Machine Learning 2024 - Sheet 2.2

Isabel Valera

Notation. The input feature vector of the i-th sample, i.e., $\mathbf{x}_i \in \mathbb{R}^D$, can be used to construct feature matrix for N samples represented as $\mathbf{X} \in \mathbb{R}^{N \times D}$. The input vector can be represented by a basis function $\Phi(\mathbf{x}_i) \in \mathbb{R}^M$. The feature matrix for N samples is then represented as $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$. The target vector of the i-th sample, i.e., $y_i \in \mathbb{R}$, can be used to construct column vector for N samples represented as $\mathbf{Y} \in \mathbb{R}^N$. $\mathbf{\Sigma}$ or $\mathbf{\Sigma}_W$ refers to within-class covariance, whereas $\mathbf{\Sigma}_B$ refers to between-class covariance

Exercise 1: Sigmoid: the beginning ⚡

Show that the logistic sigmoid function given in 1 satisfies the following properties:

i) $\sigma(-a) = 1 - \sigma(a)$

ii) $\sigma^{-1}(y) = \ln \frac{y}{1-y}$

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{1}$$

Exercise 2: Sigmoid: the posterior ⚡⚡

Assume that we have a classifier that can decide if an input belongs to $\mathcal{C}_1$ or $\mathcal{C}_2$. We can write the posterior probability for class $\mathcal{C}_1$ given a sample $\mathbf{x}$:

$$p(\mathcal{C}_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)} \tag{2}$$

$$= \frac{1}{1 + \exp -a} = \sigma(a) \tag{3}$$

Where a is given by 4

$$a = \ln \frac{p(\mathbf{x} \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} \mid \mathcal{C}_2)p(\mathcal{C}_2)} \tag{4}$$

Note: Assume that class-conditional densities are Gaussians and all classes have same covariance matrix $\mathbf{\Sigma}$.

i) Derive the result of equation 5 for the posterior class probability in the two-class generative model with Gaussian densities.

$$p(\mathcal{C}_1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) \tag{5}$$

In your results, verify Equations 6 and 7 for the parameters $\mathbf{w}$ and b.

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{6}$$

$$b = -\frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \tag{7}$$

ii) Comment on the effect of prior probabilities in the derived result.

Solutions:

ii) Prior densities are only effecting the bias term of our classifier. Therefore, we can say that the prior densities are only effecting the decision threshold, i.e. moving the decision boundary.

Exercise 3: Sigmoid: the derivative ⚡

Verify the relation given in 8 for the derivative of the logistic sigmoid function defined by equation 1.

$$\frac{d\sigma}{da} = \sigma(1-\sigma) \tag{8}$$

Exercise 4: Sigmoid: the error ⚡⚡

By making use of the result 8 for the derivative of the logistic sigmoid, show that the derivative of the error function given in 9 for the logistic regression model is given by equation 10.

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{w}) = -\ln p(\mathbf{Y} \mid \mathbf{w}) = -\sum_{n=1}^{N} [y_n \ln \sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n) + (1-y_n)\ln(1-\sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n)))] \tag{9}$$

$$\nabla_{\mathbf{w}} L(\boldsymbol{\Phi}, \mathbf{y}, \mathbf{w}) = \sum_{n=1}^{N} (\sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n)) - y_n)\Phi(\mathbf{x}_n) \tag{10}$$

Exercise 5: Binary cross entropy loss ⚡⚡

We have a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ in which features $\mathbf{x}$ are measurements taken from an product and y are labels showing if the product is damaged $y = 1$ or undamaged $y = 0$. For this task, we can learn logistic regression model with the loss

$$L(\boldsymbol{\Phi}, \mathbf{y}, \mathbf{w}) = -\ln p(\mathbf{y} \mid \mathbf{w}) = -\sum_{n=1}^{N} [y_n \ln \sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n)) + (1-y_n)\ln(1-\sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n)))] \tag{11}$$

where σ is a sigmoid activation, $p(\hat{y} = 1|\mathbf{x}_n) = \sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n))$ and the optimal parameters can be written as

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} L(\mathbf{w}) \tag{12}$$

i) Assume that our model works but it predicts too many false predictions with the label “damaged”. Propose a cost sensitive cost cross entropy loss so that we can solve this issue.

ii) After deriving a cost sensitive loss, propose an adjustment for the costs to improve the precision of the model.

iii) Given the prediction outputs in Table 1, construct the confusion matrix and compute accuracy, precision, recall, F1 and F2 scores.

	1	2	3	4	5	6
True label	0	0	0	1	1	1
Prediction	0	1	1	1	1	1

Table 1: True labels and predictions for six sample from the test set.

Solutions:

i)

$$L(\mathbf{X}, \mathbf{y}, \mathbf{w}) = -\sum_{n=1}^{N} \Big[C_1 y_n \ln p(\hat{y} = 1|\mathbf{x}_n) + C_2(1 - y_n)\ln(1 - p(\hat{y} = 1|\mathbf{x}_n)) \Big]$$

where $p(\hat{y} = 1|\mathbf{x}_n) = \sigma(\mathbf{w}^\top \Phi(\mathbf{x}_n))$. Observe that $C_1 = 1, C_2 = 1$ recovers to original cross entropy loss. Where C_1 is the cost of misclassifying a positive instance as negative, and C_2 is the cost of misclassifying a negative instance as positive.

ii) If we want to improve the precision, i.e., $PPV = \frac{TP}{TP+FP}$. We would like to have less false positives, thus $C_2 > C_1$.

iii) Let's compute the metrics

	Pos. prediction	Neg. prediction
Pos. label	3	0
Neg. label	2	1

Table 2: Confusion matrix.

True Positive $= 3$, **True Negative** $= 1$, **False Positive** $= 2$, **False Negative** $= 0$

Accuracy $= 4/6$, **Precision** $= 3/5$, **Recall** $= 3/3$, **F1-score** $= 2\dfrac{0.6 \times 1}{0.6 + 1}$

For F2-score, we pick $\beta = 2$, $F2 - score = 5\frac{0.6\times 1}{4\times 0.6+1}$

Exercise 6: Linearly separable ⚡⚡⚡

Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\mathbf{w}$ whose decision boundary $\mathbf{w}^\top \phi(\mathbf{x}) = 0$ separates the classes $(\mathcal{C}_1, \mathcal{C}_2)$ and then taking the magnitude of $\mathbf{w}$ to infinity.

Exercise 7: Linear Discriminant Analysis ⚡⚡⚡⚡

Suppose we have features $\mathbf{x} \in \mathbb{R}^d$, a two-class response C_1, C_2, with class sizes N_1, N_2, and the target coded as $-\frac{N}{N_1}, \frac{N}{N_2}$. Assume that class conditional probabilities are Gaussians with a common covariance matrix. $\mathbf{\Sigma}$ stands for within class covariance, $\mathbf{\Sigma}_B$ stands for between class covariance; with the definitions

$$\mathbf{\Sigma} = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k}^{N_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \tag{13}$$

$$\mathbf{\Sigma}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top \tag{14}$$

i) Show that the LDA rule classifies to class 2 if the equation below holds and to class 1 otherwise.

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) > \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \ln \frac{N_2}{N_1} \tag{15}$$

ii) Consider minimization of the least squares criterion $\sum_{i=1}^{N}(y_i - \mathbf{w}^\top \mathbf{x}_i - w_0)^2$. Show that the solution w satisfies

$$\left(\boldsymbol{\Sigma} + \frac{N_1 N_2}{N}\boldsymbol{\Sigma}_B\right)\mathbf{w} = N(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

with the definitions states in Equation 14.

iii) Hence show that $\boldsymbol{\Sigma}_B \mathbf{w}$ is in the direction $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and thus $\mathbf{w} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Therefore the least-squares regression coefficient is identical to the LDA coefficient, up to a scalar multiple.

iv) Show that the maximum of $J(\mathbf{w}) = \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_B \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}$ is given by $\boldsymbol{\Sigma}_B \mathbf{w} = \lambda \boldsymbol{\Sigma} \mathbf{w}$ where $\lambda = \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_B \mathbf{w}}{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}$. Show that $\mathbf{w} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.

v) Find the solution w_0 (up to the same scalar multiple as before), and hence the predicted value $f(\mathbf{x}) = w_0 + \mathbf{x}^\top \mathbf{w}$. Consider the following rule: classify to class 2 if $f(\mathbf{x}) > 0$ and class 1 otherwise. Show this is the not the same as the LDA rule unless the classes have equal numbers of observations.

Exercise 8: One-of-K ⚡⚡⚡⚡⚡

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditional densities $p(\Phi|\mathcal{C}_k)$ where Φ is the input feature vector. Suppose we are given a training data set $\{\Phi(\mathbf{x}_n), \mathbf{y}_n\}$ where $n = 1, ..., N$ and $\mathbf{y}_n$ is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $y_{nj} = \boldsymbol{I}_{jk}$ if pattern n is from class $\mathcal{C}_k$ where $\boldsymbol{I}_{K\times K}$ identity matrix.

i) Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is $\pi_k = \frac{N_k}{N}$ where N_k is the number of data points assigned to class $\mathcal{C}_k$.

ii) Consider the same classification model and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix so that $p(\Phi|\mathcal{C}_k) = \mathcal{N}(\Phi|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. Show that the maximum likelihood solution for the mean of the Gaussian distribution for class $\mathcal{C}_k$ is given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} y_{nk}\Phi_n$$

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*. Springer series in statistics New York, 2001.