Lab6: Logistic Regression and Metrics

DataLab

2023.10.19
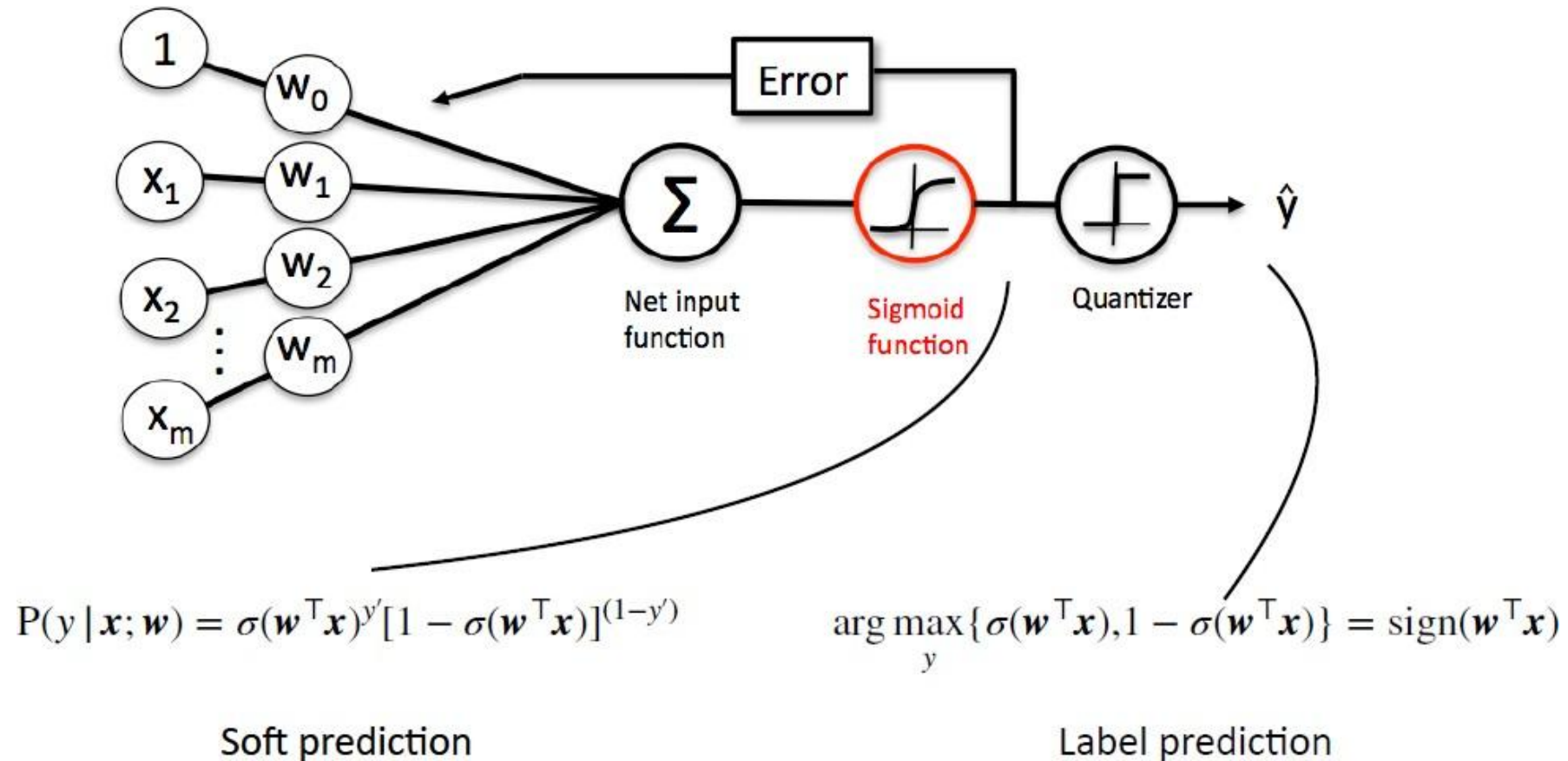
Outline

- Brief Review: Logistic Regression

- Common Evaluation Metrics for Binary Classification

  - Confusion Matrix

  - Soft Classifiers – ROC Curve

Outline

- Brief Review: Logistic Regression

- Common Evaluation Metrics for Binary Classification
  - Confusion Matrix
  - Soft Classifiers – ROC Curve

Logistic Regression

$$P(y \mid \boldsymbol{x}; \boldsymbol{w}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x})^{y'}[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{(1-y')}$$

$$\arg\max_{y}\{\sigma(\boldsymbol{w}^\top \boldsymbol{x}), 1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})\} = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x})$$

Soft prediction

Label prediction

Logistic Regression + Regularization

- 把 regularization term 加到 loss 內，讓模型在學 weights 的時候，會傾向選擇比較簡單的模型。

$$\arg\max_{w} \log P(\mathbb{X} \mid w) - \frac{\alpha}{2}\|w\|^2$$

Outline

- Brief Review: Logistic Regression

- Common Evaluation Metrics for Binary Classification

   - Confusion Matrix

   - Soft Classifiers – ROC Curve

Confusion Matrix

- It is important to know how the model make wrong prediction

- In **binary classification**, confusion matrix is a common tool to analyze the predictions
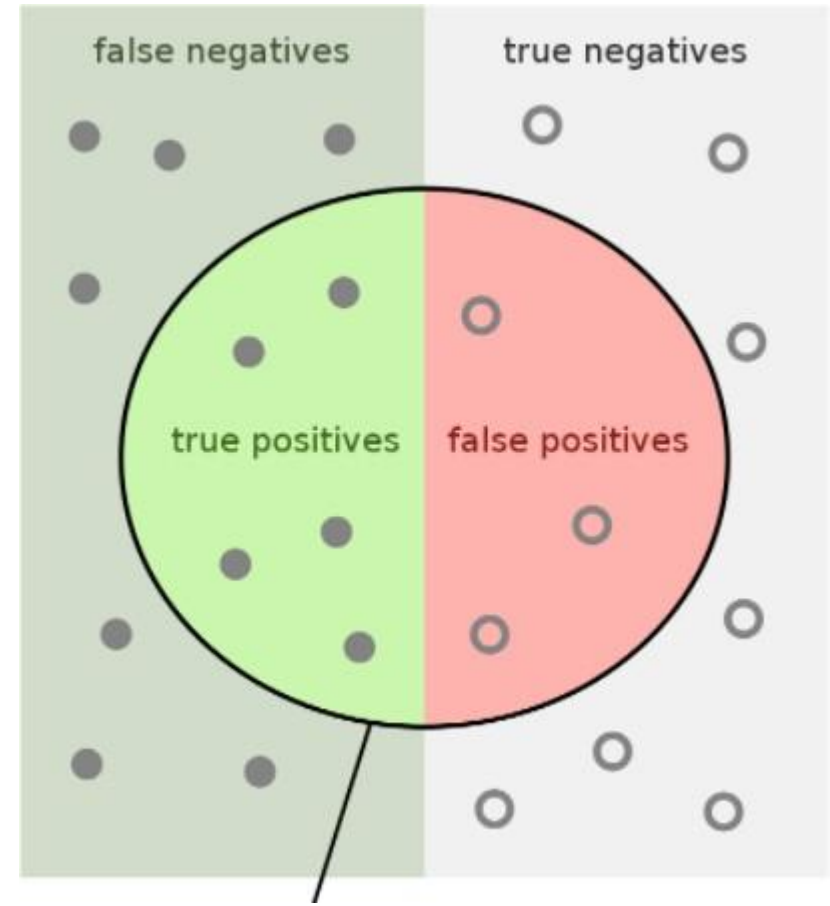
Confusion Matrix

- It is important to know how the model make wrong prediction.

- e.g. 檢測絕症
  - 寧可讓多一點人到 TP & FP，也不要讓 FN 很高
  - i.e. TPR higher

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

Predicted class

		P'	N'
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Confusion Matrix

- Precision(PRE) & Recall Rate(REC)

$$PRE = \frac{TP}{TP + FP}, \quad \text{(the higher, the better)}$$

$$REC = \frac{TP}{TP + FN} = TPR. \quad \text{(the higher, the better)}$$

- F-1 Score

$$F_1 = 2\frac{(PRE * REC)}{PRE + REC}, \quad \text{(the higher, the better)}$$

Predicted class

		P'	N'
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Confusion Matrix

- Precision(PRE) & Recall Rate(REC)

$$PRE = \frac{TP}{TP + FP}, \quad \text{(the higher, the better)}$$

$$REC = \frac{TP}{TP + FN} = TPR. \quad \text{(the higher, the better)}$$

- F-1 Score

$$F_1 = 2\frac{(PRE * REC)}{PRE + REC}, \quad \text{(the higher, the better)}$$

Outline

- Brief Review: Logistic Regression

- Common Evaluation Metrics for Binary Classification
  - Confusion Matrix
  - Soft Classifiers – ROC Curve

ROC Curve

- ROC curve analyze the performance for **every threshold in soft classifiers**

- In X-axis $\quad FPR = \dfrac{FP}{FP + TN}$
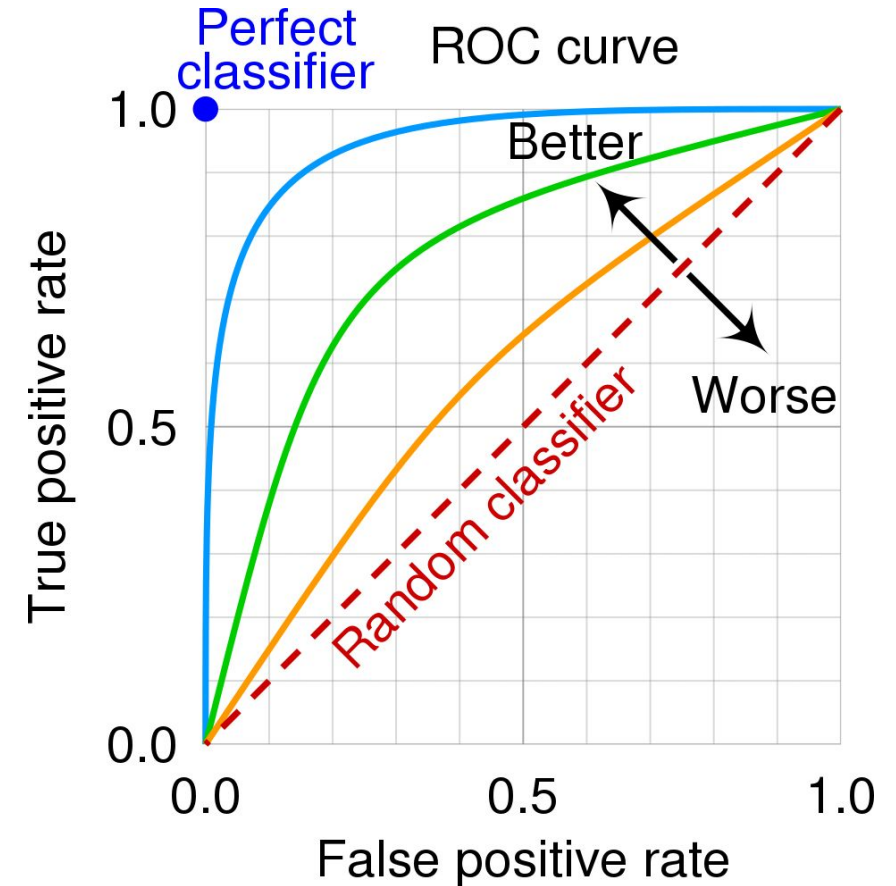
- In Y-axis $\quad TPR = \dfrac{TP}{TP + FN}$

$$
\begin{array}{c|c}
1 & \\
1 & \\
0.87 & \theta \\
\hline
0.64 & \Downarrow \\
\vdots & \\
-0.88 & \\
-0.93 & \\
-1 &
\end{array}
$$

ROC Curve

- ROC curve analyze the performance for **every threshold in soft classifiers**
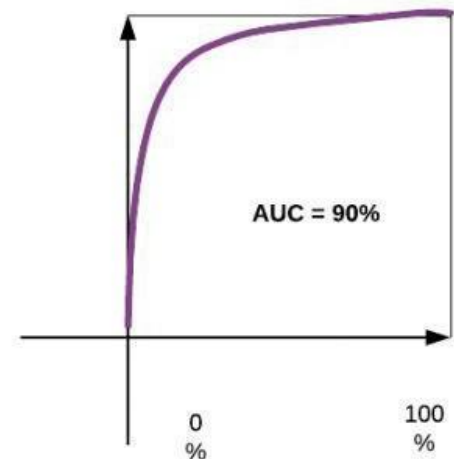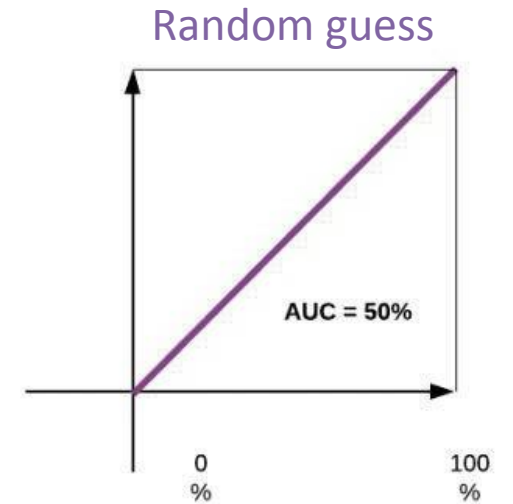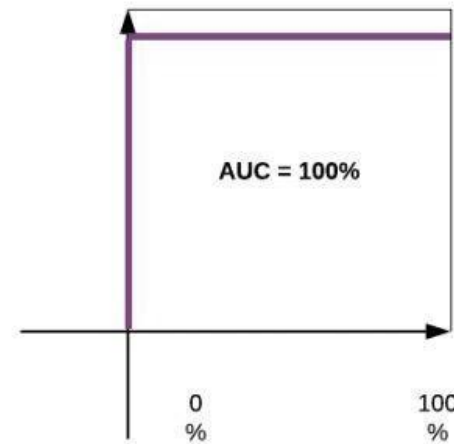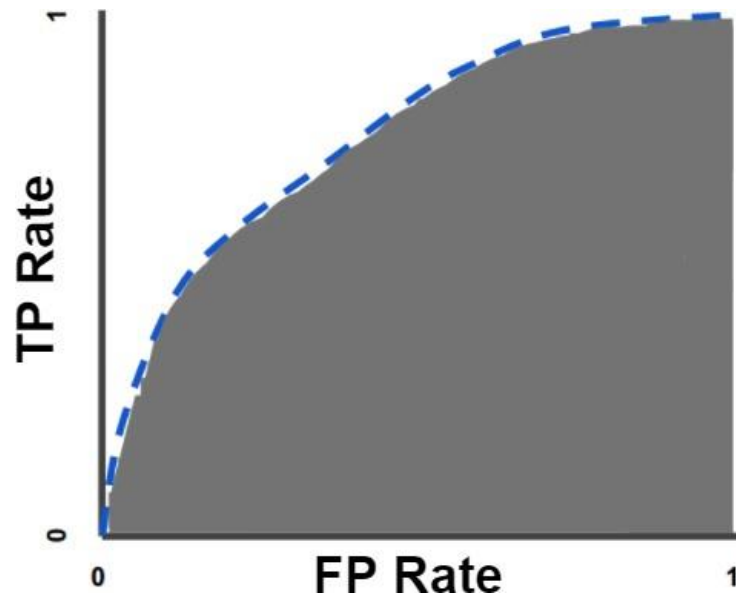
- In X-axis $FPR = \dfrac{FP}{FP + TN}$

- In Y-axis $TPR = \dfrac{TP}{TP + FN}$

AUC

- AUC (Area Under the ROC Curve)
  - ROC can be quantified using AUC

https://developers.google.com/machine-learning/crash- course/classification/roc-and-auc
https://medium.com/acing-ai/what-is-auc-446a71810df9
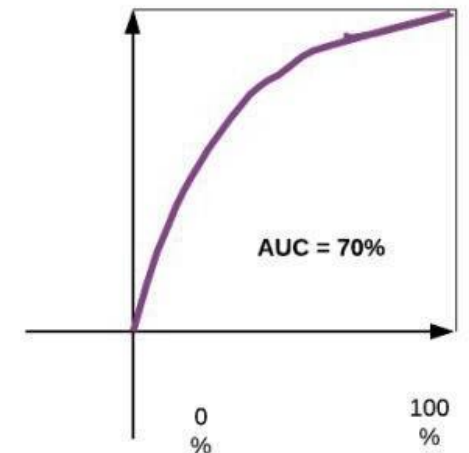
Homework

- Homework: Lab06
  - Lab06: Logistic Regression, Metrics

Reference

- https://bookdown.org/ccwang/medical_statistics6/section-43.html
- https://bookdown.org/ccwang/medical_statistics6/bernoulli.html
- https://bookdown.org/ccwang/medical_statistics6/binomial.html
- https://bookdown.org/ccwang/medical_statistics6/likelihood-definition.html
- https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- https://commons.wikimedia.org/w/index.php?curid=109730045
- https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- https://medium.com/acing-ai/what-is-auc-446a71810df9
- https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation