Classification (part 2)

Multiclass logistic regression

Recap: logistic function

Represent $p(\mathbf{x}) = P(Y = 1 \,|\, \mathbf{X} = \mathbf{x})$ by sigmoid function

$$\sigma(t) = \frac{1}{1 + e^{-t}} .$$

where t is an arbitrary function of $\mathbf{x}$ (e.g. the output of a deep neural network)

Recap: cross-entropy loss

The cross-entropy loss is a popular loss function for binary classification:

$$L\left(y, \hat{p}(\mathbf{x})\right) = \underbrace{\boxed{- y \cdot \log\left(\hat{p}\right)}}_{\text{when } y=1} \underbrace{\boxed{- (1-y) \cdot \log\left(1-\hat{p}\right)}}_{\text{when } y=0} \tag{1}$$

Recap: decision cutoff

Let $\hat{p}(\mathbf{x})$ be an estimator of $p(\mathbf{x})$, we can classify $\mathbf{X}$ by the following classification rule:

$$h(\mathbf{X}) = \begin{cases} 1, & \text{if } \hat{p}(\mathbf{X}) \geq \delta \\ 0, & \text{if } \hat{p}(\mathbf{X}) < \delta \end{cases}$$

- The hyperparameter $\delta \in (0,1)$ is called a classification threshold or a decision cutoff

- δ is often selected by optimizing an appropriate metric

Recap: metrics for a binary classifier

- Confusion matrix
- Accuracy
- Sensitivity (recall)
- Precision
- Specificity
- ROC curve and AUC

Agenda

- Multiclass logistic regression
  - sigmoid function $\longrightarrow$ softmax functions
  - cross-enropy loss $\longrightarrow$ categorical cross-entropy loss
- Evaluating multiclass classifiers

Multiclass logistic regression

Categorical model

Recall that in a multiclass classification problem $y \in \{1,2,\ldots,K\}$ for $K \geq 2$. Here our idea is to model $p_k(\mathbf{x}) = P(Y = k \mid \mathbf{X} = \mathbf{x})$ and predict the value of Y by the values of $p_k(\mathbf{x})$.

1. Since $p_k(\mathbf{x})$ is a conditional probability, we need $0 \leq p_k(\mathbf{x}) \leq 1$.

2. We also need $\sum_{k=1}^{K} p_k(\mathbf{x}) = 1$.

Softmax functions

We may satisfy the above two constraints by modeling $p_k(\mathbf{x})$ as softmax functions:

$$p_k(\mathbf{x}) = \frac{\exp\left[m_k(\mathbf{x})\right]}{\sum_{j=1}^{K} \exp\left[m_j(\mathbf{x})\right]} \tag{2}$$

where $m_j(\mathbf{x})$ is an arbitrary function of $\mathbf{x}$ (e.g. the output of a neural network), often denoted as $m_j\left(\mathbf{x}; \boldsymbol{\theta}_j\right)$.

Implementation details

Exponential functions grow very fast, and thus $e^{m_j(\mathbf{x})}$ may easily overflow. Fortunately, we have

$$\frac{\exp\left[m_k(\mathbf{x})\right]}{\sum_{j=1}^{K}\exp\left[m_j(\mathbf{x})\right]} = \frac{C\cdot\exp\left[m_k(\mathbf{x})\right]}{C\cdot\sum_{j=1}^{K}\exp\left[m_j(\mathbf{x})\right]}$$

$$= \frac{\exp\left[m_k(\mathbf{x})+\tilde{C}\right]}{\sum_{j=1}^{K}\exp\left[m_j(\mathbf{x})+\tilde{C}\right]}$$

Thus we can avoid numerical overflow by using $\tilde{C} = -\max\left\{m_j(\mathbf{x})\right\}$.

One–hot encoding

Let $\tilde{\mathbf{y}} = \left[y^{(1)}, y^{(2)}, \ldots, y^{(K)}\right]^\top$ be the one–hot encoding of $y \in \{1, 2, \cdots, K\}$, where

$$y^{(k)} \triangleq I(y = k) = \begin{cases} 1, & \text{if } y = k \\ 0, & \text{otherwise} \end{cases}$$

只有一個1、其他$k - 1$個都是0

Categorical cross-entropy loss

The categorical cross-entropy loss for multiclass classification is defined as

$$L\left(\tilde{\mathbf{y}}, \hat{\mathbf{p}}\right) = -\sum_{k=1}^{K} \tilde{y}^{(k)} \log \hat{p}_k, \tag{3}$$

where $\hat{\mathbf{p}} = \left[\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K\right]^\top$.

When $K = 2$

When $K = 2$, equation (3) can be written as

$$L\left(y, \hat{\mathbf{p}}\right) = -I(y = 1) \cdot \hat{p}_1 - I(\cancel{y = 2}) \cdot \hat{p}_2$$

$$= -I(y = 1) \cdot \hat{p}_1 - I(y = 0) \cdot \left(1 - \hat{p}_1\right)$$

$$= -y \cdot \hat{p} - (1 - y) \cdot \left(1 - \hat{p}\right)$$

$\hat{p}_1 + \hat{p}_2 = 1$

Multiclass logistic regression

By combining equations (2) and (3), multiclass logistic regression estimates $p_k(\mathbf{x})$ for $k = 1,2,\ldots,K$ by

$$\min_{\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_K} - \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{y}_i^{(k)} \log \hat{p}_k(\mathbf{x}_i)$$

where $\hat{p}_k(\mathbf{x}_i)$'s are defined in equation (2).

Multiclass logistic regression by MLE

- Assume $Y_i \,|\, \mathbf{X}_i = \mathbf{x}_i$ is independently sampled from categorical distribution with parameters $\mathbf{p}(\mathbf{x}_i)$.

- The likelihood function becomes

$$\prod_{i=1}^{n} \left[\prod_{k=1}^{K} p_k(\mathbf{x}_i) \right]^{y_i^{(k)}}$$

- The log-likelihood becomes

$$\ell = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} y_i^{(k)} \log p_k(\mathbf{x}_i) \right]$$

Classification rule

In multiclass logistic regression, we often predict Y by

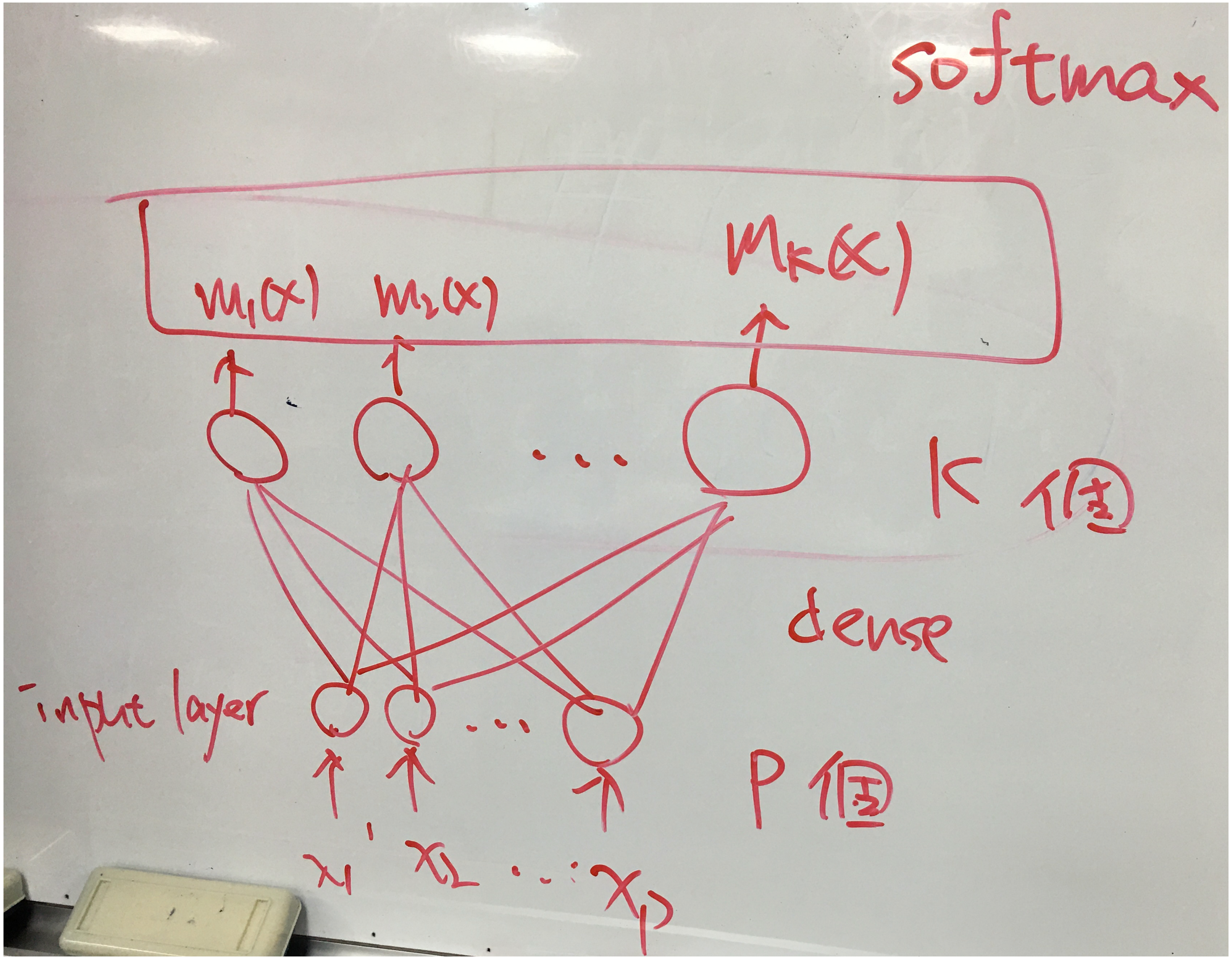
$$\hat{Y} = \underset{k=1,2,\ldots,K}{\arg\max}\, \hat{P}(Y = k \mid \mathbf{X} = \mathbf{x})$$
$$= \underset{k=1,2,\ldots,K}{\arg\max}\, \hat{p}_k(\mathbf{x})$$

Multiclass logistic regression in TensorFlow

- When the response variables are already one-hot encoded, specify the loss function as tf.keras.losses.CategoricalCrossentropy

- When the response variables are numeric or categorical (e.g. strings), specify the loss function as tf.keras.losses.SparseCategoricalCrossentropy

Example: linear multiclass logistic regression



L196

```
model = tf.keras.Sequential(name='iris')
model.add(layers.Dense(nc, activation='softmax', input_shape=(p,)))
```

```
model.compile(optimizer='sgd',
              loss=tf.keras.losses.SparseCategoricalCrossentropy())
```

```
y_pred = np.argmax(model.predict(X_test), axis=-1)
```

model.predict會輸出類神經網路輸出層的輸出 (e.g. $\hat{p}_k$)

Evaluating multiclass classifiers

Confusion matrix

		True/Actual		
		Cat ( )	Fish ( )	Hen ( )
Predicted	Cat ( )	4	6	3
	Fish ( )	1	2	0
	Hen ( )	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Accuracy

$$\text{Accuracy} = \frac{\text{trace of confusion matrix}}{\text{sample size}}$$

Precisions and recalls

Precisions and recalls are defined class-by-class:

		True/Actual		
		Cat (🐱)	Fish (🐠)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐠)	1	2	0
	Hen (🐔)	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

		True/Actual		
		Cat (🐱)	Fish (🐠)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐠)	1	2	0
	Hen (🐔)	1	2	6

		Cat (🐱)	Fish (🐠)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐠)	1	2	0
	Hen (🐔)	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Combining per-class scores

- macro-score = arithmetic mean of the per-class scores
- weighted-score = weighted average (by number of samples from that class) of the per-class scores
- micro-score: let micro-TP = sum of per-class TPs, ...

Summary

- We extend logistic regression to multiclass classification problems
  - sigmoid function $\longrightarrow$ softmax functions
  - cross-enropy loss $\longrightarrow$ categorical cross-entropy loss
  - decision rule

- Metrics for multiclass classifiers:
    - accuracy
    - per–class precisions and recalls
    - micro, macro, and weighted scores

Readings

- Video from 機器學習基石
- Metrics for multiclass classification