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: Dimitri Lezcano (dlezcan1), Partner 2: Harrison Khoo (hkhoo2)

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.

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

To minimize $J(\beta)$, we set $\frac{\partial J(\beta)}{\partial \beta} = 0$

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

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).

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 $(2\pi)^{-k/2}\det(\mathbf{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$.

$$
\begin{aligned}
posterior &= likelihood \times prior \\
&= \frac{1}{2\pi^{k/2}\sqrt{\sigma^2}} exp\{\frac{-1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\sigma^2(\mathbf{y}-\mathbf{X}\beta)\} \times \frac{1}{2\pi^{k/2}\sqrt{\tau}} exp\{\frac{-1}{2}\beta^T\tau\beta\} \\
&= \frac{1}{4\pi^k\sigma\sqrt{\tau}} exp\{\frac{-\sigma^2}{2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) - \frac{\tau}{2}\beta^T\beta\}
\end{aligned}
$$

Since exp(.) is convex, we know that $argmax\ posterior = argmax\ \ln\{posterior\}$

$$
\begin{aligned}
argmax\ \ln\{posterior\} &= argmax\ \frac{-\sigma^2}{2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) - \frac{\tau}{2}\beta^T\beta \\
&= argmin\ (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) + \frac{\tau}{\sigma^2}\beta^T\beta
\end{aligned}
$$

This is the same form as equation (1) and minimizes in the same fashion. Here, $\lambda = \frac{\tau}{\sigma^2}$.

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:

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

Consider a fixed input/output pair x_0, y_0, and the MSE $\mathrm{E}[(y_0 - \hat{f}^{\text{split}}(x_0))^2]$ and $\mathrm{E}[(y_0 - \hat{f}^{\text{whole}}(x_0))^2]$ 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 $\beta^{(2)}$.

$\hat{f}^{whole}(x_0) = x_0^T \beta^{whole} + \epsilon$

$$\begin{aligned}
\mathrm{E}[(y_0 - \hat{f}^{whole}(x_0))^2] &= \mathrm{E}[(y_0 - (x_0^T\beta^{whole} + \epsilon))^2] \\
&= \mathrm{E}[y_0^2] + \mathrm{E}[(x_0^T\beta^{whole})^2] + E[\epsilon^2] - 2\,\mathrm{E}[y_0 x_0^T \beta^{whole}] \\
&\quad - 2\,\mathrm{E}[y_0\epsilon] - 2\,\mathrm{E}[x_0^T\beta^{whole}\epsilon] \\
&= \mathrm{Var}[x_0^T\beta^{whole} + \epsilon] + \mathrm{E}[y_0^2] + \mathrm{E}[(x_0^T\beta^{whole})]^2 + E[\epsilon]^2 \\
&\quad - 2\,\mathrm{E}[y_0 x_0^T\beta^{whole}] - 2\,\mathrm{E}[y_0\epsilon] - 2\,\mathrm{E}[x_0^T\beta^{whole}\epsilon] \\
&= \mathrm{Var}[x_0^T\beta^{whole}] + \mathrm{Var}[\epsilon] + \Big(\mathrm{E}[y_0 - (x_0^T\beta^{whole} + \epsilon)]\Big)^2
\end{aligned}$$

Now for $\hat{f}^{split}$

$$\begin{aligned}
\mathrm{E}[(y_0 - \hat{f}^{split}(x_0))^2] &= \mathrm{E}[(y_0 - \frac{1}{2}(x_0^T\beta^{(1)} + \epsilon^{(1)} + x_0^T\beta^{(2)} + \epsilon^{(2)}))^2] \\
&= \frac{1}{4}\left(\mathrm{Var}[x_0^T\beta^{(1)}] + \mathrm{Var}[x_0^T\beta^{(2)}]\right) + \frac{1}{2}\,\mathrm{Var}[\epsilon] \\
&\quad + \left(\mathrm{E}[y_0 - \frac{1}{2}(x_0^T\beta^{(1)} + \epsilon^{(1)} + x_0^T\beta^{(2)} + \epsilon^{(2)})]\right)^2 \\
&= \frac{1}{4}\left(\mathrm{Var}[x_0^T\beta^{(1)}] + \mathrm{Var}[x_0^T\beta^{(2)}]\right) + \frac{1}{2}\,\mathrm{Var}[\epsilon] \\
&\quad + \frac{1}{4}\left(\mathrm{E}[y_0 - (x_0^T\beta^{(1)} + \epsilon^{(1)})]^2 + \mathrm{E}[y_0 - (x_0^T\beta^{(2)} + \epsilon^{(2)})]^2\right)
\end{aligned}$$

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

$$\text{Var}[\hat{f}^{whole}(x_0)] = \text{Var}[x_0^T \beta^{whole}] + \text{Var}[\epsilon]$$
$$\text{Var}[\hat{f}^{split}(x_0)] = \frac{1}{4}\left(\text{Var}[x_0^T \beta^{(1)}] + \text{Var}[x_0^T \beta^{(2)}]\right) + \frac{1}{2}\text{Var}[\epsilon]$$

If $\beta^{whole} = \beta^{(1)} = \beta^{(2)} = \beta$, then we can relate the two variances.

$$\text{Var}[x_0^T \beta] = \text{Var}[\hat{f}^{whole}(x_0)] - \text{Var}[\epsilon]$$
$$\text{Var}[\hat{f}^{split}(x_0)] = \frac{1}{2}\left(\text{Var}[\hat{f}^{whole}(x_0)] - \text{Var}[\epsilon]\right) + \frac{1}{2}\text{Var}[\epsilon]$$
$$= \frac{1}{2}\text{Var}[\hat{f}^{whole}(x_0)]$$

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

For small n (small $[D]$), dataset imbalance can greatly increase the variance of the split model. For large n, the variance of each of the individual models $\hat{f}^{whole}$, $\hat{f}^{(1)}$, and $\hat{f}^{(2)}$ will essentially there by allowing for a smaller variance in the split model by about a factor of 2.

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: $p(Y = 1 \mid \mathbf{X}) = \frac{1}{1+\exp\left\{\log\left(\frac{p(Y=0)p(\mathbf{X}|Y=0)}{p(Y=1)p(\mathbf{X}|Y=1)}\right)\right\}}$.

$\prod_i p(X_i \mid Y) = p(\mathbf{X} \mid Y) \because (X_i \perp\!\!\!\perp X_j \mid Y)\ \forall i \neq j$. Thus, $p(Y \mid \mathbf{X}) = \frac{p(Y)p(\mathbf{X} \mid Y)}{\sum_Y p(Y)p(\mathbf{X} \mid Y)}$. Performing some algebra

$$p(Y = 1|\mathbf{X}) = \frac{p(Y = 1)p(\mathbf{X}|Y = 1)}{\sum_Y p(Y)p(\mathbf{X}|Y)} = \frac{1}{1 + \frac{p(Y=0)p(\mathbf{X}|Y=0)}{p(Y=1)p(\mathbf{X}|Y=1)}} = \frac{1}{1 + \exp\left[\log\left(\frac{p(Y=0)p(\mathbf{X}|Y=0)}{p(Y=1)p(\mathbf{X}|Y=1)}\right)\right]}$$

Using $X_i \mid Y \sim \mathcal{N}(\mu_{i,y}, \sigma_i^2)$ and the above, we have that

$$\begin{aligned}\log\left(\frac{p(\mathbf{X} \mid Y = 0)}{p(\mathbf{X} \mid Y = 1)}\right) &= \sum_i -\frac{1}{2\sigma_i^2}\left((x_i - \mu_{i,0})^2 - (x_i - \mu_{i,1})^2\right) \\ &= -\sum_i \frac{\mu_{i,1} - \mu_{i,0}}{\sigma_i^2} x_i + \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{2\sigma_i^2}\end{aligned}$$

Define $\beta_{int} = \log\left(p(\mathbf{X} \mid Y = 1)/p(\mathbf{X} \mid Y = 0)\right) + \left(\mu_{i,0}^2 - \mu_{i,1}^2\right)/2\sigma_i^2$, and $\beta_i = \left(\mu_{i,1} - \mu_{i,0}\right)/\sigma_i^2$. Therefore, we can rewrite $p(Y = 1 \mid \mathbf{X})$ in the logistic regression form:

$$p(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + \exp\left(-\beta_{int} - \sum_{i=1} \beta_i x_i\right)}$$