


STA 521: Homework 2

Release date: **Tuesday, Sept 20**

Due by: **3 PM, Tuesday, Oct 04**


The honor code

- (a) Please state the names of people who you worked with for this homework. You can also provide your comments about the homework here.



- (b) Please type/write the following sentences yourself and sign at the end. We want to make it *extra* clear that nobody cheats even unintentionally.

*I hereby state that all of my solutions were entirely in my words and were written by me.
I have not looked at another student's solutions and I have fairly credited all external
sources in this write up.*



Submission instructions

- It is a good idea to revisit your notes, slides and reading; and synthesize their main points BEFORE doing the homework.
- A .Rnw file corresponding to the homework is also uploaded for you. You may use that to write-up your solutions. Alternately, you can typeset your solutions in latex or submit neatly handwritten/scanned solutions.
- **For the parts that ask you to implement/run some R code, your answer should look something like this (code followed by result):**

```
> myfun<- function(){  
+ show('this is a dummy function')  
+ }  
> myfun()  
  
[1] "this is a dummy function"
```

Note that this is automatically generated if you use the R knitr environment.

- You need to submit the following:
 1. A pdf of your write-up to “HW2 write-up” that includes some of the code snippets and plots as asked in different parts.
 2. A Rmd or Rnw file, that has all your entire code, to “HW2 code”.
- Ensure a proper submission to gradescope, otherwise it will not be graded.

Homework Overview

This homework revisits ordinary least squares and other regression methods. Some problems are from the **ISL** book and the **ESL** book (available on the course website).

1 True or False (8 pts)

Examine whether the following statements are true or false and *provide one line justification*. Linear model in the following statements refers to the linear model studied in the lectures.

- (a) Under the linear model, the OLS estimator of the regression coefficients is unbiased.
- (b) For the linear model, bias of the ridge regression increases and the variance decreases as we increase the regularization parameter λ .
- (c) Every eigenvalue of an idempotent matrix is always either zero or one. (Recall that A square matrix $M \in \mathbb{R}^{m \times m}$ is called *idempotent* if $M^2 = M$.)

- (d) Let X be an $n \times p$ matrix of full rank. Let $H = X(X^\top X)^{-1}X^\top$. The matrix H is symmetric, idempotent and PSD.
- (e) Let $Q = \mathbb{I}_n - H$ where H is defined in the previous part. When $p \leq n$, we have $\text{trace}(Q) = n$.
- (f) An outlier in linear regression always has high leverage score.
- (g) Given p predictors x_1, \dots, x_p . The more predictors we include in the OLS, the smaller test error will be.
- (h) The two letters “gg” in the name of the R package ggplot2 stands for “Grammar of Graphics”.

2 Understanding orthogonal projection (13 points)

In this problem, we will understand orthogonal projection in quite some detail. The length of the problem should not bother you as several parts are relatively easy.

- (a) For any vector $\mathbf{x} = [x_1, x_2, x_3]^\top \in \mathbb{R}^3$, what is its projection on the vector $[1, 0, 0]^\top$?
- (b) What is the orthogonal projection of $\mathbf{x} = [x_1, x_2, x_3]^\top \in \mathbb{R}^3$ on the subspace spanned by the vectors $[1, 0, 0]^\top$ and $[0, 1, 0]^\top$?
- (c) Write the projection matrices in the previous two cases.
- (d) Create the previous projection matrices in R and compute the projections of two random vectors whose entries are drawn independently from the uniform distribution on $[0, 1]$. Do not forget to set the seed.
- (e) Let’s make the problem more general. Write the expression for the orthogonal projection of a vector $\mathbf{x} \in \mathbb{R}^d$ along any given vector $\mathbf{a} \in \mathbb{R}^d$. What is the projection matrix in this case?
- (f) Create a function in R that takes in input two vectors \mathbf{x}, \mathbf{a} and computes the projection from previous part. Call the function with $\mathbf{x} = [3, 2, -1]^\top$ and $\mathbf{a} = [1, 0, 1]^\top$.
- (g) Given two orthogonal vectors \mathbf{a}_1 and \mathbf{a}_2 in \mathbb{R}^d , what is the orthogonal projection of a generic vector \mathbf{x} on to the subspace spanned by these two vectors?
- (h) Now let’s make the problem a bit more challenging. Suppose that the two vectors \mathbf{a}_1 and \mathbf{a}_2 are not orthogonal. How will you compute the orthogonal projection of \mathbf{x} in this case? It may be useful to revise Gram Schmidt Orthogonalization.
- (i) Implement your method in the previous part in R for $\mathbf{x} = [3, 2, -1]^\top$ and $\mathbf{a}_1 = [1, 0, 1]^\top$ and $\mathbf{a}_2 = [1, -1, 0]^\top$.
- (j) Can you generalize the answer from previous part to the case to compute the orthogonal projection along a k -dimensional subspace spanned by the vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ which need not be orthogonal?

(k) Define the matrix

$$\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_k] \in \mathbb{R}^{d \times k}$$

such that the columns are independent. Then the orthogonal projection of any vector $\mathbf{x} \in \mathbb{R}^d$ onto the k -dimensional subspace spanned by the vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ is given by

$$\mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{x}. \quad (1)$$

- (l) Compute the projection for $\mathbf{x} = [3, 2, -1]^\top$ on the space spanned by the vectors $\mathbf{a}_1 = [1, 0, 1]^\top$ and $\mathbf{a}_2 = [1, -1, 0]^\top$ using the expression in previous part using R. Compare it with your answer from part (i).
- (m) How does the expression in the equation (1) simplify if the vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ are orthogonal?

3 A closer look at EM (11 pts)

In this question, we consider a simple mixture model and work our way through a derivation of the EM updates.

We work with the following simple two mixture model:

$$\begin{aligned} Z &\sim \text{Bernoulli}(1 - w) + 1 \\ X|Z = 1 &\sim \mathcal{N}(\mu_1, 1), \quad \text{and} \\ X|Z = 2 &\sim \mathcal{N}(\mu_2, 1), \end{aligned} \quad (2)$$

where Z denotes the label of the Gaussian from which X is drawn. Given a set of observations only for X (i.e., the labels are unobserved), our goal is to infer the maximum-likelihood parameters for μ_1, μ_2 and w . Note that to simplify your calculations, we have fixed the variance parameter and assumed it to be known.

- (a) (3 points) Let $\theta = (\mu_1, \mu_2, w)$ denote the parameters of the model. **Write down the expressions of the joint likelihood $p(X = x, Z = 1; \theta)$ and $p(X = x, Z = 2; \theta)$. What is the marginal likelihood $p(X = x; \theta)$ and the log-likelihood $\ell(X = x; \theta)$? Given n i.i.d. samples $\{x_1, \dots, x_n\}$, write the expression for the log-likelihood $\ell(X_1 = x_1, \dots, X_n = x_n; \theta)$.**
- (b) (2 points) The EM algorithm is an iterative algorithm to maximize likelihood which at each iteration does one E-step and one M-step. At iteration t , given the current parameter estimate $\theta^t = (\mu_1^t, \mu_2^t, w^t)$, **Derive the expressions for $p(Z = 1|X = x_i; \theta^t)$ and $p(Z = 2|X = x_i; \theta^t)$. Then describe the E-step computations using these quantities.** (see Lecture 05)
- (c) (3 points) Using the expression of the conditional expectation $\sum_{i=1}^n \mathbb{E}_{Z|x_i, \theta^t}[\log(p(x_i, Z; \theta))]$ from the previous part, **derive the expressions for its gradients with respect to μ_1, μ_2, w . By setting these gradients to zero, show that the M-step updates are given by**

$$\mu_1^{t+1} = \frac{\sum_{i=1}^n q_i^{t+1} x_i}{\sum_{i=1}^n q_i^{t+1}}, \quad \mu_2^{t+1} = \frac{\sum_{i=1}^n (1 - q_i^{t+1}) x_i}{\sum_{i=1}^n (1 - q_i^{t+1})}, \quad \text{and} \quad w^{t+1} = \frac{\sum_{i=1}^n q_i^{t+1}}{n},$$

where $q_i^{t+1} = p(Z = 1 \mid X = x_i; \theta^t)$ derived in (b). This complete the derivation of EM updates.

- (d) (3 points) We now see EM in action and compare it with K-means. (Only plots are required in the write-up submission.) Generate 1000 samples from the following mixture of two Gaussians in two dimensions:

$$\begin{aligned} Z &= \text{Bernoulli}(0.5) + 1 \\ X|Z = 1 &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\ X|Z = 2 &\sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}\right) \end{aligned} \quad (3)$$

where I_2 denotes the identity matrix in two dimensions. **Generate both (Z, X) for the data and scatter plot the X values with color based on Z . Run K-means on X with $K = 2$ and report the cluster centers and scatter plot the data with estimated labels. Run Gaussian mixture model with EM on X to fit two clusters too (use Mclust package with $G = 2$) and report the mean parameters and scatter plot the data with estimated labels. Note that we need 3 plots for this part. Justify qualitatively why the cluster centers obtained by K-means and EM and the estimated label are different.**

4 OLS statistical properties (4 pts)

1. (2 pts) Consider a linear model where we observe the samples (x_i, y_i) for $i = 1, \dots, n$, that are generated as follows

$$y_i = x_i^\top \beta^* + \epsilon, \quad (4)$$

where the error $\epsilon \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. **Show that the OLS estimate $\hat{\beta}$ on data $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^n$ satisfies**

$$\hat{\beta} \sim \mathcal{N}\left(\beta^*, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right).$$

Also conclude that $\mathbb{E}[\mathbf{X}\hat{\beta}] = \mathbf{X}\beta^*$.

2. (2 pts) In the Gaussian linear model described above, **show that $\mathbb{E}[RSS] = \sigma^2(n - d)$.**
3. (optional, not graded) **ESL book Ex. 3.3 (a)** (part (b) is not needed). *Hint:* The notion unbiased linear estimator is explained in Section 3.2.2 around equation (3.19).
4. (optional, not graded) **ESL book Ex. 2.9.** We expect a solution that **does not** use the explicit data generation process. *Partial credit is given if you use the explicit data generation process.*

5 Theory of ridge regression (14 pts)

Given the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the responses $\mathbf{y} \in \mathbb{R}^n$, ridge regression solves the following penalized least squares problem:

$$\min_{\theta} \mathbf{X}\theta - \mathbf{y}^2 + \lambda\theta^2 \quad (5)$$

Let $\mathbf{X}^\top \mathbf{X}$ have the following eigen-decomposition: $\mathbf{X}^\top \mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{U}^\top$ where \mathbf{D} is a diagonal matrix with non-negative entries.

Let $\hat{\theta}$ denote the solution for the problem (5) above.

1. (1 pt) **Show that for any $\lambda > 0$, the solution $\theta^{\text{RR}}(\lambda)$ is unique and derive its expression.**
2. (2 pts) For *all the following parts*, we assume the linear regression model: That is the data is generated as follows:

$$\mathbf{y} = \mathbf{X}\theta^* + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Using the previous part, **show that the distribution of the ridge-estimate is given as follows:**

$$\hat{\theta} \sim \mathcal{N}(\mathbf{W}_\lambda \theta^*, \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1})$$

where $\mathbf{W}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{X}$.

3. (4 pts) Let $\mathbf{X}^\top \mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{U}^\top$ with $\mathbf{D}_{ii} = d_i$. **Show the following:**

(a) **The squared bias is given by**

$$\text{squared bias} = \mathbb{E}[\hat{\theta}] - \theta^{*2} = \sum_{i=1}^p \frac{\lambda^2}{(d_i + \lambda)^2} (v_i)^2,$$

where $v = \mathbf{U}^\top \theta^*$.

(b) **The (scalar) variance is given by**

$$\text{scalar-variance} = \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2 = \sigma^2 \sum_{i=1}^p \frac{d_i}{(d_i + \lambda)^2}.$$

4. (2 pts) **What is the value of the squared-bias and variance at $\lambda = 0$ and as $\lambda \rightarrow \infty$. Do you see a trade-off in the squared-bias and variance change as λ increases?**
5. (2 pts) Define the moment matrix \mathbf{M} :

$$\mathbf{M}(\lambda) = \mathbb{E}[(-\theta^*)(-\theta^*)^\top].$$

Recall the mean-squared error

$$\text{MSE}() := \mathbb{E}[-\theta^{*2}].$$

Show that $\mathbb{E}[-\theta^{*2}] = \text{trace}(\mathbf{M}(\lambda))$. Moreover, show that

$$\text{MSE}() = \text{squared-bias} + \text{scalar-variance}$$

6. (3 pts) Recall that when $\mathbf{X}^\top \mathbf{X}$ is invertible, the OLS-estimator is unbiased. For this case, show that its mean squared error satisfies $\text{MSE}(\theta^{\text{OLS}}) = \text{trace}(\mathbf{M}(0))$. Furthermore, show that there exists a range of $\lambda > 0$ for which

$$\text{MSE}() < \text{MSE}(\theta^{\text{OLS}}).$$

Conclude that there always exists a range of $\lambda > 0$, for which the MSE is smaller for ridge regression when compared to OLS in the Gaussian linear model.

6 Apply linear regression to a dataset (Optional, not graded)

This problem applies linear regression to the **Carseats** dataset in the ISL book. Use the following code to load the **Carseats** dataset.

```
> # install.packages("ISLR")
> library("ISLR")
> head(Carseats)
```

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education
1	9.50	138	73	11	276	120	Bad	42	17
2	11.22	111	48	16	260	83	Good	65	10
3	10.06	113	35	10	269	80	Medium	59	12
4	7.40	117	100	4	466	97	Medium	55	14
5	4.15	141	64	3	340	128	Bad	38	13
6	10.81	124	113	13	501	72	Bad	78	16

	Urban	US
1	Yes	Yes
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes
5	Yes	No
6	No	Yes

Answer the questions in ISL 3.7.10 (a)-(h)