ASSIGNMENT 9 SOLUTION

Nayan Man Singh Pradhan

Excercise 1. (Classification by linear regression on the indicator matrix)

We would like to solve a simple classification problem using paper and pencil. To this end, you are given the training data

$$\mathcal{T} = \{(0.3, 1), (1.8, 1), (1.5, 1), (4.8, 2), (2.6, 2)\} .$$

The objective is to predict the class labels for the two evaluation points $x_1 = 2.4$ and $x_2 = 6.2$. Use linear regression on the indicator matrix to build the necessary predictor and evaluate it at x_1 and x_2. In addition, evaluate the *training error* using the 0-1 loss.

(4 Points)

$$T' = \left\{\left(0.3, \begin{pmatrix}1\\0\end{pmatrix}\right), \left(1.8, \begin{pmatrix}1\\0\end{pmatrix}\right), \left(1.5, \begin{pmatrix}1\\0\end{pmatrix}\right), \left(4.8, \begin{pmatrix}0\\1\end{pmatrix}\right), \left(2.6, \begin{pmatrix}0\\1\end{pmatrix}\right)\right\}$$

$$T^{(1)} = \{(0.3, 1), (1.8, 1), (1.5, 1), (4.8, 0), (2.6, 0)\}$$

$$T^{(2)} = \{(0.3, 0), (1.8, 0), (1.5, 0), (4.8, 1), (2.6, 1)\}$$

for $T^{(1)}$

$$X = \begin{pmatrix} 1 & 0.3 \\ 1 & 1.8 \\ 1 & 1.5 \\ 1 & 4.8 \\ 1 & 2.6 \end{pmatrix}, \quad X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1.8 & 1.5 & 4.8 & 2.6 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1.8 & 1.5 & 4.8 & 2.6 \end{pmatrix}_{2\times5} \begin{pmatrix} 1 & 0.3 \\ 1 & 1.8 \\ 1 & 1.5 \\ 1 & 4.8 \\ 1 & 2.6 \end{pmatrix}_{5\times2} = \begin{pmatrix} 5 & 11 \\ 11 & 35.38 \end{pmatrix}$$

$$X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1.8 & 1.5 & 4.8 & 2.6 \end{pmatrix}_{2\times5} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}_{5\times1} = \begin{pmatrix} 3 \\ 3.6 \end{pmatrix}$$

$$X^T X \hat{\beta} = X^T y$$

$$\begin{pmatrix} 5 & 11 \\ 11 & 35.38 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3.6 \end{pmatrix}$$

$$\beta_1 = 1.190, \quad \beta_2 = -0.268, \quad \therefore \hat{\beta} = \begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix}$$

$$Z = \begin{pmatrix} 1 & 2.4 \\ 1 & 6.2 \end{pmatrix}$$

$$\hat{y}_1 = Z\hat{\beta} = \begin{pmatrix} 1 & 2.4 \\ 1 & 6.2 \end{pmatrix}\begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix} = \begin{pmatrix} 1.190 - 2.4 \times 0.268 \\ 1.190 - 6.2 \times 0.268 \end{pmatrix} = \begin{pmatrix} 0.5468 \\ -0.4716 \end{pmatrix}$$

for $T^{(2)}$

$$X = \begin{pmatrix} 1 & 0.3 \\ 1 & 1.8 \\ 1 & 1.5 \\ 1 & 4.8 \\ 1 & 2.6 \end{pmatrix}, \quad X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1.8 & 1.5 & 4.8 & 2.6 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1.8 & 1.5 & 4.8 & 2.6 \end{pmatrix}_{2\times 5}\begin{pmatrix} 1 & 0.3 \\ 1 & 1.8 \\ 1 & 1.5 \\ 1 & 4.8 \\ 1 & 2.6 \end{pmatrix}_{5\times 2} = \begin{pmatrix} 5 & 11 \\ 11 & 35.38 \end{pmatrix}$$

$$X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.3 & 1.8 & 1.5 & 4.8 & 2.6 \end{pmatrix}_{2\times 5}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}_{5\times 1} = \begin{pmatrix} 2 \\ 7.4 \end{pmatrix}$$

$$X^T X \hat{\beta} = X^T y$$

$$\begin{pmatrix} 5 & 11 \\ 11 & 35.38 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 7.4 \end{pmatrix}$$

$\beta_1 = -0.190$, $\beta_2 = 0.268$, $\hat{\beta} = \begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix}$

$$Z = \begin{pmatrix} 1 & 2.4 \\ 1 & 6.2 \end{pmatrix}$$

$$\hat{y}_2 = Z\hat{\beta} = \begin{pmatrix} 1 & 2.4 \\ 1 & 6.2 \end{pmatrix}\begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix} = \begin{pmatrix} -0.190 + 2.4 \times 0.268 \\ -0.190 + 6.2 \times 0.268 \end{pmatrix} = \begin{pmatrix} 0.4532 \\ 1.4716 \end{pmatrix}$$

Copying the result of $\hat{y}_1$ from the top, $\hat{y}_1 = \begin{pmatrix} 0.5468 \\ -0.4716 \end{pmatrix}$

For point $x_1 = 2.4$, we see $y_{11} > y_{21}$, ie $0.5468 > 0.4532$.
$\therefore$ x_1 is classified into class 1.

For point $x_2 = 6.2$, we see $y_{22} > y_{12}$, ie $1.4716 > -0.4716$.
$\therefore$ x_2 is classified into class 2.

Training Error Using 0-1 loss

i) $y_1 = (1 \;\; 0.3)\begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix} = 1.1096$, $y_2 = (1 \;\; 0.3)\begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix} = -0.1096$

∴ Classified class = 1, Actual class = 1, 0-1 loss = 0

ii) $y_1 = (1 \;\; 1.8)\begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix} = 0.7076$, $y_2 = (1 \;\; 1.8)\begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix} = 0.2924$

∴ Classified class = 1, Actual class = 1, 0-1 loss = 0

iii) $y_1 = (1 \;\; 1.5)\begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix} = 0.788$, $y_2 = (1 \;\; 1.5)\begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix} = 0.212$

∴ Classified class = 1, Actual class = 1, 0-1 loss = 0

iv) $y_1 = (1 \;\; 4.8)\begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix} = -0.0964$, $y_2 = (1 \;\; 4.8)\begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix} = 1.0964$

∴ Classified class = 2, Actual class = 2, 0-1 loss = 0

v) $y_1 = (1 \;\; 2.6)\begin{pmatrix} 1.190 \\ -0.268 \end{pmatrix} = 0.4932$, $y_2 = (1 \;\; 2.6)\begin{pmatrix} -0.190 \\ 0.268 \end{pmatrix} = 0.5068$

∴ Classified class = 2, Actual class = 2, 0-1 loss = 0

∴ Training Error $= \frac{1}{5}(0+0+0+0+0)$

$= 0$

Excercise 2. (Classification by linear discriminant analysis)

We again start from the training data set

$$\mathcal{T}_{train} = \{(0.3, 1), (1.8, 1), (1.5, 1), (4.8, 2), (2.6, 2)\}$$

for a paper and pencil classification task. In addition, you are given the validation set

$$\mathcal{T}_{val} = \{(1.6, 1), (1.9, 2), (2.5, 2)\} .$$

a) Use linear discriminant analysis to build a classifier based on the training data.

b) Evaluate the generalization error for the just constructed predictor using the 0-1 loss and the validation set approach.

(4 Points)

a) $T_{train} = \{(0.3,1),(1.8,1),(1.5,1),(4.8,2),(2.6,2)\}$

$N_1 = 3 \qquad N_2 = 2 \qquad N = 5$

$p_a(1) = \frac{3}{5}$, $p_a(2) = \frac{2}{5}$

$\mu_1 = \frac{1}{3}(0.3 + 1.8 + 1.5) = 1.2$

$\mu_2 = \frac{1}{2}(4.8 + 2.6) = 3.7$

$$\hat{\Sigma} = \sum_{g=1}^{2} \sum_{(x,g)\in T_{train}} \frac{(x-\mu_g)^2}{(N-r)} = \frac{1}{3}\Big((0.3-1.2)^2 + (1.8-1.2)^2 + (1.5-1.2)^2 + (4.8-3.7)^2 + (2.6-3.7)^2\Big)$$

$$= 1.22\overline{6}$$

$$\hat{g} = \underset{g\in R_a}{\operatorname{argmax}} \left(\ln \hat{p_a}(g) + x^T \hat{\Sigma}^{-1} \hat{\mu}_g - \frac{1}{2}\hat{\mu}_g^T \Sigma^{-1} \hat{\mu}_g\right)$$

$$= \underset{g\in R_a}{\operatorname{argmax}} \left(\ln \hat{p_a}(g) + x^T (1.226)^{-1} \hat{\mu}_g - \frac{1}{2}\hat{\mu}_g^T (1.226)^{-1} \hat{\mu}_g\right)$$

$g=1$: $\ln\left(\frac{3}{5}\right) + x^T\left(\frac{75}{92}\right) \times 1.2 - \frac{1}{2} \times 1.2 \times \left(\frac{75}{92}\right) \times 1.2$

$g=2$: $\ln\left(\frac{2}{5}\right) + x^T\left(\frac{75}{92}\right) \times 3.7 - \frac{1}{2} \times 3.7 \times \left(\frac{75}{92}\right) \times 3.7$

b) $\mathcal{T}_{val} = \{(1.6, 1), (1.9, 2), (2.5, 2)\}$.

for (1.6, 1),

$g=1$: $\ln\left(\frac{3}{5}\right) + 1.6 \times \left(\frac{75}{92}\right) \times 1.2 - \frac{1}{2} \times 1.2 \times \left(\frac{75}{92}\right) \times 1.2 = 0.467$

$g=2$: $\ln\left(\frac{2}{5}\right) + 1.6 \times \left(\frac{75}{92}\right) \times 3.7 - \frac{1}{2} \times 3.7 \times \left(\frac{75}{92}\right) \times 3.7 = -1.670$

$\therefore \underset{g \in R}{\text{argmax}} = 1$.

Since actual class = 1, 0-1 loss = 0

for (1.9, 2)

$g=1$: $\ln\left(\frac{3}{5}\right) + 1.9 \times \left(\frac{75}{92}\right) \times 1.2 - \frac{1}{2} \times 1.2 \times \left(\frac{75}{92}\right) \times 1.2 = 0.760$

$g=2$: $\ln\left(\frac{2}{5}\right) + 1.9 \times \left(\frac{75}{92}\right) \times 3.7 - \frac{1}{2} \times 3.7 \times \left(\frac{75}{92}\right) \times 3.7 = -0.765$

$\therefore \underset{g \in R}{\text{argmax}} = 1$

Since actual class = 2, 0-1 loss = 1

for (2.5, 2)

$g=1$: $\ln\left(\frac{3}{5}\right) + 2.5 \times \left(\frac{75}{92}\right) \times 1.2 - \frac{1}{2} \times 1.2 \times \left(\frac{75}{92}\right) \times 1.2 = 1.347$

$g=2$: $\ln\left(\frac{2}{5}\right) + 2.5 \times \left(\frac{75}{92}\right) \times 3.7 - \frac{1}{2} \times 3.7 \times \left(\frac{75}{92}\right) \times 3.7 = 1.044$

$\therefore \underset{g \in R}{\text{argmax}} = 1$.

Since actual class = 2, 0-1 loss = 1

Generalization Error = $\frac{1}{3}(0+1+1) = 0.67$ //

Excercise 3. (Training of logistic regression)

Prove Lemma 8.2 from the lecture.

(4 Points)

Method 12 ((Multinomial) Logistic regression) We start from given training data $\{(\boldsymbol{x}_i, g_i)\}_{i=1}^N$, $g_i \in \{1, \dots, r\}$. In *(multinomial)*[2)] *logistic regression*, we construct for $g = 1, \dots, r$ approximations to the class posterior $p(g|\boldsymbol{x})$. These are obtained by first introducing r linear models

$$s_g(\boldsymbol{x}) = \beta_0^{(g)} + \sum_{i=1}^{D} \beta_i^{(g)} x_i \qquad \text{for } g = 1, \dots, r\,,$$

hence one model per class (posterior). Together, the set of coefficients of all models is

$$\mathcal{B} = \left\{ \beta_i^{(g)} \right\}_{g=1,\dots r, i=0,\dots,D}\,.$$

Then, we apply to each of the linear models the so-called *softmax function* and approximate the class posteriors by

$$p_{\mathcal{B}}(g|\boldsymbol{x}) \approx \frac{\exp(s_g(\boldsymbol{x}))}{\sum_{h=1}^{r} \exp(s_h(\boldsymbol{x}))} \qquad \text{for } g = 1, \dots, r\,.$$

It remains to find "appropriate" weights $\mathcal{B}$. △

Theorem 8.4 With the setting of Method 12, a given training set $\{(\boldsymbol{x}_i, g_i)\}_{i=1}^N$ and the use of the cross entropy loss, the functional that we need to minimize to compute the coefficients $\mathcal{B}$ is given by

$$\begin{aligned} J(\mathcal{B}) &= \sum_{i=1}^{N} L_{CE}(p(\cdot|\boldsymbol{x}_i), p_{\mathcal{B}}(\cdot|\boldsymbol{x}_i)) \\ &= -\sum_{i=1}^{N} \sum_{g=1}^{r} p(g|\boldsymbol{x}_i) \log(p_{\mathcal{B}}(g|\boldsymbol{x}_i)) \\ &= -\sum_{i=1}^{N} \left[\left(\sum_{g=1}^{r} p(g|\boldsymbol{x}_i) s_g(\boldsymbol{x}) \right) - \log \left(\sum_{h=1}^{r} \exp(s_h(\boldsymbol{x})) \right) \right] \end{aligned}$$

Lemma 8.2 With the setting from Theorem 8.4, the gradient of the functional $J(\mathcal{B})$ with respect to a fixed class label g, which we call "∇_g", is given by

$$\nabla_g J(\mathcal{B}) = \begin{pmatrix} \frac{\partial}{\partial \beta_0^{(g)}} J(\mathcal{B}) \\ \vdots \\ \frac{\partial}{\partial \beta_D^{(g)}} J(\mathcal{B}) \end{pmatrix} = \frac{1}{N} \sum_{i=1}^{N} (p(g|\boldsymbol{x}_i) - s_g(\boldsymbol{x}_i)) \begin{pmatrix} 1 \\ \boldsymbol{x}_i \end{pmatrix}\,.$$