MM 225 – AI and Data Science

Day 28: Logistic Regression 2

Instructors: Hina Gokhale, MP Gururajan, N. Vishwanathan

10 OCTOBER 2024

Good news!

LSE function for multiple regression is a convex function:

$$\boldsymbol{SSE} = \sum \left(y^{(i)} - \boldsymbol{w} \cdot \boldsymbol{x}^{(\boldsymbol{i})} \right)^2$$

Cross entropy error function is also convex

$$\mathcal{E}(\mathbf{w}) = -\sum y^{(i)} \log\left(\psi\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(\boldsymbol{i})} \right) \right) + \left(1 - y^{(i)} \right) log\left(1 - \psi\left(\boldsymbol{w} \cdot \boldsymbol{x}^{(\boldsymbol{i})} \right) \right)$$

$$\psi(z) = \frac{1}{1 + e^{-z}}$$

Regression

$$\boldsymbol{SSE} = \sum \left(y^{(i)} - \boldsymbol{w} \cdot \boldsymbol{x}^{(i)}\right)^2 = \left[\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} \boldsymbol{w} - 2\left(\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y}\right)^{\boldsymbol{T}} \boldsymbol{w}\right] + const$$

Initialize with $\boldsymbol{w}_1$ at random

set t = 1 and choose λ

continue until convergence

1. compute the gradient $\nabla \boldsymbol{SSE} = \boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{X}\boldsymbol{w}_{\boldsymbol{t}} - \boldsymbol{y})$
2. update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \lambda \nabla \boldsymbol{SSE}$
3. $t \leftarrow t + 1$

Logistic Regression

$$\boldsymbol{\mathcal{E}}(\mathbf{w}) = -\sum y^{(i)} \log\left(\psi\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(\boldsymbol{i})}\right)\right) + \left(1 - y^{(i)}\right) log\left(1 - \psi\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(\boldsymbol{i})}\right)\right)$$

$\psi(z) = \frac{1}{1+e^{-z}}$ then $\frac{d\psi(z)}{dz} = \psi(z)(1-\psi(z))$

It can be shown that

$$\frac{\partial\boldsymbol{\mathcal{E}}(\mathbf{w})}{\partial\boldsymbol{w}} = \sum_i \left[\psi\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(\boldsymbol{i})}\right) - y^{(i)}\right]\boldsymbol{x}^{(\boldsymbol{i})}$$

Algorithm for Logistic Regression

1. Choose λ
2. Initiate $\boldsymbol{w}_1$
3. Iterate as $\boldsymbol{w}_{\boldsymbol{t+1}} = \boldsymbol{w}_t - \lambda \sum_i \left[\psi\left(\boldsymbol{w}_t \cdot \boldsymbol{x}^{(\boldsymbol{i})}\right) - y^{(i)}\right]\boldsymbol{x}^{(\boldsymbol{i})}$
4. Iterate until convergence

Choosing λ in practice

Try out $\lambda = 0.001, 0.01, 0.1$ on a test data set. Choose the λ that gives stable and fast convergence.

To reach true minimum, reduce λ by factor of 10 as learning saturates.

Validation

Validation with test data

Model estimation using training data

Apply the same model to test data and see if the results are “same”!

What is meant by “same”?

Misclassification Quantification

Total data size = P + N		Predicted Condition	
		Positive(PP)	Negative (PN)
Actual Condition	Positive (P)	**True Positive (TP)**	**False Negative (FN)**
	Negative (N)	**False Positive (FP)**	**True Negative (TN)**

$$\text{TPR = True Positive Rate} = \frac{\mathbf{TP}}{\mathbf{TP+FN}}$$ = Sensitivity / Hit Rate / Recall

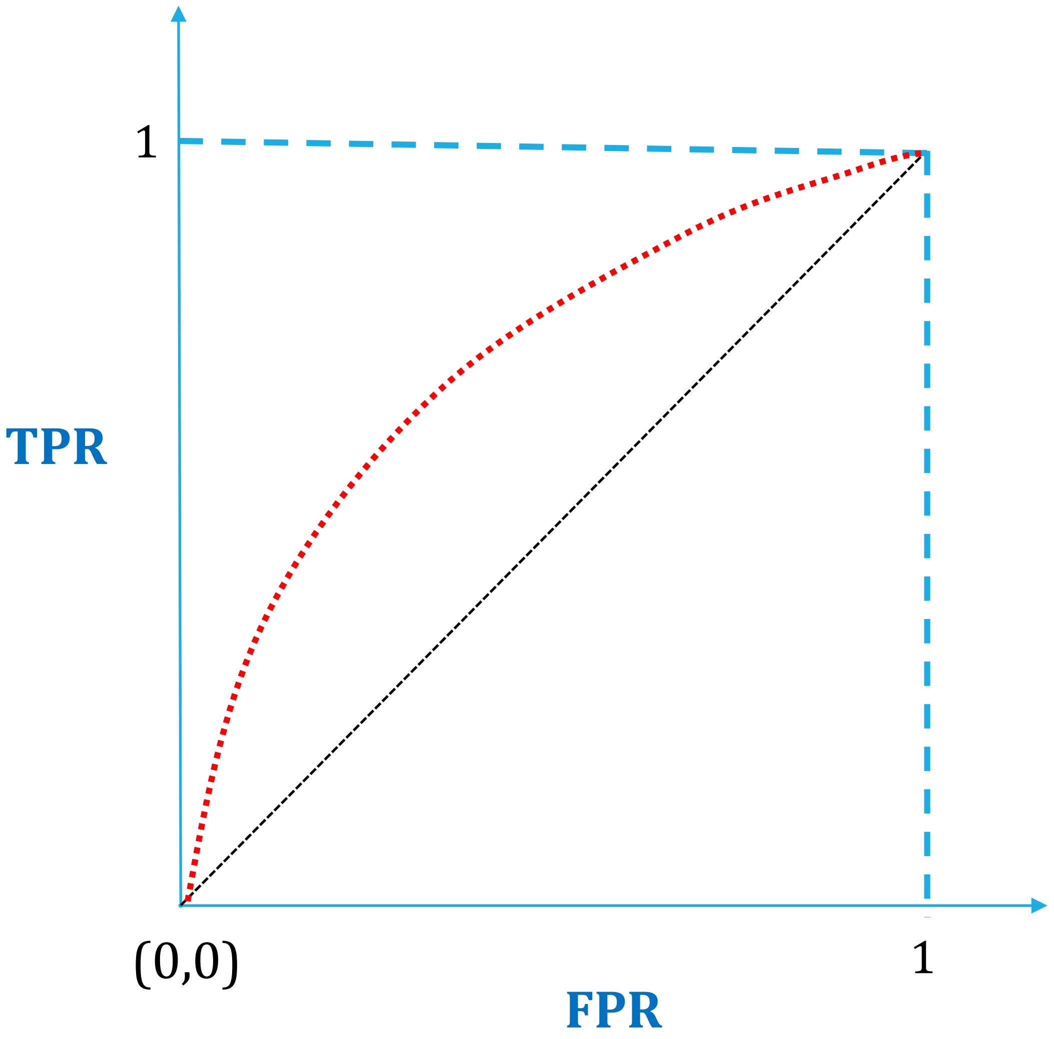
$$\text{FPR = False Positive Rate} = \frac{\mathbf{FP}}{\mathbf{FP+TN}}$$ = Probability of false alarm = 1- specificity

ROC Curve & AUC

ROC Curve = Receiving Operating Characteristic Curve

ROC is a plot of FPR vs. TPR calculated at different threshold values.

Area Under the ROC Curve is called AUC



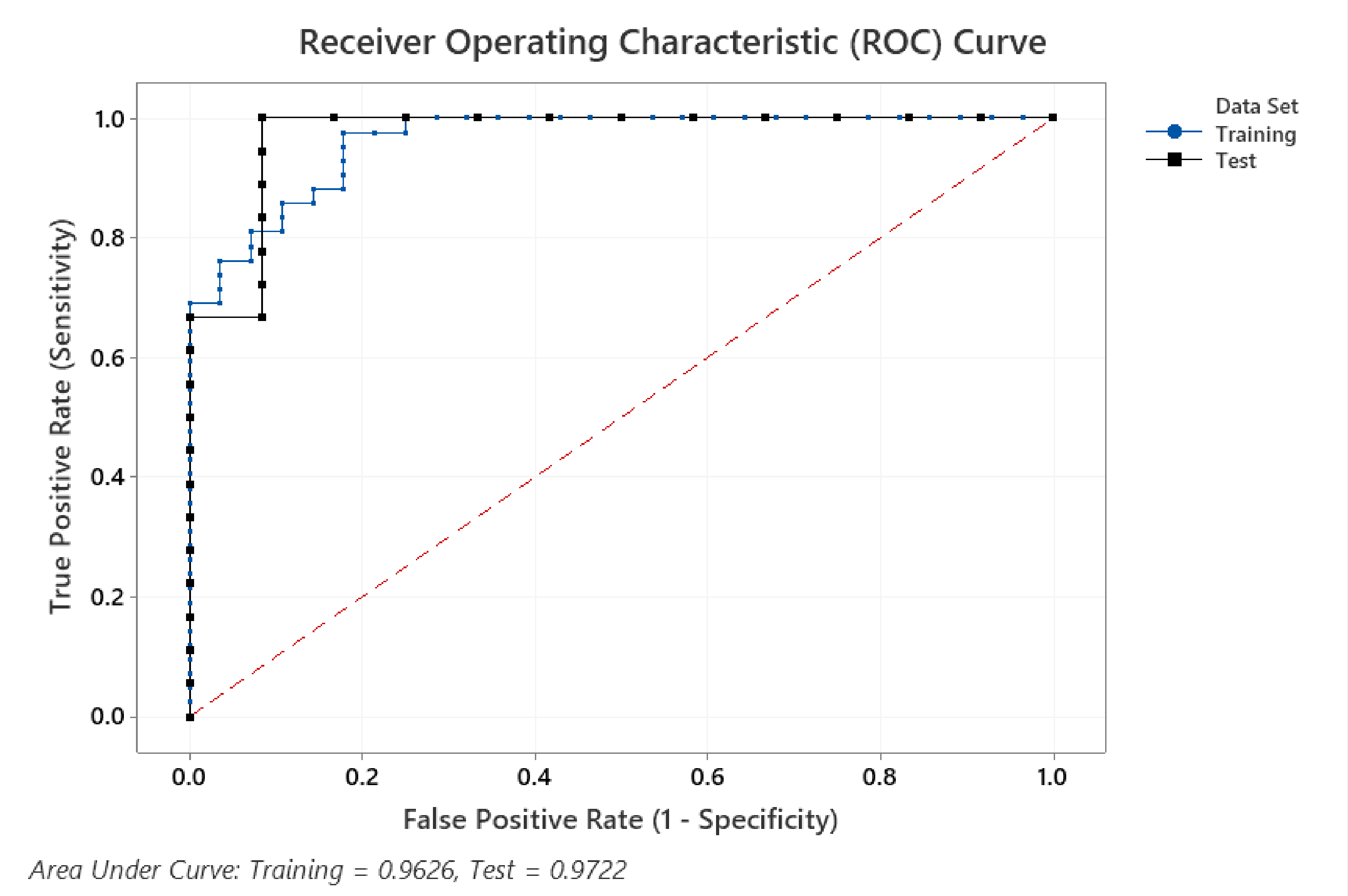
What is a good fit?

ROC is above the FPR = TPR line

AOC is more than 50%

ROC and AOC for Test data is close to that of Training data.

Large data set - Validation



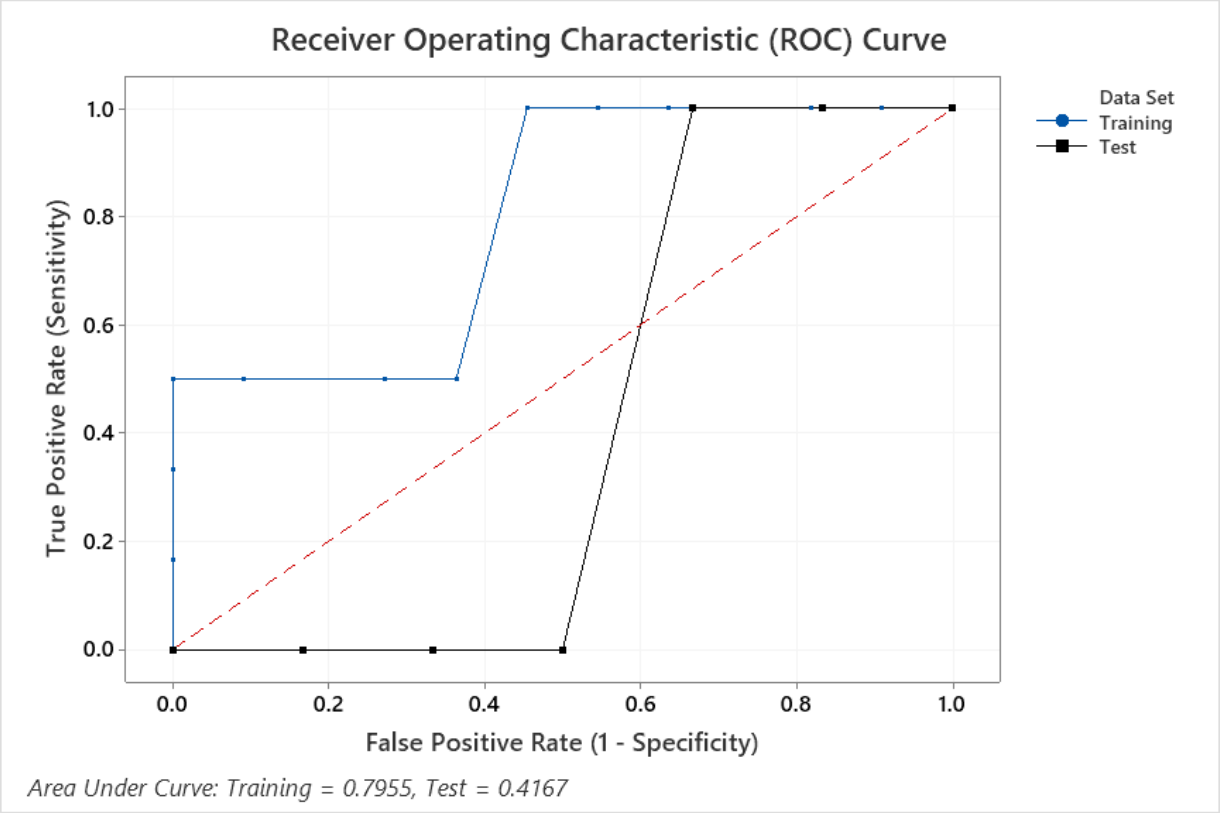
Validation

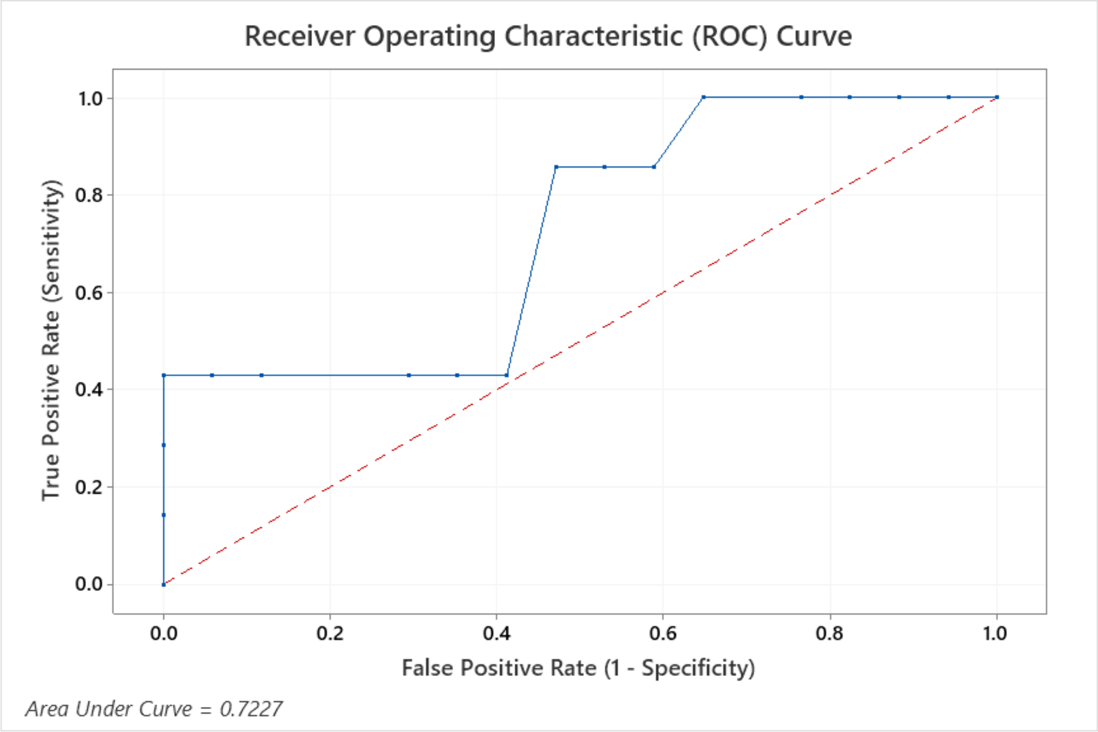
If ROC and AUC for Training Data and Test data are close then the model is validated.

If not valid:

- Too many features? Reduce the number of features.
- Correlated features? Keep only uncorrelated features.
- Training and test data sets chosen randomly? If not, correct it.
- Is data too small? Then avoid dividing the data. Keep only training data for estimation and make sure that the model fits well.

Small data – Caution!





Summary

Algorithm for Logistic Regression and choice of λ.

Validation of the model : Confusion Matrix

- ROC – Receiving Operating Characteristic Curve
- AUC – Area under the ROC Curve

Thank you…