# CS 446 / ECE 449 — Homework 4

*yiminc2*

**Instructions.**

- Homework is due **Friday, Oct 31**, at 11:59 **PM** CST; you have **3** late days in total for **all Homeworks**.
- Everyone must submit individually at Gradescope under `HW4` and `HW4 - Programming Assignment`.
- The "written" submission at `HW4` **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use LaTeX, markdown, google docs, MS word, whatever you like; but it must be typed!
- When submitting at `HW4`, Gradescope will ask you to **mark out boxes around each of your answers**; please do this precisely!
- Please make sure your NetID is clear and large on the first page of the homework.
- Your solution **must** be written in your own words. Please see the course webpage for full **academic integrity** information. You should cite any external reference you use.
- We reserve the right to reduce the auto-graded score for `HW4 - Programming Assignment` if we detect funny business (e.g., your solution lacks any algorithm and hard-codes answers you obtained from someone else, or simply via trial-and-error with the autograder).
- When submitting to `HW4 - Programming Assignment`, only upload `hw4_q3.py` and `hw4_utils.py`. Additional files will be ignored.

## 1. Bias-Variance in Ridge Regression. (23 pt)

Recall from the lecture, the Expected Test Error can be decomposed as follows:

$$\mathbb{E}_{x,y,\mathcal{D}}[(h_{\mathcal{D}}(x)-y)^2] = \underbrace{\mathbb{E}_{x,\mathcal{D}}[(h_{\mathcal{D}}(x)-\bar{h}(x))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_{x}[(\bar{h}(x)-\bar{y}(x))^2]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_{x,y}[(\bar{y}(x)-y)^2]}_{\text{Noise}}$$

Consider fixed (non-random) scalar features $\{x^{(i)}\}_{i=1}^N$. The labels are generated as $y^{(i)} = w^* x^{(i)} + \epsilon^{(i)}$ where $w^*$ is fixed and $\epsilon^{(i)}$ are i.i.d noises from Gaussian distribution $N(0, \sigma^2)$. Note that $w^*$ is unknown and $\epsilon^{(i)}$ is independent of $x^{(i)}$. Therefore, we can define the observed dataset as $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$.

Ridge regression optimizes the following objective for a dataset $\mathcal{D}$ with $\lambda \geq 0$:

$$w_{\mathcal{D}} = \arg\min_{w} \frac{1}{N}\sum_{i=1}^{N}(wx^{(i)} - y^{(i)})^2 + \lambda w^2$$

For simplicity, the intercept term is omitted from this problem. The closed-form solution of ridge regression is given as:

$$w_{\mathcal{D}} = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}y^{(i)}}{\lambda + \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}}$$

(a) Consider the expected label $\bar{y}(x) = \mathbb{E}_{y|x}[y]$. Show that $\bar{y}(x) = w^* x$. Similarly, consider the noise term:

$$\text{Noise} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y^{(i)}|x^{(i)}}[(\bar{y}(x^{(i)}) - y^{(i)})^2]$$

Show that Noise $= \sigma^2$. (3 pt)

(b) From the lecture, given a machine learning algorithm $\mathcal{A}$, then $h_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$. For our case, $h_{\mathcal{D}}(x) = w_{\mathcal{D}}x$. Consider the expected predictor $\bar{h} = \mathbb{E}_{\mathcal{D}\sim P^N}[h_{\mathcal{D}}]$, then in our case $\bar{w} = \mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}]$. Let $s^2 = \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}$, show that:

$$\bar{w} = \frac{s^2}{\lambda + s^2} w^*$$

(3 pt)

(c) Consider the squared bias term:

$$\text{Bias}^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{w}x^{(i)} - \bar{y}(x^{(i)}))^2$$

Show that:

$$\text{Bias}^2 = \left(\frac{\lambda}{\lambda + s^2}\right)^2 w^{*2} s^2$$

(3 pt)

(d) Consider the variance term:

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)} - \mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}])^2\right]$$

Show that:

$$\text{Variance} = \frac{s^4\sigma^2}{N(\lambda + s^2)^2}$$

(5 pt)

(e) What happens to the Bias$^2$ and Variance term when $\lambda \to 0$ and $\lambda \to \infty$. Your answer should demonstrate that the bias and variance are monotonic with respect to $\lambda$, but in different directions. Therefore, changing $\lambda$ controls the trade-offs. In practice, since we don't know $w^*$ and the true distribution of $\epsilon$, we cannot infer the optimal value of $\lambda$. Therefore, we use model selection to determine the best value for $\lambda$. (3 pt)

(f) Alternatively, we can consider an equivalent form of ridge regression:

$$w_{\mathcal{D}} = \arg\min_{w} \frac{1}{N}\sum_{i=1}^{N}(wx^{(i)} - y^{(i)})^2 \quad \text{so that} \quad w^2 \leq R$$

The regularization constraint forces the weight $w$ to be inside a ball around the origin with radius $\sqrt{R}$. Use the triangle inequality to show that:

$$|w_{\mathcal{D}} - \bar{w}|^2 \leq 4R$$

From there, we can see that the maximum Euclidean distance between any two points in the ball can at most be $2\sqrt{R}$. (3 pt)

(g) Show that ridge regression bounds the variance by $4Rs^2$

$$\text{Variance} \leq 4Rs^2$$

Note that this bound does not depend on $w^*$ or $\epsilon$, but it can be loose compared to the actual value of variance. (3 pt)

---

**Solutions:**

---

(a) i. Prove that $\bar{y}(x) = w^*x$

$$\begin{aligned}
\bar{y}(x) &= \mathbb{E}_{y|x}[y] \\
&= \mathbb{E}_{y|x}[w^*x + \epsilon] \\
&= w^*x + \mathbb{E}_{y|x}[\epsilon] \quad (\epsilon \sim N(0, \sigma^2)) \\
&= w^*x
\end{aligned}$$

ii. Prove Noise $= \sigma^2$

$$\begin{aligned}
Noise &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{y^{(i)}|x^{(i)}}[(\bar{y}(x^{(i)}) - y^{(i)})^2] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{y^{(i)}|x^{(i)}}[\bar{y}(x^{(i)})^2] - 2\mathbb{E}_{y^{(i)}|x^{(i)}}[\bar{y}(x^{(i)})y^{(i)}] + \mathbb{E}_{y^{(i)}|x^{(i)}}[y^{(i)2}]\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{y^{(i)}|x^{(i)}}[(w^*x^{(i)})^2] - 2\mathbb{E}_{y^{(i)}|x^{(i)}}[w^*x^{(i)}(w^*x^{(i)} + \epsilon)] + \mathbb{E}_{y^{(i)}|x^{(i)}}[(w^*x^{(i)} + \epsilon)^2]\right] \\
&\text{Plug in } \bar{y}(x^{(i)}) = w^*x^{(i)} \quad \mathbb{E}\left[\epsilon^{(i)}\right] = 0 \quad \mathbb{E}\left[(\epsilon^{(i)2}\right] = \sigma^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[(w^*x^{(i)})^2 - 2(w^*x^{(i)})^2 + (w^*x^{(i)})^2 + \sigma^2\right] \\
&= \sigma^2
\end{aligned}$$

(b) Show that $\bar{w} = \frac{s^2}{\lambda + s^2} w^*$

$$
\begin{aligned}
\bar{w} &= \mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)} y^{(i)}}{\lambda + \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}}\right] \quad (s^2 = \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}) \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)} y^{(i)}}{\lambda + s^2}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[\frac{x^{(i)} y^{(i)}}{\lambda + s^2}\right] \quad (y^{(i)} = w^* x^{(i)} + \epsilon^{(i)}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[\frac{x^{(i)} w^* x^{(i)} + x^{(i)} \epsilon^{(i)}}{\lambda + s^2}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[\frac{x^{(i)2} w^*}{\lambda + s^2}\right] + \mathbb{E}_{\mathcal{D}}\left[\frac{x^{(i)} \epsilon^{(i)}}{\lambda + s^2}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[\frac{x^{(i)2} w^*}{\lambda + s^2}\right] + \frac{1}{\lambda + s^2}\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[x^{(i)} \epsilon^{(i)}\right] \quad (\mathbb{E}_{\mathcal{D}}\left[\epsilon^{(i)}\right] = 0) \\
&= w^* \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[\frac{x^{(i)2}}{\lambda + s^2}\right] \\
&= w^* \mathbb{E}_{\mathcal{D}}\left[\frac{s^2}{\lambda + s^2}\right] \\
&= \frac{s^2}{\lambda + s^2} w^*
\end{aligned}
$$

(c) Show that $Bias^2 = \left(\frac{\lambda}{\lambda+s^2}\right)^2 w^{*2} s^2$

$$
\begin{aligned}
Bias^2 &= \frac{1}{N}\sum_{i=1}^{N}(\bar{w}x^{(i)} - \bar{y}(x^{(i)}))^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[(\bar{w}x^{(i)})^2 - 2\bar{w}x^{(i)}\bar{y}(x^{(i)}) + \bar{y}(x^{(i)})^2\right] \\
&= \bar{w}^2\frac{1}{N}\sum_{i=1}^{N}x^{(i)2} - \bar{w}\frac{1}{N}\sum_{i=1}^{N}2x^{(i)}\bar{y}(x^{(i)}) + \frac{1}{N}\sum_{i=1}^{N}\bar{y}(x^{(i)})^2 \\
&= \bar{w}^2 s^2 - 2\bar{w}w^* s^2 + w^{*2}s^2 \\
&= s^2\left(\bar{w}^2 - 2\bar{w}w^* + w^{*2}\right) \\
&= s^2\left(\left(\frac{s^2}{\lambda+s^2}w^*\right)^2 - 2\frac{s^2}{\lambda+s^2}w^{*2} + w^{*2}\right) \\
&= s^2 w^{*2}\left(\frac{s^2}{\lambda+s^2} - 1\right)^2 \\
&= s^2 w^{*2}\left(\frac{-\lambda}{\lambda+s^2}\right)^2 \\
&= \left(\frac{\lambda}{\lambda+s^2}\right)^2 w^{*2}s^2
\end{aligned}
$$

(d) Show that Variance $= \frac{s^4\sigma^2}{N(\lambda+s^2)^2}$

$$
\begin{aligned}
\text{Variance} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)} - \mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}])^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)})^2 - 2(w_{\mathcal{D}}x^{(i)})\mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}] + \mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}]^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)})^2\right] - 2\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)})\mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}]\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}]^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)})^2\right] &= \mathbb{E}_{\mathcal{D}}\left[w_{\mathcal{D}}^2 x^{(i)2}\right] \quad \text{(given } \{x^{(i)}\}_{i=1}^N \text{ is fixed)} \\
&= x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[w_{\mathcal{D}}^2\right] \\
&= x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[\left(\frac{\frac{1}{N}\sum_{i=1}^{N}x^{(i)}y^{(i)}}{\lambda+s^2}\right)^2\right] \\
&= x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[\left(\frac{s^2w^*}{\lambda+s^2}\right)^2\right] \\
&= x^{(i)2}\left(\bar{w}^2 + Var_D[w_D]\right)
\end{aligned}
$$

$$
\begin{aligned}
\because Var_{\mathcal{D}}[w_{\mathcal{D}}] &= \frac{1}{N^2(\lambda+s^2)^2}\sum_{i=1}^{N} Var_{\mathcal{D}}\left[x^{(i)}\epsilon^{(i)}\right] \\
&= \frac{1}{N^2(\lambda+s^2)^2}\sum_{i=1}^{N} x^{(i)2} Var_{\mathcal{D}}\left[\epsilon^{(i)}\right] \quad \left(Var_{\mathcal{D}}\left[\epsilon^{(i)}\right]=\sigma^2\right) \\
&= \frac{1}{N(\lambda+s^2)^2}s^2\sigma^2 \\
\therefore \mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)})^2\right] &= x^{(i)2}\left(\bar{w}^2+\frac{\sigma^2 s^2}{N(\lambda+s^2)^2}\right) \\
\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)})\mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}]\right] &= x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[\frac{s^2w^*}{\lambda+s^2}\mathbb{E}_{\mathcal{D}}\left[\frac{s^2w^*}{\lambda+s^2}\right]\right] \\
&= x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[\frac{s^2w^*}{\lambda+s^2}\right]\mathbb{E}_{\mathcal{D}}\left[\frac{s^2w^*}{\lambda+s^2}\right] \\
&= x^{(i)2}\bar{w}^2 \\
\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}]^2\right] &= x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}\left[\frac{s^2w^*}{\lambda+s^2}\right]\right)^2\right] \\
&= x^{(i)2}\bar{w}^2 \\
Variance &= \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}\left(\bar{w}^2+\frac{\sigma^2 s^2}{N(\lambda+s^2)^2}\right) - 2x^{(i)2}\bar{w}^2 + x^{(i)2}\bar{w}^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}\frac{\sigma^2 s^2}{N(\lambda+s^2)^2} \\
&= \frac{\sigma^2 s^4}{N(\lambda+s^2)^2}
\end{aligned}
$$

(e) What happens to the $Bias^2$ and Variance term when $\lambda \to 0$ and $\lambda \to \infty$?

**Monotonicity**

- $Bias^2$ is increasing in $\lambda$ as both the numerator $\lambda^2$ and the denominator $(\lambda + s^2)^2$ grows as $\lambda$ increases.
- Variance is decreasing in $\lambda$ as only the denominator $N(\lambda + s^2)^2$ grows, which shrinks the value of the variance.

i. When $\lambda \to 0$
$Bias^2 \to 0, \quad Variance \to \frac{\sigma^2}{N}$

ii. When $\lambda \to \infty$
$Bias^2 \to w^{*2}s^2, \quad Variance \to 0$

(f) Prove

$$|w_{\mathcal{D}} - \bar{w}|^2 \leq 4R$$

with triangular inequality.

- Given the interval $\left[-\sqrt{R}, \sqrt{R}\right]$, for every dataset $\mathcal{D}$, the optimizer $w_{\mathcal{D}}$ satisfies $\mid w_{\mathcal{D}} \mid \leq \sqrt{R}$.
- $\mid w_{\mathcal{D}} \mid \leq \sqrt{R}$ also implies that the expected predictor $\bar{w}$ should abide by the rule $\mid \bar{w} \mid \leq \sqrt{R}$

- Given triangle inequality, we also know that $\mid w_{\mathcal{D}} - \bar{w} \mid \leq \mid w_{\mathcal{D}} \mid + \mid \bar{w} \mid$.
- Therefore, we can derive that:

$$\mid w_{\mathcal{D}} - \bar{w} \mid \leq \mid w_{\mathcal{D}} \mid + \mid \bar{w} \mid \leq \sqrt{R} + \sqrt{R}$$

$$\mid w_{\mathcal{D}} - \bar{w} \mid^2 \leq 4R$$

(g) Show that ridge regression bounds the variance by $4Rs^2$

- First, rewrite the original variance formula as follows:

$$\begin{aligned} Variance &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}}x^{(i)} - \mathbb{E}_{\mathcal{D}}[w_{\mathcal{D}}x^{(i)}])^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}} - \bar{w})^2\right] \end{aligned}$$

- Given that $\mid w_{\mathcal{D}} - \bar{w} \mid^2 \leq 4R$, we can derive the inequality:

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} x^{(i)2}\mathbb{E}_{\mathcal{D}}\left[(w_{\mathcal{D}} - \bar{w})^2\right] \leq \frac{1}{N}\sum_{i=1}^{N} x^{(i)2}4R = 4Rs^2 \\ &\implies Variance \leq 4Rs^2 \end{aligned}$$

## 2. Optimal Classifier under Squared Loss. (12 pt)

Let $h_D(\boldsymbol{x})$ be a predictor trained on a dataset $D$, which maps an input feature vector $\boldsymbol{x} \in \mathbb{R}^d$ to a predicted output. The output variable is denoted by $y \in \mathbb{R}$.

Consider the expected squared error loss, which measures the performance of our predictor. This expectation is taken over the joint distribution $P$ of input data $\boldsymbol{x}$ and the true labels $y$, and distribution of dataset $D$ samples from $P^N$, where $D$ has $N$ data points:

$$L = E_{(\boldsymbol{x},y)\sim P, D\sim P^N}\left[(h_D(\boldsymbol{x}) - y)^2\right]$$

Your task is to:

(a) **Find the Optimal Classifier**: Derive the predictor $h_{opt}(\boldsymbol{x})$ that minimizes this expected loss. Note that the optimal predictor should not be dependent on any specific dataset $D$. (6 pt)
*Hint:* One route you can take is applying the law of total expectation and minimizing the inner expectation for a fixed classifier $h_D(\boldsymbol{x})$.

(b) **Find the Optimal Error Rate**: Derive the minium achievable error, or irreducible error, after you derive the optimal classifier. (6 pt)

---

**Solutions:**

---

(a) Find the Optimal Classifier
**Ans:** $h_{opt}(x) = \mathbb{E}\left[y \mid x\right]$

- Given the law of total expectation

$$\begin{aligned} L &= \mathbb{E}_{(x,y)\sim P, D\sim P^N}\left[(h_D(\boldsymbol{x}) - y)^2\right] \\ &= \mathbb{E}_{(x,D)}\left[\mathbb{E}_{(y|x)}(h_D(x) - y)^2 \mid x\right] \end{aligned}$$

- For fixed $x$ and $D$

$$\begin{aligned} &\mathbb{E}_{y|x}\left[(h_D(x) - y)^2 \mid x\right] \\ &= \mathbb{E}_{y|x}\left[h_D(x)^2 - 2h_D(x)y + y^2\right] \\ &= h_D(x)^2 - 2h_D(x)\mathbb{E}\left[y \mid x\right] + \mathbb{E}\left[y^2 \mid x\right] \\ &= h_D(x)^2 - 2h_D(x)\mathbb{E}\left[y \mid x\right] + Var(y \mid x) + \mathbb{E}\left[y \mid x\right]^2 \\ &= \Big(h_D(x) - \mathbb{E}\left[y \mid x\right]\Big)^2 + Var(y \mid x) \end{aligned}$$

- Thus

$$\begin{aligned} L &= \mathbb{E}_{(x,D)}\left[\Big(h_D(x) - \mathbb{E}\left[y \mid x\right]\Big)^2 + Var(y \mid x) \mid x\right] \\ &= \mathbb{E}_{(x,D)}\left[\Big(h_D(x) - \mathbb{E}\left[y \mid x\right]\Big)^2 \mid x\right] + \underbrace{\mathbb{E}_x\left[Var(y \mid x)\right]}_{\text{Independent of } h_D} \\ &= \mathbb{E}_{(x,D)}\left[\Big(h_D(x) - \mathbb{E}\left[y \mid x\right]\Big)^2 \mid x\right] \end{aligned}$$

- Finally, we derive $h_{opt}(x)$ by setting $L = 0$

$$0 = \mathbb{E}_{(x,D)}\left[\left(h_D(x) - \mathbb{E}\left[y \mid x\right]\right)^2 \mid x\right]$$
$$h_D(x) = \mathbb{E}\left[y \mid x\right] = h_{opt}(x)$$

(b) Find the Optimal Error Rate
**Ans:** $\mathbb{E}_x\left[Var(y \mid x)\right]$
Plug the $h_{opt}(x) = \mathbb{E}\left[y \mid x\right]$ back to L, and get the answer:

$$\begin{aligned} L^* &= \mathbb{E}_{(x,D)}\left[\left(\mathbb{E}\left[y \mid x\right] - \mathbb{E}\left[y \mid x\right]\right)^2 \mid x\right] + \mathbb{E}_x\left[Var(y \mid x)\right] \\ &= \mathbb{E}_x\left[Var(y \mid x)\right] \end{aligned}$$

# 3. Model Selection. (19 pt)

In this problem, you will implement a model selection pipeline using k-fold cross-validation to find the best hyper-parameters for polynomial regression with regularization. You can see more detailed instructions in the code file `hw4_q3.py`.

**Submission Instruction** If you want to implement any helper function of your own, please make sure you either put it directly in `hw4_q3.py` or put them into `hw4_utils.py` and submit `hw4_utils.py` with `hw4_q3.py` to Gradescope!

(a) **K-Fold Cross-Validation (8 pt)**
Implement `cross_validate_model(X, y, model, k_folds)` that

- Splits the data into $k$ folds using `KFold` with `shuffle=True` and `random_state=42`
- For each fold, trains the model on $k-1$ folds and evaluates on the remaining fold
- Returns the mean and standard deviation of validation mean squared error across all folds

**Remark 1:** For `model`, you can train the `model` by calling `model.fit(X,y)` on data `(X,y)`. In addition, you can call `model.predict(X)` to obtain the prediction from `model`.
**Remark 2:** For each iteration during k-fold cross validation, please make sure you make a copy of `model` by `model_copy = deepcopy(model)` and then train `model_copy` instead of `model`. Otherwise, you will be training a model from previous iteration.

(b) **Model Selection (11 pt)**
Implement `select_best_model(X_train, y_train)` that sweeps through different polynomial degrees and regularization strengths (for Ridge and Lasso regression) to perform k-fold cross validation with $k = 5$. The function should return the model with lowest cross-validation error.
**Remark 1:** You can use `LinearRegression()` to initialize the Linear Regression model.
**Remark 2:** You can use `Ridge(alpha=alpha, random_state=42)` to initialize the Ridge Regression model.
**Remark 3:** You can use `Lasso(alpha=alpha, random_state=42, max_iter=2000)` to initialize the Lasso Regression model.

---

**Solutions:**

---