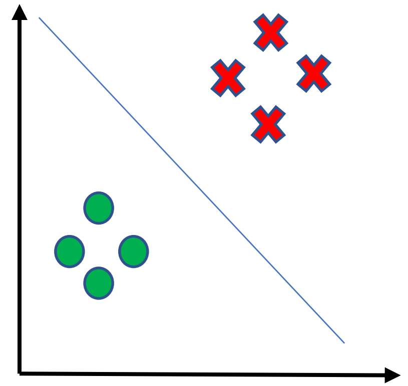
Logistic Regression

Spring 2024

Multi-class Classification

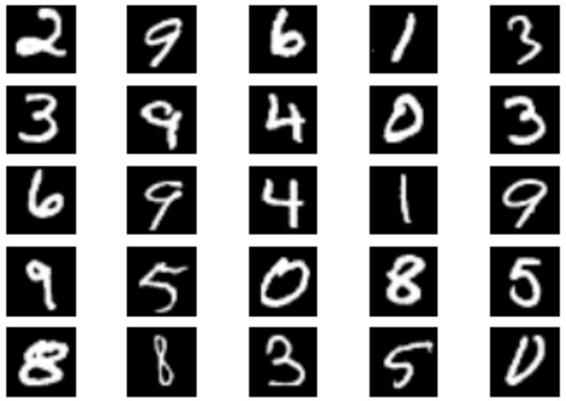
- Binary Classification
  - Only two classes
  - E.g. fraud detection



Given n samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Learn a mapping function $x_i \xrightarrow{f(x)} \begin{cases} 0 \\ 1 \end{cases}$

- Multi-class Classification
  - Multiple classes



Given n samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Learn a mapping function $x_i \xrightarrow{f(x)} \begin{cases} 0 \\ 1 \\ 2 \\ \\ K\text{-}1 \end{cases}$

Multi-class Classification

- Each class has a linear mapping function

$$z_{i,1} = \mathbf{w}_1^T \mathbf{x}_i = w_{1,0} + w_{1,1} x_{i,1} + w_{1,2} x_{i,2} + \cdots w_{1,d} x_{i,d}$$

$$z_{i,2} = \mathbf{w}_2^T \mathbf{x}_i = w_{2,0} + w_{2,1} x_{i,1} + w_{2,2} x_{i,2} + \cdots w_{2,d} x_{i,d}$$

$$z_{i,3} = \mathbf{w}_3^T \mathbf{x}_i = w_{3,0} + w_{3,1} x_{i,1} + w_{3,2} x_{i,2} + \cdots w_{3,d} x_{i,d}$$

$$\cdots$$

$$z_{i,K} = \mathbf{w}_K^T \mathbf{x}_i = w_{K,0} + w_{K,1} x_{i,1} + w_{K,2} x_{i,2} + \cdots w_{K,d} x_{i,d}$$

  - Each class has a parameter vector

$$\mathbf{w}_k \in \mathbb{R}^{d+1}$$

Multi-class Classification

- Each class has a linear mapping function

$$z_{i,1} = \mathbf{w}_1^T \mathbf{x}_i = w_{1,0} + w_{1,1} x_{i,1} + w_{1,2} x_{i,2} + \cdots w_{1,d} x_{i,d}$$

$$z_{i,2} = \mathbf{w}_2^T \mathbf{x}_i = w_{2,0} + w_{2,1} x_{i,1} + w_{2,2} x_{i,2} + \cdots w_{2,d} x_{i,d}$$

$$z_{i,3} = \mathbf{w}_3^T \mathbf{x}_i = w_{3,0} + w_{3,1} x_{i,1} + w_{3,2} x_{i,2} + \cdots w_{3,d} x_{i,d}$$

...

$$z_{i,K} = \mathbf{w}_K^T \mathbf{x}_i = w_{K,0} + w_{K,1} x_{i,1} + w_{K,2} x_{i,2} + \cdots w_{K,d} x_{i,d}$$

K parameter vectors

- Comparison with binary classification

$$z_i = \mathbf{w}^T \mathbf{x}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$

Only one parameter vector

Multi-class Classification

- Matrix representation

$$\begin{bmatrix} z_{i,1} \\ z_{i,2} \\ z_{i,3} \\ \cdots \\ z_{i,K} \end{bmatrix} = \begin{bmatrix} w_{1,0} & w_{1,1} & w_{1,2} & \cdots & w_{1,d} \\ w_{2,0} & w_{2,1} & w_{2,2} & \cdots & w_{2,d} \\ w_{3,0} & w_{3,1} & w_{3,2} & \cdots & w_{3,d} \\ \cdots & & & & \\ w_{K,0} & w_{K,1} & w_{K,2} & \cdots & w_{K,d} \end{bmatrix} \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ \cdots \\ x_{i,d} \end{bmatrix}$$

Parameter matrix

Feature vector of the i-th sample

$$\mathbf{z}_i = W^T \mathbf{x}_i$$

Multi-class Classification

- How to get the class probability?
  - K output (K classes)

$$p(y=1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(\mathbf{w}^T\mathbf{x})}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

$$p(y=0|\mathbf{x}) = 1 - p(y=1|\mathbf{x}) = 1 - \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

Binary classification with sigmoid function

$$p(y=1|x) = ?$$

$$p(y=2|x) = ?$$

$$p(y=3|x) = ?$$

$\ldots$

$$p(y=K|x) = ?$$

Multi-class Classification

- How to get the class probability?
  - Softmax function

Given a vector $\mathbf{z} = [z_1, z_2, \cdots, z_K] \in \mathbb{R}^K$,

$$\text{softmax}(\mathbf{z}) = \begin{bmatrix} \frac{\exp(z_1)}{\sum_{i=1}^K \exp(z_i)} \\ \frac{\exp(z_2)}{\sum_{i=1}^K \exp(z_i)} \\ \frac{\exp(z_3)}{\sum_{i=1}^K \exp(z_i)} \\ \cdots \\ \frac{\exp(z_K)}{\sum_{i=1}^K \exp(z_i)} \end{bmatrix} \begin{matrix} \rightarrow \text{Prob for class 1} \\ \rightarrow \text{Prob for class 2} \\ \rightarrow \text{Prob for class 3} \\ \\ \rightarrow \text{Prob for class K} \end{matrix}$$

- Properties:

$$0 < \frac{\exp(z_j)}{\sum_{i=1}^K \exp(z_i)} < 1$$

$$\sum_{j=1}^n \frac{\exp(z_j)}{\sum_{i=1}^K \exp(z_i)} = 1$$

Select the largest probability.
The corresponding class is the prediction

Multi-class Classification

- Step 1:

$$\mathbf{z}_i = \begin{bmatrix} z_{i,1} \\ z_{i,2} \\ z_{i,3} \\ \cdots \\ z_{i,K} \end{bmatrix} = \begin{bmatrix} w_{1,0} & w_{1,1} & w_{1,2} & \cdots & w_{1,d} \\ w_{2,0} & w_{2,1} & w_{2,2} & \cdots & w_{2,d} \\ w_{3,0} & w_{3,1} & w_{3,2} & \cdots & w_{3,d} \\ \cdots & & & & \\ w_{K,0} & w_{K,1} & w_{K,2} & \cdots & w_{K,d} \end{bmatrix} \begin{bmatrix} 1 \\ x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ \cdots \\ x_{i,d} \end{bmatrix}$$

- Step 2:

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{z}_i)$$

Note that $\hat{\mathbf{y}}_i$ is a vector, NOT a scalar

Multi-class Classification

- Loss function
  - Label matrix
    - The label of each sample is a one-hot vector

Label vector of 3rd sample

$$Y = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \cdots, \mathbf{y}_n] = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{K \times n}$$

Label vector of 1st sample

Multi-class Classification

- Loss function
  - Likelihood function for the i-th sample

$$Y = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \cdots, \mathbf{y}_n] = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{K\times n}$$

$$p(1|x_i)^{y_{1i}} p(2|x_i)^{y_{2i}} \cdots p(K|x_i)^{y_{Ki}} = \prod_{k=1}^{K} p(k|x_i)^{y_{ki}}$$

  - Maximize the likelihood function for all samples

$$\max_W \prod_{i=1}^{n} \prod_{k=1}^{K} p(k|x_i)^{y_{ki}}$$

$$\min_W -\log \prod_{i=1}^{n} \prod_{k=1}^{K} p(k|x_i)^{y_{ki}}$$

Multi-class Classification

- Loss function

$$\min_W L(W) \triangleq -\log \prod_{i=1}^{n} \prod_{k=1}^{K} p(k|\mathbf{x}_i)^{y_{ki}}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ki} \log p(k|\mathbf{x}_i)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ki} \{\mathbf{w}_k^t \mathbf{x}_i - \log(\sum_{k=1}^{K} \exp(\mathbf{w}_k^t \mathbf{x}_i))\}$$

  - With regularization term

$$\min_W \textstyle\sum_{i=1}^{n} \sum_{k=1}^{K} \left(y_{ki} \log(\sum_{k=1}^{K} \exp(\mathbf{w}_k^t \mathbf{x}_i)) - y_{ki} \mathbf{w}_k^t \mathbf{x}_i \right) + \lambda \|W\|_F^2$$

Multi-class Classification

- Binary classification
  - Model

$$z_i = \mathbf{w}^T \mathbf{x}_i$$

$$\hat{y}_i = \text{sigmoid}(z_i)$$

  - Loss function

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \{\log(1 + \exp(\mathbf{w}^T \mathbf{x}_i)) - y_i \mathbf{w}^T \mathbf{x}_i\}$$

- Multi-class classification
  - Model

$$\mathbf{z}_i = W^T \mathbf{x}_i$$

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{z}_i)$$

  - Loss function

$$\min_W \sum_{i=1}^{n} \sum_{k=1}^{K} \left(y_{ki} \log(\textstyle\sum_{k=1}^{K} \exp(\mathbf{w}_k^t \mathbf{x}_i)) - y_{ki} \mathbf{w}_k^t \mathbf{x}_i \right)$$

Evaluation for multi-class classification

- Binary classification
  - F1 score: harmonic mean of precision and recall

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

- Multi-class classification
  - Micro/Macro recall
  - Micro/Macro precision
  - Micro/Macro f1-score

Evaluation for multi-class classification

- Micro averaging
  - Collect decisions for all classes, compute contingency table and evaluate

Class 1	
10 (TP)	10 (FP)
10 (FN)	970 (TN)

Class 2	
90 (TP)	10 (FP)
10 (FN)	890 (TN)

Class 3	
40 (TP)	10 (FP)
10 (FN)	940 (TN)

Micro	
140 (TP)	30 (FP)
30 (FN)	2800 (TN)

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{micro precision} = \frac{10+90+40}{(10+10)+(90+10)+(40+10)} = 0.82$$

Evaluation for multi-class classification

- Macro averaging:
  - Compute performance for each class, then average

Class 1	
10 (TP)	10 (FP)
10 (FN)	970 (TN)

Class 2	
90 (TP)	10 (FP)
10 (FN)	890 (TN)

Class 3	
40 (TP)	10 (FP)
10 (FN)	940 (TN)

$$\text{precision}_1 = \frac{10}{10+10} = 0.5$$

$$\text{precision}_2 = \frac{90}{90+10} = 0.9$$

$$\text{precision}_3 = \frac{40}{40+10} = 0.8$$

$$\text{macro precision} = \frac{\text{precision}_1 + \text{precision}_2 + \text{precision}_3}{3} = 0.73$$

Evaluation for multi-class classification

- Exercise: micro-recall, macro-recall $\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$

Class 1	
10 (TP)	10 (FP)
10 (FN)	970 (TN)

Class 2	
90 (TP)	10 (FP)
10 (FN)	890 (TN)

Class 3	
40 (TP)	10 (FP)
10 (FN)	940 (TN)

Micro	
140 (TP)	30 (FP)
30 (FN)	2800 (TN)

Evaluation for multi-class classification

- Micro-f1

Class 1	
10 (TP)	10 (FP)
10 (FN)	970 (TN)

Class 2	
90 (TP)	10 (FP)
10 (FN)	890 (TN)

Class 3	
40 (TP)	10 (FP)
10 (FN)	940 (TN)

Micro	
140 (TP)	30 (FP)
30 (FN)	2800 (TN)

$$\text{micro precision} = \frac{10+90+40}{(10+90+40)+(10+10+10)}$$

$$\text{micro recall} = \frac{10+90+40}{(10+90+40)+(10+10+10)}$$

$$\text{micro-f1} = \frac{2 \times \text{micro precision} \times \text{micro recall}}{\text{micro recall} + \text{micro precision}}$$

Review for multiclass classification

- Macro-f1

Class 1	
10 (TP)	10 (FP)
10 (FN)	970 (TN)

Class 2	
90 (TP)	10 (FP)
10 (FN)	890 (TN)

Class 3	
40 (TP)	10 (FP)
10 (FN)	940 (TN)

$$\text{precision}_1 = \frac{10}{10+10}$$

$$\text{precision}_2 = \frac{90}{90+10}$$

$$\text{precision}_3 = \frac{40}{40+10}$$

$$\text{recall}_1 = \frac{10}{10+10}$$

$$\text{recall}_2 = \frac{90}{90+10}$$

$$\text{recall}_3 = \frac{40}{40+10}$$

$$\text{macro precision} = \frac{\text{precision}_1+\text{precision}_2+\text{precision}_3}{3}$$

$$\text{macro recall} = \frac{\text{recall}_1+\text{recall}_2+\text{recall}_3}{3}$$

$$\text{macro-f1} = \frac{\text{f1}_1+\text{f1}_2+\text{f1}_3}{3}$$

Evaluation of multi-class classification

- Imbalance between classes
  - One class has many more samples than other classes
  - Use micro averaging

Class 1	
1 (TP)	1 (FP)
7 (FN)	1 (TN)

Class 2	
10 (TP)	90 (FP)
890 (FN)	10 (TN)

Class 3	
1 (TP)	1 (FP)
7 (FN)	1 (TN)

$$\text{precision}_1 = \frac{1}{1+1} = 0.5$$

$$\text{precision}_2 = \frac{10}{90+10} = 0.1$$

$$\text{precision}_3 = \frac{1}{1+1} = 0.5$$

$$\text{macro precision} = \frac{0.5+0.1+0.5}{3} = 0.36$$

$$\text{micro precision} = \frac{1+10+1}{(1+10+1)+(1+90+1)} = 0.11$$