Homework 1
BMI534 Introduction to Machine Learning,
CS534 Machine Learning
Spring 2024

The homework is due on **Feb. 13th at 11:59 PM ET** on Canvas.

For coding assignments, we will support Python 3.8. Any necessary packages need to be provided with `requirements.txt` and the code should be executable without any further dependencies. Please include the following *signed honor statement*:

```
/* THIS CODE IS MY OWN WORK, IT WAS WRITTEN WITHOUT
 CONSULTING CODE WRITTEN BY OTHER STUDENTS OR LARGE
 LANGUAGE MODELS LIKE CHATGPT. Your_Name_Here */

I collaborated with the following classmates for this
          homework:  <names of classmates>
                        Or
I have completed this homework without collaborating
                with any classmates.
```

This homework assignment tests knowledge in Statistical Decision Theory, Linear Regression, and Naïve Bayes. Show your work, include the code in a zipped file, and include a filled and signed copy of the honor pledge.

---

1 (Code) Regularizing linear regression (20 pts)

In this problem, we have a dataset that contains $p = 100$ features, but the underlying model that relates X, Y involves only 5 of these

$$Y = \beta_{j_1} X_{j_1} + \beta_{j_2} X_{j_2} + \ldots + \beta_{j_5} X_{j_5} + \epsilon \tag{1}$$

for some $j_1, j_2, \ldots, j_3 \in \{1, \ldots, 100\}$.

LASSO regression incorporates a model penalty in the loss function that effectively encourages a *sparse* solution, forcing many model weights β_j towards zero

$$Loss(\beta) = \sum_{i=1}^{N} (y_i - \beta x_i)^2 + \lambda |\beta|_1. \tag{2}$$

Since the L_1 norm is not differentiable, LASSO models can be derived using the sub-gradient method for gradient-descent. This sub-gradient method represents the gradient of the penalty term using a *soft thresholding* operation

$$S(\lambda, \beta_j) = \begin{cases} \beta_j - \lambda & \text{if } \beta_j > \lambda \\ 0 & \text{if } -\lambda \leq \beta_j \leq \lambda \\ \beta_j + \lambda & \text{if } \beta_j < -\lambda. \end{cases} \tag{3}$$

This sub-gradient term is combined with the gradient of the cost term $\sum_{i=1}^{N}(y_i - \beta x_i)^2$ for gradient descent.

(a) **(6 pts)** Find a LASSO model $\hat{\beta}$ to the training data using gradient descent. Use the penalty weight $\lambda = 1$ and learning rate $\gamma = 5e-3$ to train the model for 10000 gradient updates. Plot the final model coefficients. Can you guess the indices of the nonzero model weights $\{j_1, \ldots, j_5\}$?

(b) **(6 pts)** Fit a Least Squares model without regularization to the training set. Plot the final model coefficients.

(c) **(8 pts)** Compare the mean-square error of the LASSO and ordinary least squares models on the testing set.

2 (Written) Decision Theory (20 pts)

In classification, the loss function we usually want to minimize is the 0/1 loss:

$$l(f(x), y) = \mathbf{1}\{f(x) \neq y\} \tag{4}$$

where $f(x), y \in \{0, 1\}$ (i.e. binary classification). In this problem, we will consider the effect of using an asymmetric loss function:

$$l_{\alpha,\beta}(f(x), y) = \alpha\mathbf{1}\{f(x) = 1, y = 0\} + \beta\mathbf{1}\{f(x) = 0, y = 1\} \tag{5}$$

Under this loss function, the two types of errors received different weights, determined by $\alpha, \beta > 0$.

(a) **(6 pts)** Determine the Bayes optimal classifier $f(x)$, i.e. the classifier that achieves minimum Bayes risk assuming $P(x, y)$ is known, for the loss $l_{\alpha,\beta}$, where $\alpha, \beta > 0$.

(b) **(6 pts)** Suppose that the class $y = 0$ is extremely uncommon (i.e. $P(y = 0)$ is small). This means that the classifier $f(x) = 1$ for all x will have a good (low) risk. To overcome this unbalance problem, we may try to put the two classes on even footing by considering the following risk:

$$R = P(f(x) = 1|y = 0) + P(f(x) = 0|y = 1) \tag{6}$$

Show how this risk is equivalent to choosing a certain α, β and minimizing the risk where the loss function is $l_{\alpha,\beta}$?

(c) **(8 pts)** Consider the following classification problem. We first choose the label $Y \sim Bernoulli(\frac{1}{2})$, which is 1 with probability $\frac{1}{2}$. If $Y = 1$, then $X \sim Bernoulli(p)$; otherwise, $X \sim Bernoulli(q)$. Assume that $p > q$. What is the Bayes Optimal classifier, and what is its risk?

3 (Written) Naïve Bayes (20 pts)

Consider the spam-mail example that was covered in the lecture. For your benefit, it is restated with slightly different notations but essentially the same way. Given labeled training data $D = \{X_i, y_i\}$, where class label $y_i = c \in \{0, 1\}$, such that $y_i = 1$ for spam mail and $y_i = 0$ for ham mail and where $X_i = [x_{i1}, x_{i2}, \cdots, x_{iK}]$ is an indicator vector of all the words in the dictionary *i.e.* $x_{ij} = 1$ if the j word exist in the mail; otherwise, $x_{ij} = 0$ if it does not exist. There are K words in the dictionary. The probability of the ith word in class c is denoted as $\theta_{i,c}$ (to be clear class refers to either spam ($c = 1$) or ham ($c = 0$), and it was estimated (Bayes' estimate, conditional mean) from training D as

$$\hat{\theta}_{i,c} = \frac{n_{i,c} + 1}{n_c + 2} \tag{7}$$

where n_c is the count of class c mails and $n_{i,c}$ is the count of ith word in class c. Then, using $0 - 1$ Loss and minimizing the Bayes Risk (Mean risk) defined as

$$E[L(\hat{y}, y)] = \sum_{\hat{y}} \sum_{y} p(\hat{y}, y) L(\hat{y}, y) \tag{8}$$

where

$$L(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{if } \hat{y} \neq y \end{cases} \tag{9}$$

the decision rule (Bayes classifier)

$$\hat{y} = \arg\max_{\hat{y}} p(y = \hat{y}|X) \tag{10}$$

was derived

(a) **(5 pts)** What assumptions are made in the estimation of Equation 7? Be sure to be exact in your statement. (e.g. something has this probability distribution and the estimate is the $(\cdots)$ of the distribution).

(b) **(5 pts)** Derive the decision rule Equation 10 from Equation 8.

(c) **(5 pts)** Write the decision rule in terms of $p(X|y)$ and $p(y)$. Then assume that the occurrence of a word is independent of one another and write the decision rule in terms of $p(x_i|y)$ and $p(y)$, where $X = [x_i, \cdots, x_K]$.

(d) **(5 pts)** Compute the likelihood as

$$p(X|y, D) = \int_\theta p(X, \theta|y, D)d\theta \tag{11}$$

$$= \int_\theta p(X|\theta, y, D)p(\theta|D)d\theta \tag{12}$$

$$= \int_\theta p(X|\theta, y)p(\theta|D)d\theta \tag{13}$$

where it is assumed $p(X|\theta, y, D) = p(X|\theta, y)$. Assume that the occurrence of a word is independent of one another. You may need the following: $Beta(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$ and $\Gamma(n+1) = n\Gamma(n)$.

4 (Written) Linear Regression (20 pts)

The pth ordered linear regression model has the form

$$f(\mathbf{x}) = \sum_{j=0}^{p} x_j w_j = \mathbf{w}^T \mathbf{x} \tag{14}$$

where $\mathbf{x} = [x_0, x_1, \cdots, x_p]^T$ and $\mathbf{w} = [w_0, w_1, \cdots, w_p]^T$ denote input and coefficient vector, respectively. Assume $x_0 = 1$. Given the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T$ and $\mathbf{y} = [y_1, y_2, \cdot, y_N]^T$,

(a) **(3 pts)** Derive the ordinary least squared (OLS) estimate of $\mathbf{w}$, as a function of $\mathbf{X}$ and $\mathbf{y}$. Draw geometric interpretation and explain the meaning of the OLS estimate of $\mathbf{w}$.

(b) **(3 pts)** Assume that

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) \tag{15}$$

Derive the maximum likelihood (ML) estimate of $\mathbf{w}$ as a function of $\mathbf{X}$ and $\mathbf{y}$.

(c) **(3 pts)** Are the OLS estimate and the ML estimate of $\mathbf{w}$ same? What are the conditions the OLS estimate and the ML estimate are same?

(d) **(3 pts)** Ridge regression solves the following constrained optimization problem.

$$\hat{\mathbf{w}} = \arg\min_w \sum_{i=1}^N (y_i - w^T x_i)^2 \tag{16}$$

$$s.t.\ \|w\|^2 \leq s \tag{17}$$

Derive the ridge regression estimate of $\mathbf{w}$ as a function of $\mathbf{X}$ and $\mathbf{y}$.

(e) **(3 pts)** Assume a Gaussian prior,

$$\mathbf{w} \sim N(\mathbf{0}, \alpha^2\mathbf{I})$$

Derive the maximum a posterior (MAP) estimates of $\mathbf{w}$ as a function of $\mathbf{X}$ and $\mathbf{y}$.

(f) **(5 pts)** Lasso regression solves the following constrained optimization problem.

$$\hat{\mathbf{w}} = \arg\min_w \sum_{i=1}^N (y_i - w^T x_i)^2 \tag{18}$$

$$s.t.\ \|w\| \leq s \tag{19}$$

Explain briefly the difference between the ridge and lasso regression in terms of the estimate for $\mathbf{w}$.

5 (Written) Maximum Likelihood (20 pts)

You are playing a game with two coins. Coin 1 has a θ probability of heads. Coin 2 has a 2θ probability of heads. You flip these coins several times and record your results:

Coin	Result
1	Head
2	Tail
2	Tail
2	Tail
2	Head

(a) **(6 pts)** What is the likelihood of the data given θ?

(b) **(6 pts)** What is the maximum likelihood estimation for θ?

(c) **(8 pts)** A uniform distribution in the range of $[0, \theta]$ is given by

$$p(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

What is the maximum likelihood estimator of θ? (*Hint*: think of two cases, where $\theta < max(x^i)$ and $\theta \geq max(x^i)$ separately.)