CS 475 Machine Learning: Homework 2 Analytical (35 points)

Assigned: Friday, September 24, 2021

Due: Friday, October 8, 2021, 11:59 pm US/Eastern

Partner 1: Chang Yan (cyan13), Partner 2: Jingguo Liang (jliang35)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box. You can change the size of the box if you need more space.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not add text outside of the answer boxes. You are allowed to make boxes larger if needed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

1 Ridge Regression

1. Assume $\mathbf{X}$ is a dataset of n rows of k feature values each, and $\mathbf{y}$ is the corresponding vector of outcome values. Assume the data is centered, meaning that $\mathrm{E}[Y] = 0$, and for each $X_i \in \mathbf{X}$, $\mathrm{E}[X_i] = 0$. Consider the following modified squared loss for a linear regression model:

$$J(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \|\beta\|_2^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_i \beta_i^2. \tag{1}$$

Note that since $\mathrm{E}[Y] = 0$, the linear regression does not need an intercept parameter.

Assuming that $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ ($\mathbf{I}$ is the identity matrix) is invertible, find the values of β that minimize this loss. Please show your work.

$$\begin{aligned}
\frac{dJ(\beta)}{d\beta} &= (\frac{d}{d\beta})(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_i \beta_i^2 \\
&= (\frac{d}{d\beta})(\mathbf{y}^T - \beta^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_i \beta_i^2 \\
&= (\frac{d}{d\beta})\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda\beta^T\beta \\
&= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta + 2\lambda\beta
\end{aligned}$$

Let $\frac{dJ(\beta)}{d\beta} = 0$

$$\begin{aligned}
-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta + 2\lambda\beta &= 0 \\
2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta &= 2\mathbf{X}^T\mathbf{y} \\
\beta &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

2. Show that the ridge regression minimizer in the previous question is the mode of the posterior distribution, under a Gaussian prior on β given by $\mathcal{N}(0, \tau \cdot \mathbf{I})$, and Gaussian likelihood $Y = \mathcal{N}(\mathbf{X} \cdot \beta, \sigma^2\mathbf{I})$. The mode β^* of the posterior are the settings of parameters that maximize the posterior distribution (e.g. the maximum a posteriori (MAP) parameter estimates).

$$
\begin{aligned}
p_{posterior}(\beta) &= \frac{\mathcal{L}_{[D]}(\beta) \cdot p_{prior}(\beta)}{\int_\beta \mathcal{L}_{[D]}(\beta) \cdot p_{prior}(\beta)} \\
&= \frac{\frac{1}{\sigma\mathbf{I}\sqrt{2\pi}^k} e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2}} \cdot \frac{1}{\tau\mathbf{I}\sqrt{2\pi}^k} e^{-\frac{||\beta||_2^2}{2(\tau)^2}}}{\int_\beta \frac{1}{\sigma\mathbf{I}\sqrt{2\pi}^k} e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2}} \cdot \frac{1}{\tau\mathbf{I}\sqrt{2\pi}^k} e^{-\frac{||\beta||_2^2}{2(\tau)^2}}} \\
&= \frac{e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2}} \cdot e^{-\frac{||\beta||_2^2}{2(\tau)^2}}}{\int_\beta e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2}} \cdot e^{-\frac{||\beta||_2^2}{2(\tau)^2}}} \\
&= \frac{e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} - \frac{||\beta||_2^2}{2(\tau)^2}}}{\int_\beta e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} - \frac{||\beta||_2^2}{2(\tau)^2}}}
\end{aligned}
$$

We know the denominator of the above function is just a normalization constant, and exp(.) is a monotonic, concave function. So maximizing the above function is equivalent to maximizing:

$$-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} - \frac{||\beta||_2^2}{2(\tau)^2}$$

And maximizing the above equation is equivalent to minimizing:

$$\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} + \frac{||\beta||_2^2}{2(\tau)^2}$$

take the derivative with respect to β:

$$
\begin{aligned}
&(\frac{d}{d\beta})\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} + \frac{||\beta||_2^2}{2(\tau)^2} \\
=&(\frac{d}{d\beta})\frac{1}{2\sigma^2}(\mathbf{y}^T - \beta^T\mathbf{X}^T)(\mathbf{y}-\mathbf{X}\beta) + \frac{1}{2\tau^2}\beta^T\beta \\
=&(\frac{d}{d\beta})\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta) + \frac{1}{2\tau^2}\beta^T\beta \\
=&-\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} + (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})\beta
\end{aligned}
$$

To find the mode β^* we set it to zero:

$$-\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} + (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})\beta^* = 0$$

(Continued on next page)

so we have:

$$(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})\beta^* = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}$$
$$(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})\beta^* = \mathbf{X}^T\mathbf{y}$$
$$\beta^* = (\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

We can see the mode of the posterior distribution is the same as ridge regression minimizer

3. Find the relationship between the regularization parameter λ in the ridge formula, and the variance parameters τ and σ^2.

   Show your work!

   Hints:

   - The posterior takes the form of $\frac{\mathcal{L}_{[D]}(\beta)\cdot p(\beta)}{\int \mathcal{L}_{[D]}(\beta)\cdot p(\beta) d\beta}$. It often suffices to only think about the numerator, and let the denominator be whatever normalizing function that makes the whole expression integrate to 1.
   - In class we used the fact that log(.) is a concave function to conclude maximizing the likelihood is equivalent to maximizing the log likelihood. For this problem it might be useful to use the fact that exp(.) is a convex function.
   - The multivariate normal distribution on k variables with mean vector μ and covariance matrix Σ has the density $(2\pi)^{-k/2}\det(\mathbf{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$.

We know from question 1 ridge regression that:

$$\beta = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

We know from question 2 MAP that:

$$p_{posterior}(\beta) = \frac{e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} - \frac{||\beta||_2^2}{2(\tau)^2}}}{\int_\beta e^{-\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} - \frac{||\beta||_2^2}{2(\tau)^2}}}$$

$$MAP(\beta^*) : (\frac{d}{d\beta})\frac{(\mathbf{y}-\mathbf{X}\cdot\beta)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\cdot\beta)}{2} + \frac{||\beta||_2^2}{2(\tau)^2} = 0$$

$$\beta^* = (\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

As they are equal, we have:

$$\lambda = \frac{\sigma^2}{\tau^2}$$

2 Splitting Data And Combining Predictors

Assume a linear regression model $Y = X^T \cdot \beta + \epsilon$, where ϵ is an arbitrary distribution.

Given a dataset $[D]$ of size n draw from the true observed data distribution $p_0(X, Y)$, imagine training two predictors. The first predictor, $\hat{f}^{\text{whole}}$ simply minimizes the squared loss on $[D]$. The second predictor $\hat{f}^{\text{split}}$ splits $[D]$ into two halves $[D]_1, [D]_2$ each of size $n/2$, trains two separate models: $\hat{f}^{(1)}$ by minimizing squared loss on $[D]_1$, and $\hat{f}^{(2)}$ by minimizing squared loss on $[D]_2$, and then averages the predictions of these two models:

$$\hat{f}^{\text{split}}(x) = \frac{1}{2}\left(\hat{f}^{(1)}(x) + \hat{f}^{(1)}(x)\right).$$

Consider a fixed input/output pair x_0, y_0, and the MSE $\text{E}[(y_0 - \hat{f}^{\text{split}}(x_0))^2]$ and $\text{E}[(y_0 - \hat{f}^{\text{whole}}(x_0))^2]$ of both predictors, with the expectation taken over $p([D])$.

1. Write out the bias/variance decomposition of both the MSE of both predictors, expressing this decomposition in terms of E[.] and $Var(.)$ of random quantities, e.g. parameters of the models fit using $[D]$ drawn from $p([D])$. You can call the parameters of $\hat{f}^{\text{whole}}$ by β^{whole}, parameters of $\hat{f}^{\text{split}}$ by $\beta^{(1)}$ and $\beta^{(2)}$.

$$\begin{aligned}
E[(y_0 - \hat{f}^{\text{whole}}(x_0))^2] &= E[y_0^2] - E[2y_0\hat{f}^{\text{whole}}(x_0)] + E[\hat{f}^{\text{whole}}(x_0)^2] \\
&= y_0^2 - 2y_0E[\hat{f}^{\text{whole}}(x_0)] + E[\hat{f}^{\text{whole}}(x_0)^2] \\
&= y_0^2 - 2y_0E[\hat{f}^{\text{whole}}(x_0)] + Var[\hat{f}^{\text{whole}}(x_0)] + E[\hat{f}^{\text{whole}}(x_0)]^2 \\
&= (y_0 - E[\hat{f}^{\text{whole}}(x_0)])^2 + Var[\hat{f}^{\text{whole}}(x_0)] \\
&= E[y_0 - \hat{f}^{\text{whole}}(x_0)]^2 + Var[\hat{f}^{\text{whole}}(x_0)] \\
&= E[y_0 - (x_0^T \cdot \beta^{\text{whole}} + \epsilon)]^2 + Var[x_0^T \cdot \beta^{\text{whole}} + \epsilon]
\end{aligned}$$

Similarly,

$$\begin{aligned}
E[(y_0 - \hat{f}^{\text{split}}(x_0))^2] &= E[y_0 - \hat{f}^{\text{split}}(x_0)]^2 + Var[\hat{f}^{\text{split}}(x_0)] \\
&= E[y_0 - \frac{1}{2}(\hat{f}^{(1)}(x_0) + \hat{f}^{(2)}(x_0))]^2 + Var[\frac{1}{2}(\hat{f}^{(1)}(x_0) + \hat{f}^{(2)}(x_0))] \\
&= E[y_0 - (x_0^T \cdot \frac{1}{2}(\beta^{(1)} + \beta^{(2)}) + \epsilon)]^2 + Var[x_0^T \cdot \frac{1}{2}(\beta^{(1)} + \beta^{(2)}) + \epsilon]
\end{aligned}$$

2. Compare the variance of $\hat{f}^{\text{whole}}$ with the variance of $\hat{f}^{\text{split}}$.

$$
\begin{aligned}
Var[\hat{f}^{\text{whole}}] &= Var[x_0^T \cdot \beta^{\text{whole}} + \epsilon] \\
&= (x_0^T)^2 \cdot Var[\beta^{\text{whole}}] + Var[\epsilon]
\end{aligned}
$$

$$
\begin{aligned}
Var[\hat{f}^{\text{split}}] &= Var[x_0^T \cdot \frac{1}{2}(\beta^{(1)} + \beta^{(2)}) + \epsilon] \\
&= \frac{1}{4}(x_0^T)^2(Var[\beta^{(1)}] + Var[\beta^{(2)}]) + Var[\epsilon]
\end{aligned}
$$

I am assuming that $Var[\beta^{\text{whole}}]$, $Var[\beta^{(1)}]$, $Var[\beta^{(2)}]$ are the same or similar. Then the variance of $\hat{f}^{\text{whole}}$ will be greater than the variance of $\hat{f}^{\text{split}}$, since $Var[\beta^{\text{whole}}] > \frac{1}{4}(Var[\beta^{(1)}] + Var[\beta^{(2)}])$

3. What are the advantages and disadvantages of using $\hat{f}^{\text{whole}}$ versus $\hat{f}^{\text{split}}$, if both are unbiased estimators.

The advantage of using $\hat{f}^{\text{split}}$ compared with $\hat{f}^{\text{whole}}$ is that it has a smaller variance and thus produces more stable result. The disadvantage of using $\hat{f}^{\text{split}}$ compared with $\hat{f}^{\text{whole}}$ is that calculating $\hat{f}^{\text{whole}}$ only involves the prediction of one model and is thus more simple and less time-comsuming, while $\hat{f}^{\text{split}}$ needs to train two models, which also means we need doubled space to store parameters. Also, for each splitted model, only looking into part of the data set may not be as representative as looking into the whole, and the model parameters may converges slower as there are less data points each model can look at (also meaning longer training time).

3 Naive Bayes and Logistic Regression

A Naive Bayes classifier uses the conditional probability $p(Y \mid \mathbf{X})$ to predict the value of Y given $\mathbf{X}$ (for Y with a finite set of values). This conditional probability is obtained from the following model: $p(Y, \mathbf{X}) = p(Y) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y)$. Thus,

$$p(Y \mid \mathbf{X}) = \frac{p(Y) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y)}{\sum_Y p(Y) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y)}.$$

Assume Y has only two values (0 and 1), and for each $X_i \in \mathbf{X}$,

$$X_i \mid Y = 0 \sim \mathcal{N}(\mu_{i0}, \sigma_i^2),$$
$$X_i \mid Y = 1 \sim \mathcal{N}(\mu_{i1}, \sigma_i^2).$$

1. Show that $p(Y = 1 \mid \mathbf{X})$ has the same parametric form as a logistic regression model. Hint:
   - It might be convenient for you to first show that: $p(Y = 1 \mid \mathbf{X}) = \frac{1}{1+\exp\left\{\log\left(\frac{p(Y=0)p(\mathbf{X}|Y=0)}{p(Y=1)p(\mathbf{X}|Y=1)}\right)\right\}}$.

First using the two values of Y (0,1) to show the hint result:

$$\begin{aligned}
p(Y=1 \mid \mathbf{X}) &= \frac{p(Y=1) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y=1)}{\sum_Y p(Y) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y)} \\
&= \frac{p(Y=1) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y=1)}{p(Y=1) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y=1) + p(Y=0) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y=0)} \\
&= \frac{1}{1 + \frac{p(Y=0) \prod_{X_i \in \mathbf{X}} p(X_i|Y=0)}{p(Y=1) \prod_{X_i \in \mathbf{X}} p(X_i|Y=1)}} \\
&= \frac{1}{1 + e^{log(\frac{p(Y=0) \prod_{X_i \in \mathbf{X}} p(X_i|Y=0)}{p(Y=1) \prod_{X_i \in \mathbf{X}} p(X_i|Y=1)})}}
\end{aligned}$$

Then we try to simplify the expression on the exponential:

$$\begin{aligned}
& log(\frac{p(Y=0) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y=0)}{p(Y=1) \prod_{X_i \in \mathbf{X}} p(X_i \mid Y=1)}) \\
=& log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} log\frac{p(X_i \mid Y=0)}{p(X_i \mid Y=1)} \\
=& log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} log\frac{\frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(X_i-\mu_{i0})^2}{2\sigma_i^2}}}{\frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(X_i-\mu_{i1})^2}{2\sigma_i^2}}} \\
=& log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} -\frac{(X_i-\mu_{i0})^2 - (X_i-\mu_{i1})^2}{2\sigma_i^2} \\
=& log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} -\frac{X_i^2 - 2X_i\mu_{i0} + \mu_{i0}^2 - X_i^2 + 2X_i\mu_{i1} - \mu_{i1}^2}{2\sigma_i^2}
\end{aligned}$$

(Continued on next page)

$$
\begin{aligned}
&= log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} \frac{2X_i(\mu_{i0} - \mu_{i1}) - \mu_{i0}^2 + \mu_{i1}^2}{2\sigma_i^2} \\
&= log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} \frac{-\mu_{i0}^2 + \mu_{i1}^2}{2\sigma_i^2} + \sum_{X_i \in \mathbf{X}} \frac{X_i(\mu_{i0} - \mu_{i1})}{\sigma_i^2}
\end{aligned}
$$

we know that all of $log\frac{p(Y=0)}{p(Y=1)}$, $\sum_{X_i \in \mathbf{X}} \frac{-\mu_{i0}^2 + \mu_{i1}^2}{2\sigma_i^2}$ and $\frac{(\mu_{i0} - \mu_{i1})}{\sigma_i^2}$ are just constants, so we can let:

$$\beta_0 = log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} \frac{-\mu_{i0}^2 + \mu_{i1}^2}{2\sigma_i^2}$$

$$\beta_i = \frac{(\mu_{i0} - \mu_{i1})}{\sigma_i^2}$$

Thus, the expression (whcih is actually the decision boundary) can be written as:

$$\beta_0 + \sum_{X_i \in \mathbf{X}} \beta_i X_i$$

put the expression back into $p(Y = 1 \mid \mathbf{X})$ we have:

$$
\begin{aligned}
p(Y = 1 \mid \mathbf{X}) &= \frac{1}{1 + e^{log(\frac{p(Y=0)\prod_{X_i \in \mathbf{X}} p(X_i|Y=0)}{p(Y=1)\prod_{X_i \in \mathbf{X}} p(X_i|Y=1)})}} \\
&= \frac{1}{1 + e^{log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} \frac{-\mu_{i0}^2 + \mu_{i1}^2}{2\sigma_i^2} + \sum_{X_i \in \mathbf{X}} \frac{X_i(\mu_{i0} - \mu_{i1})}{\sigma_i^2}}} \\
&= \frac{1}{1 + e^{\beta_0 + \sum_{X_i \in \mathbf{X}} \beta_i X_i}}
\end{aligned}
$$

where $\beta_0 = log\frac{p(Y=0)}{p(Y=1)} + \sum_{X_i \in \mathbf{X}} \frac{-\mu_{i0}^2 + \mu_{i1}^2}{2\sigma_i^2}$ and $\beta_i = \frac{(\mu_{i0} - \mu_{i1})}{\sigma_i^2}$ are both constants.
We can see this is exactly in the same parametric form as a logistice regression model.