EECS545 Machine Learning
Homework #1 Solutions

2024/2/3

1 [31 points] Derivation and Proof

(a) [**8 points**] Derive the solution for w_0 and w_1.

The loss function is:

$$L(w_0, w_1) = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - w_0 - w_1 x^{(i)})^2$$

We will compute the gradient $\frac{\partial L}{\partial w_0}$ and $\frac{\partial L}{\partial w_1}$ and use the fact that the optimal solution(s) will make the gradient zero.

The gradient for w_0 is:

$$\frac{\partial L}{\partial w_0} = -\sum_{i=1}^{N}(y^{(i)} - w_0 - w_1 x^{(i)}).$$

To make $\partial L/\partial w_0 = 0$, w_0 should satisfy $\sum_{i=1}^{N}(y^{(i)} - w_0 - w_1 x^{(i)}) = \sum_{i=1}^{N} y^{(i)} - N w_0 - w_1 \sum_{i=1}^{N} x^{(i)} = 0$.

Therefore, the solution is

$$w_0 = \frac{\sum_{i=1}^{N} y^{(i)}}{N} - w_1 \frac{\sum_{i=1}^{N} x^{(i)}}{N} = \overline{Y} - w_1 \overline{X}.$$

The gradient for w_1 is:

$$\frac{\partial L}{\partial w_1} = -\sum_{i=1}^{N} x^{(i)}(y^{(i)} - w_0 - w_1 x^{(i)}).$$

To make $\partial L/\partial w_1 = 0$, we need $\sum_{i=1}^{N}(x^{(i)} y^{(i)} - w_0 x^{(i)} - w_1 (x^{(i)})^2) = 0$.

Replace w_0 with the solution we computed earlier: $w_0 = \overline{Y} - w_1 \overline{X}$:

$$\sum_{i=1}^{N}(x^{(i)} y^{(i)} - (\overline{Y} - w_1 \overline{X}) x^{(i)} - w_1 (x^{(i)})^2) = 0$$

$$\iff \sum_{i=1}^{N}(x^{(i)} y^{(i)} - \overline{Y} x^{(i)}) - w_1 \sum_{i=1}^{N}((x^{(i)})^2 - \overline{X} x^{(i)}) = 0$$

Therefore, the solution is

$$w_1 = \frac{\sum_{i=1}^{N}(x^{(i)} y^{(i)} - \overline{Y} x^{(i)})}{\sum_{i=1}^{N}((x^{(i)})^2 - \overline{X} x^{(i)})} = \frac{\sum_{i=1}^{N} x^{(i)} y^{(i)} - N\overline{YX}}{\sum_{i=1}^{N}(x^{(i)})^2 - N\overline{X}^2} = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)} y^{(i)} - \overline{YX}}{\frac{1}{N}\sum_{i=1}^{N}(x^{(i)})^2 - \overline{X}^2}.$$

□

(b) **[14 points]**

i. [6 points] Prove A is PD if and only if $\lambda_i > 0$ for each i.
($\Longrightarrow$) Assume A is a positive definite matrix, i.e, $\mathbf{x}^\top A\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$.
We want to show all eigenvalues λ_i are positive.
Given the eigendecomposition $A = \mathbf{U}\Lambda\mathbf{U}^\top$ where for each row vector in $\mathbf{u}_i$ in $\mathbf{U}$, we have $\lambda_i\mathbf{u}_i = A\mathbf{u}_i$ (definition of eigenvector) and $\lambda_i = \lambda_i\mathbf{u}_i^\top\mathbf{u}_i$ because $\mathbf{u}_i^\top\mathbf{u}_i = 1$. Then,

$$\lambda_i = \mathbf{u}_i^\top(\lambda_i\mathbf{u}_i) = \mathbf{u}_i^\top(A\mathbf{u}_i) > 0.$$

($\Longleftarrow$) Assume $\lambda_i > 0$ for any i. We will show that $z^\top Az > 0$ for any $z \neq \mathbf{0}$:
For any non-zero vector $z \neq \mathbf{0}$,

$$z^\top Az = z^\top(\mathbf{U}^\top\Lambda\mathbf{U})z = z^\top\left(\sum_{i=1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top\right)z = \sum_{i=1}^{d}\lambda_i(z^\top\mathbf{u}_i)(\mathbf{u}_i^\top z) = \sum_{i=1}^{d}\lambda_i(z^\top\mathbf{u}_i)^2 > 0.$$

Therefore A is a positive definite matrix. □

ii. [8 points] Consider the real symmetric matrix $\Phi^\top\Phi$. With any $z \in \mathbb{R}^d$,

$$z^\top\Phi^\top\Phi z = (\Phi z)^\top(\Phi z) \geq 0,$$

therefore $\Phi^\top\Phi$ is PSD. Its eigendecomposition becomes: $\Phi^\top\Phi = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top$.
(i) With ridge regression, we consider the matrix $\Phi^\top\Phi + \beta\mathbf{I}$ and try to derive an eigendecomposition form:

$$\begin{aligned}
&\Phi^\top\Phi + \beta\mathbf{I} \\
&= \sum_{i=1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top + \beta\mathbf{I} && \because \text{eigendecomposition of } \Phi^\top\Phi \\
&= \sum_{i=1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top + \beta\sum_{i=1}^{d}\mathbf{u}_i\mathbf{u}_i^\top && \because \sum_{i=1}^{d}\mathbf{u}_i\mathbf{u}_i^\top = \mathbf{I} \\
&= \sum_{i=1}^{d}(\lambda_i + \beta)\mathbf{u}_i\mathbf{u}_i^\top.
\end{aligned}$$

□

(ii) If $\beta > 0$, ridge regression makes the matrix $\Phi^\top\Phi + \beta\mathbf{I}$ positive definite because all $(\lambda_i + \beta)$ are positive and applying (i). Note that $\lambda_i > 0$ holds because we already showed that $\Phi^\top\Phi$ is PSD. [Remark: Without showing that $\Phi^\top\Phi$ is PSD or that $\lambda_i > 0$, the proof becomes incomplete.] □

(c) **[9 points]** The log-likelihood of the data can be re-written as:

$$\sum_n\Big(\mathbb{I}(y^{(n)} = 1)\log P(y^{(n)} = 1 \mid \mathbf{x}^{(n)}) + \mathbb{I}(y^{(n)} = -1)\log P(y^{(n)} = -1 \mid \mathbf{x}^{(n)})\Big)$$

First, we plug in the definition class posterior probability of logistic regression model:
$P(y^{(n)} = 1 \mid \mathbf{x}^{(n)}) = \dfrac{1}{1 + \exp(-\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))}$ then we have:

$$\sum_n \mathbb{I}(y^{(n)} = 1)\log\left[\frac{1}{1 + \exp(-\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))}\right] + \mathbb{I}(y^{(n)} = -1)\log\left[1 - \frac{1}{1 + \exp(-\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))}\right]$$

In other words, this term can be written in case by case as follows:

- if $y^{(n)} = 1$, then

  $$\log\left[\frac{1}{1+\exp(-\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))}\right] = -\log\left[1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))\right]$$

- if $y^{(n)} = -1$, then

  $$\log\left[1-\frac{1}{1+\exp(-\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))}\right] = \log\left[\frac{1}{1+\exp(\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))}\right] = -\log\left[1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))\right]$$

Therefore, the objective can be compactly written as

$$\sum_n -\log\left[1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))\right].$$

(we want to **maximize** the class posterior probability above, so it is same as **minimize** the class posterior of the following equation)

$$\sum_n \log\left[1+\exp(-y^{(n)}\mathbf{w}^\top\phi(\mathbf{x}^{(n)}))\right].$$

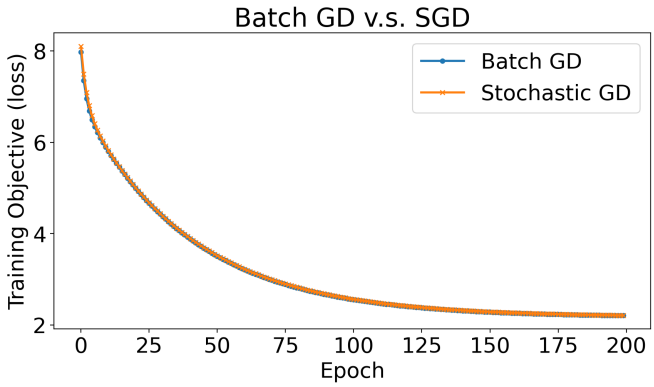
□

2 [39 points] Linear regression on a polynomial

2.1 GD and SGD

(a) **[12 points]** Autograder.

(b) **Answer [3 points]:** The following image is not required, just instructive. As long as the answer is reasonable and consistent with the output from your code, we will not penalize or deduct any points.



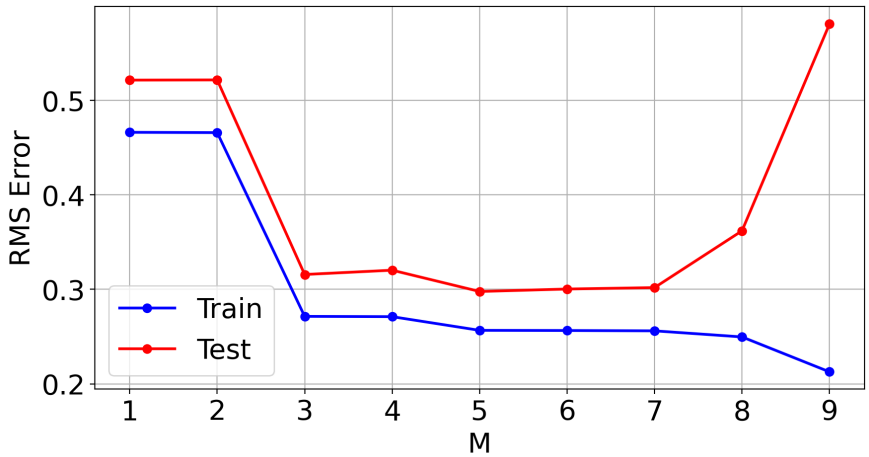
Both GD and SGD shows similar test objective $E(\mathbf{w}_{\text{test}})$, but SGD shows slightly smaller value (2.7017 from GD and 2.6796 from SGD).

However, GD is much faster than SGD in our environment (0.00 seconds vs 0.06 seconds). We believe it is because GD can compute the weight for each data point in parallel. The trend may be different if the dataset is much larger than the current dataset (e.g., more than 1M data samples).

2.2 Over-fitting study

(a) **[8 points]**: Autograder.

(b) **Answer [2 points]:** The plot should look like this:



(c) **Answer [2 points]:** For the figure above, $M = 5$ minimizes the test error (while also performing reasonably well on training error).

2.3 Regularization (Ridge Regression)

(a) **[8 points]**: Autograder.

(b) **[2 points]** The following plot:

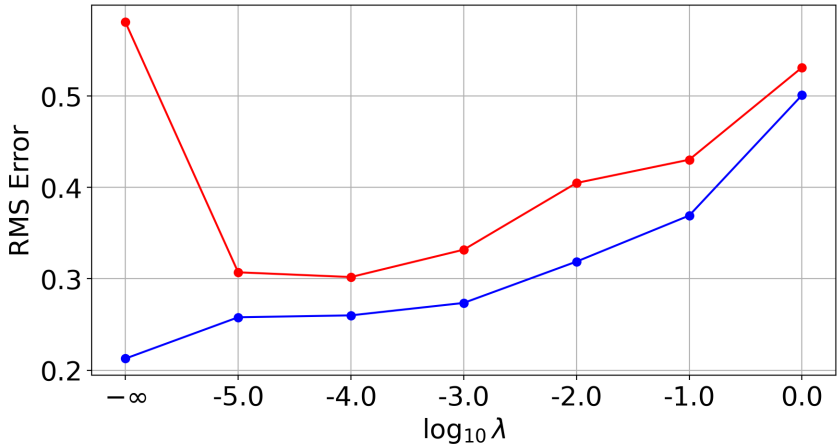


Figure 1: x axis is $\log_{10}(\lambda)$ value and y axis is error on the training and test dataset.

(c) **Answer [2 points]:** $\lambda = 10^{-4}$ (or $\lambda = 10^{-5}..10^{-4}$) seems to be the "sweet spot" for this particular problem, as this range of λ minimizes the test error.

3 [30 points] Locally weighted linear regression

(a) **[3 points]** Derive $(\mathbf{w}^\top X - \mathbf{y}^\top)R(\mathbf{w}^\top X - \mathbf{y}^\top)^\top$.

Let $\mathbf{z} = \mathbf{w}^\top X - \mathbf{y}^\top$, i.e. $z_i = \mathbf{w}^\top \mathbf{x}^{(i)} - y^{(i)}$. Then we have:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} r^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} - y^{(i)})^2 \tag{1}$$

$$= \sum_{i=1}^{N} \frac{1}{2} r^{(i)} z_i^2 \tag{2}$$

$$= \mathbf{z}R\mathbf{z}^\top \tag{3}$$

$$= (\mathbf{w}^\top X - \mathbf{y}^\top)R(\mathbf{w}^\top X - \mathbf{y}^\top)^\top \tag{4}$$

where $R_{(i,i)} = \frac{1}{2}r^{(i)}$, $R_{(i,j)} = 0$ for $i \neq j$.

(b) **[7 points]** Derive the closed-form solution:

$$\nabla_{\mathbf{w}} E_D(\mathbf{w}) = \nabla_{\mathbf{w}}(\mathbf{w}^\top XRX^\top \mathbf{w} - \mathbf{w}^\top XR\mathbf{y} - \mathbf{y}^\top RX^\top \mathbf{w} + \mathbf{y}^\top R\mathbf{y}) = 2XRX^\top \mathbf{w} - 2XR\mathbf{y}$$

So, $\nabla_{\mathbf{w}} E_D(\mathbf{w}) = 0$ when

$$XRX^\top \mathbf{w} = XR\mathbf{y}$$

These are the normal equations, from which we can get a closed form formula for $\mathbf{w}$:

$$\mathbf{w} = (XRX^\top)^{-1}XR\mathbf{y}$$

Note: R is a diagonal matrix. See (a) for the details.

(c) **Answer [8 points]:**

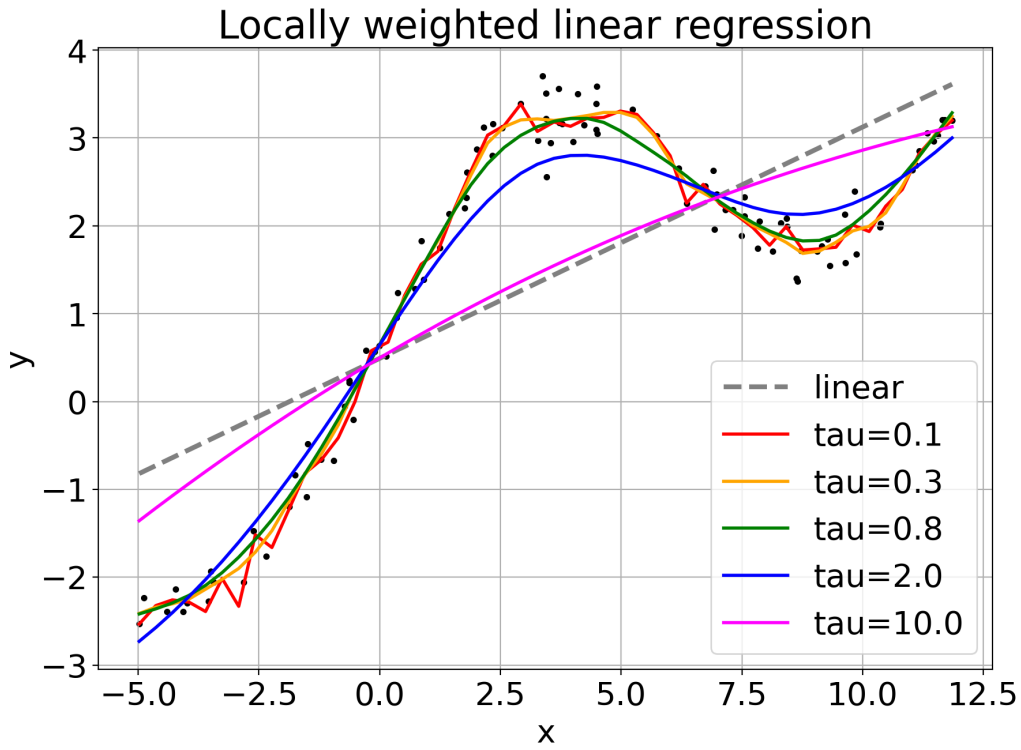
$$\begin{aligned}
\arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w}) &= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log p(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \frac{(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2}{2(\sigma^{(i)})^2} \\
&= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{N} \frac{(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2}{(\sigma^{(i)})^2} \\
&= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{N} \frac{1}{(\sigma^{(i)})^2}(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2 \\
&= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{N} r^{(i)}(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2
\end{aligned}$$

where in the last step, we substituted $r^{(i)} = \frac{1}{(\sigma^{(i)})^2}$ to get the linear regression form.

(d) **[12 points]: Programming**

i. **[8 pts]** Autograder.

ii. **[2 pts]** Plot.



iii. **[2 pts]** Discussion.

For small bandwidth parameter τ, the fitting is dominated by the closest by training samples. The smaller the bandwidth, the less training samples that are actually taken into account when doing the regression, and the regression results thus become very susceptible to noise in those few training samples. For larger τ, we have enough training samples to reliably fit straight lines, unfortunately a straight line is not the right model for these data, so we also get a bad fit for large bandwidths.