be written as

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{n} \Big(y_i \log\left(\text{sigm}(\mathbf{w}^\top \mathbf{x}_i)\right) + (1 - y_i) \log\left(1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_i)\right) \Big),$$

where $\text{sigm}(a) = \frac{1}{1+e^{-a}}$ is the *sigmoid* or *logistic* function.

(b) (*7 pts*) Show that the gradient of $\mathcal{L}$ can be written as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = -\sum_{i=1}^{n} \left(y_i - \text{sigm}(\mathbf{w}^\top \mathbf{x}_i)\right) \mathbf{x}_i.$$

(HINT: You can use the fact that $\frac{\partial \text{sigm}(z)}{\partial z} = \text{sigm}(z)(1 - \text{sigm}(z))$.)

(c) (*7 pts*) Let the $n \times n$ diagonal matrix $W_{ii} = \text{sigm}(\mathbf{w}^\top \mathbf{x}_i)(1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_i))$ and let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$. Show that the Hessian matrix is $H = XWX^\top$. For which examples is W_{ii} large, for which is it small?

(d) (*7 pts*) Using the result form the previous part, show that the Hessian H is positive-semi definite.

(e) (*7 pts*) Write down the update rule for a newton step. Show that if you use the substitution $\mathbf{z}$ where $z_i = \mathbf{x}_i^\top \mathbf{w} + \frac{1}{W_{ii}}(y_i - \text{sigm}(\mathbf{w}^\top \mathbf{x}_i))$, you arrive at

$$\mathbf{w}_{\text{new}} \leftarrow (XWX^\top)^{-1} XW\mathbf{z}.$$

Problem 4 (*10 points*) Weighted Ridge Regression

Assume that in addition to your data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ you also have weights $p_i \geq 0$ for each example. Let your loss function be

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} p_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \mathbf{w}^\top \mathbf{w}$$

(a) (*2 pts*) Rephrase the previous equation in terms of the matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{y} = [y_1, \ldots, y_n]^\top$ and the diagonal matrix $P = \text{diag}([p_1, \ldots, p_n])$, where the diag operator creates an $n \times n$ matrix with $p_1, \ldots, p_n$ on the diagonal and zeros on all off-diagonal entries.

(b) (*4 pts*) Derive a closed form solution for $\mathbf{w}$. Hint: use $\frac{\partial(\mathbf{w}^\top B \mathbf{w})}{\partial \mathbf{w}} = B\mathbf{w} + B^\top \mathbf{w}$ and $\mathbf{w}^\top \mathbf{w} = \mathbf{w}^\top I \mathbf{w}$ where I is the identity matrix.

(c) (*4 pts*) Let $\lambda = 0$. Look at the result of **Problem 3(e)** Newton's Method. Hold your breath and enjoy the moment of amazement. Why is that algorithm also called *Iteratively Reweighted Least Squares*?

Problem 5 (*10 points*) Critical Reflection

The following questions have no "right" or "wrong" answers; just explain your thought process for any decisions you make.

(a) (*5 pts*) The labels chosen for binary classification is often arbitrary, i.e., it isn't essential to the problem for emails to be labeled $y \in \{0, 1\}$ or $y \in \{-1, +1\}$. However, it can have an impact on the math, as you can see in the gradients from Problem 2(c) and Problem 3(b). Explain why you think it was or wasn't a good choice to use $y \in \{-1, +1\}$ for implementation project 1.

(b) (*5 pts*) For binary classification, we consider zero-one loss to be the "true" classification loss, and all other loss functions are surrogates that are used for practical reasons like being differentiable. What would you consider to be "true" regression loss? Explain your answer.