Optimization and Machine Learning, Spring 2021
Homework 2

(Due Thursday, Apr. 1 at 11:59pm (CST))

April 10, 2021

1. [10 points] Given a set of training data $(x_1, y_1), \cdots, (x_N, y_N)$ from which to estimate the parameters β, where each $x_i = [x_{i1}, \cdots, x_{ip}]^T$ denotes a vector of feature measurements for the ith sample. Consider a linear regression problem in which we want to "weight"different training examples differently. Specifically, suppose we aim at minimizing

$$\mathrm{RSS}(\beta) = \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{1}$$

where the example-specific weights w_i $(i = 1, 2, ..., N)$ are given. (Note: assume that the the data has been centered, and thus we do not need to consider the intercept in the linear model.)

   (a) Represent $\mathrm{RSS}(\beta)$ in a matrix form. [2 points]
   Solution: $\mathbf{W}$ is a diagonal matrix with its i-th diagonal element being $\frac{1}{2}w_i$. Suppose we have the predictions $\hat{\boldsymbol{y}} = \mathbf{X}\beta$, $\mathrm{RSS}(\beta)$ is rewritten by

$$\mathrm{RSS}(\beta) = (\mathbf{X}\beta - \mathbf{y})^T\mathbf{W}(\mathbf{X}\beta - \mathbf{y}) + \lambda\beta^T\beta = (\hat{\mathbf{y}} - \mathbf{y})^T\mathbf{W}(\hat{\mathbf{y}} - \mathbf{y}) + \lambda\beta^T\beta$$

$$\begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}^T \begin{pmatrix} \frac{1}{2}w_1 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{2}w_2 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}w_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}w_N \end{pmatrix} \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}w_1(\hat{y}_1 - y_1) \\ \vdots \\ \frac{1}{2}w_N(\hat{y}_N - y_N) \end{bmatrix}^T \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix} + \lambda\sum_{j=1}^{p}\beta_j^2 = \frac{1}{2}\sum_{i=1}^{N} w_i\left(y_i - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2, .$$

   (b) Derive the closed form of the model in (a). [3 points]
   Solution: Make partial derivative on β :

$$\begin{aligned} \frac{\partial RSS(\beta)}{\partial\beta} &= \frac{\partial}{\partial\beta}(\mathbf{X}\beta - \mathbf{Y})^{\mathrm{T}}\mathbf{W}(\mathbf{X}\beta - \mathbf{Y}) + \frac{\partial}{\partial\beta}\left(\lambda\beta^{\mathrm{T}}\beta\right) \\ &= 2\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{X}\beta - \mathbf{Y}) + 2\lambda I_p\beta \\ &= 0 \end{aligned}$$

   Therefore, we can get:

$$\begin{aligned} \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta + \lambda I_p\beta &= \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Y} \\ \Rightarrow \beta &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + \lambda I_p\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Y} \end{aligned}$$

   (c) Suppose the y_i's were observed with differing variances. To be specific, suppose that

$$p(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma_i^2}\right), \tag{2}$$

i.e., y_i has mean $x_i^T\beta$ and variance σ_i^2, where the σ_i's are fixed, known, constants). Show that finding the Maximum Likelihood Estimation (MLE) of β is equivalent to solving a weight linear regression problem in (1) with $\lambda = 0$. State clearly what the w_i's are in terms of the σ_i's. [5 points]

Solution: The log likelihood function is

$$\mathcal{L}(\beta) = \log \prod_{i=1}^{N} p\left(y_i|x_i;\beta\right) = \sum_{i=1}^{N} \log\{\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(y_i - x_i^T\beta\right)^2}{2\sigma_i^2}\right)\} = \frac{-1}{\sqrt{2\pi}\sigma_i} \sum_{i=1}^{N} \frac{\left(y_i - x_i^\top\beta\right)^2}{2\sigma_i^2}.$$

Maximizing the likelihood is equivalent to minimizing $\sum_{i=1}^{N} \frac{\left(y_i - x_i^\top\beta\right)^2}{2\sigma_i^2}$. This is equivalent to solving a weigh linear regression problem with weight $w_i = \frac{1}{\sigma_i^2}$.

2. [10 points] Suppose that we have N training samples, in which each sample is composed of p input variable and one categorical label with K states. (Note: assume that the the data has been centered, and thus we do not need to consider the intercept in the linear model.)

(a) Please solve this multi-class classification problem by least squares, and discuss its limitation. [3 points]
Solution: Input:

$$\mathbf{X} = \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_N^T & - \end{bmatrix},$$

where x_i is the i-th observation with p parameter.
Output:

$$\mathbf{Y} = \begin{bmatrix} - & y_1^T & - \\ - & y_2^T & - \\ & \vdots & \\ - & y_N^T & - \end{bmatrix},$$

where $y_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T$, with its k-th element being 1, indicating that the k-th class is associated with the i-th observation x_i.
By minimizing the following objective function,

$$\min_{\mathbf{B}} ||\mathbf{XB} - \mathbf{Y}||_F^2,$$

we can get the solution $\mathbf{B} = (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{Y}$, if $\mathbf{X}^T\mathbf{X}$ is invertible.
It probably suffers from the problem of masking when $K > p$.

(b) It is widely known that Linear Discriminant Analysis (LDA) enables to overcome this limitation. Please derive the decision boundary of LDA between an arbitrary class-pair. [3 points]
Solution: Linear discriminant analysis (LDA). In LDA, the decision boundary between two arbitrary classes A and B is

$$\mathbf{x}^T\hat{\Sigma}^{-1}(\hat{\mu}_A - \hat{\mu}_B) + \left(\ln(\frac{\Pr(A)}{\Pr(B)}) - \frac{\hat{\mu}_A\hat{\Sigma}^{-1}\hat{\mu}_A - \hat{\mu}_B\hat{\Sigma}^{-1}\hat{\mu}_B}{2}\right) = 0.$$

(c) Please revise your model in (b) by either strength or weaken its assumptions, and tell the difference between your models in (b) and (c). [4 points]
Solution: We can use quadratic discriminative analysis (QDA) for classification.
Difference:

- LDA and QDA both assume that the class conditional probability distributions are normally distributed with different means μ_k, but LDA is different from QDA in that it requires all of the distributions to share the same covariance matrix Σ and QDA requires all of the distribution to have different covariance matrix Σ_k.
- The decision boundary is linear in LDA and quadratic in QDA.
- The number of estimated parameters is $p \times (K + p)$ in LDA and $K \times p \times (p + 1)$ in QDA.

3. [10 points] Given the input variables $X \in \mathbb{R}^p$ and a categorical output variable $G \in \{0, 1\}$, the Expected Prediction Error (EPE) is defined by

$$\text{EPE} = \mathbb{E}[L(Y, \hat{G}(X))],$$

where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, G)$, and $L(G, \hat{G}(X))$ is a loss function measuring the difference between the estimated $\hat{G}(X)$ and observed G.

(a) Given the zero-one loss

$$L(k, \ell) = \begin{cases} 1 & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell, \end{cases}$$

please derive the Bayes classifier $\hat{G}(x) = \text{argmax}_{k \in \{0,1\}} \Pr(G = k | X = x)$ by minimizing EPE. [3 points]
Solution: Without loss of generality, we consider $Y \in \{1, 2, ..., M\}$, and rewrite EPE as follows

$$\begin{aligned} \text{EPE} &= \mathbb{E}[L(Y, \hat{Y}(X)] \\ &= \int_x \left[\sum_{m=1}^{M} L\left(Y = m, \hat{Y}(x)\right) \Pr(Y = m | X = x) \right] dx \\ &= \int_x \left[1 - \Pr\left(Y = \hat{Y}(x) | X = x\right) \right] dx. \end{aligned}$$

Therefore,

$$\hat{Y}(x) = \text{argmin EPE} = \text{argmax}_{m \in \{1, \dots, M\}} \Pr(Y = m | X = x).$$

(b) Please define a function which enables to map the range of an arbitrary linear function to the range of a probability. [2 points]
Solution: Given an arbitrary liner function,

$$f(X) = \beta_0 + X^\top \beta \in (-\infty, +\infty),$$

the required function can be defined by

$$\Pr(Y | X) = \frac{\exp(f(X))}{1 + \exp(f(X))} \in (0, 1).$$

(c) Based on the function you defined in (b), please approximate the Bayes classifier in (a) by a linear function between X and Y, and derive its decision boundary. [5 points]
Solution: Based on (a), we have

$$\begin{aligned} \Pr(Y = 0 | X) &= \frac{\exp(f(X))}{1 + \exp(f(X))}, \\ \Pr(Y = 1 | X) &= 1 - \Pr(Y = 0 | X) = \frac{1}{1 + \exp(f(X))}. \end{aligned}$$

Thus, using the Bayes classifier in (a), we assign the label $Y = 0$ if the following conditions hold:

$$\begin{aligned} & 1 < \frac{\Pr(Y = 0 | X)}{\Pr(Y = 1 | X)} \\ \implies & 0 < \ln \exp(f(X)) \\ \implies & 0 < f(X), \end{aligned}$$

and assign $Y = 1$ otherwise. Hence, we obtain the linear decision boundary $\{X | \beta_0 + X^\top \beta = 0\}$.