Problem Set 2 **Oke**

CEE 616: Probabilistic Machine Learning *10.09.2025*

Due October 16, 2025 at 11:59PM. Submit via Canvas as a PDF.

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]*

(b) Suppose we have features $x \in \mathbb{R}^D$, 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 *[6]*

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

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).)

Problem 2 *Logistic regression I (4 pts)*

In a binary logistic regression with a multiple predictors $\boldsymbol{x}^\intercal = (1, x_1, \ldots, x_D)$, the logistic function is given by: *[4]*

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

where $\boldsymbol{w}^\intercal = (b, w_1, \ldots, w_D)$. Using this function, show explicitly that the log-odds or logit function of $p(y = 1|\boldsymbol{x})$ is given by:

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \boldsymbol{w}^\intercal \boldsymbol{x} \tag{3}$$

Problem 3 *Logistic regression II (8 pts)*

PMLI Exercise 10.2 (a) - (d)

Problem 4 *Ridge regression (8 pts)*

PMLI Exercise 11.2

Problem 4 *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 D columns (features). The following relationships hold:

$$N = D$$
$$x_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{4}$$

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

$$\hat{\boldsymbol{w}}_j = y_j \tag{5}$$

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

$$\hat{\boldsymbol{w}}^R = \arg\min_{\boldsymbol{w}} \left[RSS^R(\boldsymbol{w})\right] = \arg\min_{\boldsymbol{w}} \left\{||(y_j - \boldsymbol{w}_j)||_2^2 + \lambda||\boldsymbol{w}_j||_2^2\right\} \tag{6}$$

Show algebraically that the ridge solution is:

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

Problem 6 *Poisson regression (12 pts)*

The Poisson regression model is given by:

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

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 $\mu_n = \exp(\boldsymbol{w}^\intercal \boldsymbol{x}_n)$.

[8] **(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 $h(y_n)$.

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

[1] **(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?