Problem Set 2 Solutions

Prof. Oke

CEE 697M: Data Mining and Machine Learning for Engineers — *03.07.2022*

Due February 25, 2022 at 11:59PM. Submit via Moodle.

The standard problems are worth a total of **48 points**.

Problem 1 *Gaussian discriminant analysis (8 pts)*

(a) Assuming a distinct class covariance $\mathbf{\Sigma}_c$ in Gaussian discriminant analysis results in a quadratic decision boundary (QDA), which can easily overfit the data. LDA, which assumes a common covariance $\mathbf{\Sigma}$ across classes, results in a linear decision boundary, and can thus prevent overfitting. List two other approaches to prevent overfitting in Gaussian discriminant analysis. *[2]*

(i) Regularization of the covariance matrices (e.g., shrinkage towards a diagonal matrix; called "diagonal LDA")

(ii) Using MAP estimation: $\hat{\Sigma}_{\text{map}} = \lambda \text{diag}(\hat{\Sigma}_{\text{mle}}) + (1 - \lambda)\hat{\Sigma}_{\text{mle}}$

(b) Suppose we have features $x \in \mathbb{R}^p$, a two-class response with class sizes n_1, n_2 and the target coded as $-n/n_1, n/n_2$. Show that the LDA rule classifies to class 2 if *[8]*

$$x^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(n_2/n_1), \quad (1)$$

and class 1 otherwise. (*Hint:* First write the priors π_1 and π_2. Then write the discriminant functions δ_1 and δ_2. Knowing that the LDA classifier assigns an observation to class 2 when $\delta_2 > \delta_1$, expand this condition to obtain (1).)

We have $K = 2$ classes. The discrimant functions are thus:

$$\delta_1(x) = x^T \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^T \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1 + \log \pi_1$$
$$\delta_2(x) = x^T \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_2^T \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 + \log \pi_2$$

And the priors are:

$$\pi_1 = \frac{n_1}{n_1 + n_2}$$
$$\pi_2 = \frac{n_2}{n_1 + n_2}$$

The LDA classifier will assign an observation to class 2 if $\delta_2 > \delta_1$:

$$\begin{aligned}
\implies x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 + \log\pi_2 &> x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1 + \log\pi_1 \\
x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 - x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1 &> -\frac{1}{2}\hat{\mu}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1 + \log\pi_1 + \frac{1}{2}\hat{\mu}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 - \log\pi_2 \\
x^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) &> \frac{1}{2}\hat{\mu}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1 - \log\frac{\pi_2}{\pi_1} \\
&> \frac{1}{2}\left[\hat{\mu}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1 + \hat{\mu}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_1\right] - \log\frac{n_2}{n_1} \\
&> \frac{1}{2}\left[\hat{\mu}_2^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) + \hat{\mu}_1^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)\right] + \log\frac{n_2}{n_1} \\
&> \frac{1}{2}\left[(\hat{\mu}_2 + \hat{\mu}_1)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)\right] + \log\frac{n_2}{n_1} \quad \square
\end{aligned}$$

Problem 2 *Logistic regression I (4 pts)*

[4pts] In simple logistic regression with a multiple predictors $X^T = (1, X_1, \ldots, X_p)$, the logistic function is given by:

$$p(X) = \frac{1}{1 + e^{-\boldsymbol{w}^\intercal \boldsymbol{x}}} \tag{2}$$

where $\beta^T = (\beta_0, \beta_1, \ldots, \beta_p)$. Using this function, show explicitly that the log-odds or logit function of $p(X)$ is given by:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta^T X \tag{3}$$

First, the odds are given by:

$$odds = \frac{p(X)}{1 - p(X)} = \frac{\frac{e^{\beta^T X}}{1 + e^{\beta^T X}}}{1 - \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}}$$

We can rearange the denominator of the odds as follows:

$$1 - \frac{e^{\beta^T X}}{1 + e^{\beta^T X}} = \frac{1 + e^{\beta^T X}}{1 + e^{\beta^T X}} - \frac{e^{\beta^T X}}{1 + e^{\beta^T X}} = \frac{1 + e^{\beta^T X} - e^{\beta^T X}}{1 + e^{\beta^T X}} = \frac{1}{1 + e^{\beta^T X}}$$

Thus, we can express the odss as:

$$\frac{p(X)}{1 - p(X)} = \frac{\frac{e^{\beta^T X}}{1 + e^{\beta^T X}}}{\frac{1}{1 + e^{\beta^T X}}} = e^{\beta^T X}$$

Then, taking the natural log, we obtain:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta^T X = \log(e^{\beta^T X}) = \beta^T X$$

Problem 3 *Logistic regression II (8 pts)*

Problem 4 *Ridge regression (8 pts)*

$$
\begin{aligned}
J(\boldsymbol{w}, w_0) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1}) + \lambda\boldsymbol{w}^\top\boldsymbol{w} \\
&= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{w} - 2w_0\boldsymbol{y}^\top\mathbf{1} + \boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + 2w_0\boldsymbol{w}^\top\boldsymbol{X}^\top\mathbf{1} + w_0\mathbf{1}^\top\mathbf{1}w_0 + \lambda\boldsymbol{w}^\top\boldsymbol{w} \\
&= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{w} - 2w_0n\overline{y} + \boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + 2w_0\mathbf{1}^\top\boldsymbol{X}\boldsymbol{w} + nw_0^2 + \lambda\boldsymbol{w}^\top\boldsymbol{w} \\
&= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{w} - 2w_0n\overline{y} + \boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + 2w_0\left(\sum_n\sum_d x_{nd}w_d\right) + nw_0^2 + \lambda\boldsymbol{w}^\top\boldsymbol{w} \\
&= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{w} - 2w_0n\overline{y} + \boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \underbrace{2w_0n\overline{\boldsymbol{x}}^\top\boldsymbol{w}}_{=0;\quad \overline{\boldsymbol{x}}=0} + nw_0^2 + \lambda\boldsymbol{w}^\top\boldsymbol{w} \\
\therefore,\quad \frac{\partial J(\boldsymbol{w}, w_0)}{\partial w_0} &= -2n\overline{y} + 2nw_0 = 0 \\
\implies \quad \hat{w}_0 &= \overline{y} \\
\text{and:}\quad \frac{\partial J(\boldsymbol{w}, w_0)}{\partial \boldsymbol{w}} &= -2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + 2\lambda\boldsymbol{w} = 0 \\
(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{w} &= \boldsymbol{X}^\top\boldsymbol{y} \\
\implies \quad \hat{\boldsymbol{w}} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}
$$

Problem 5 *Exploration of ridge regression (8 pts)*

Consider the special case of performing regression without an intercept on a design matrix $\boldsymbol{X}$ with n rows (observations) and p columns (features). The following relationships hold:

$$
\begin{gathered}
n = p \\
x_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}
\end{gathered} \tag{4}
$$

(a) Show algebraically that the least squares solution is given by: *[3]*

$$\hat{\beta}_j = y_j \tag{5}$$

We recall that the least squares solution is given by:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

We note that:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Thus,

$$\begin{aligned}
\hat{\beta} &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\
&= \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{(The transpose and inverse of an identity matrix gives the identity matrix)}
\end{aligned}$$

Thus, $\hat{\beta}_j = y_j$.

[5] **(b)** The ridge regression estimate is given by:

$$\hat{\beta}^R = \arg\min_{\beta} \left[RSS^R(\beta) \right] = \arg\min_{\beta} \left\{ \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{6}$$

Show algebraically that the ridge solution is:

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda} \tag{7}$$

We recall that the least squares solution is given by:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

We note that:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Thus,

$$\begin{aligned}
\hat{\beta} &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\
&= \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{(The transpose and inverse of an identity matrix gives the identity matrix)}
\end{aligned}$$

Thus, $\hat{\beta}_j = y_j$.

Problem 6 *Poisson regression (12 pts)*

The Poisson regression model is given by:

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathrm{Poi}(y_n| \exp(\boldsymbol{w}^\intercal \boldsymbol{x}_n)) = \frac{\exp(-\mu_n)\mu_n^{y_n}}{y_n!} \tag{8}$$

where $y_n \in \{0, 1, 2, \ldots\}$ is a count response, $\boldsymbol{x}_n$ is a vector of predictors, $\boldsymbol{w}$ is the weight vector and $\mu_n = \exp(\boldsymbol{w}^\intercal \boldsymbol{x}_n)$.

(a) Write the model in GLM form $\exp(y_n\eta_n - A(\eta_n) + h(y_n))$, identifying the canonical parameter η_n, the log partition function $A(\eta_n)$ and the base measure $h(y_n)$. *[8]*

(b) Derive the mean function $\ell^{-1}(\eta_n)$ by taking the derivative of $A(\eta_n)$. *[2]*

(c) What is the link function $\ell(\mu_n)$? *[1]*

(d) If the natural parameter of the Poisson distribution is $\eta_n = \log(\mu_n)$, is the link function canonical? *[1]*