PLN para a Ciência Política e Políticas Públicas Públicas

Professora: Lorena Barberia

Semana 9

Universidade de São Paulo

Tópicos da Aula

Universidade de São Paulo

Seleção de um Modelo Probabilístico Paramétrico de Aprendizado para Classificação

Voltando a nosso projeto, vamos utilizar hoje uma variável binaria *dummy* para identificar conteúdo (e.g., desfavorável)

- $y = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ valor anotado no *training data set*
- $\hat{y}$ = valor predito na amostra

Universidade de São Paulo

Classificação Supervisionada

- Lembrando das 03 etapas necessárias:
  1. Classificação dos documentos (Definição e Categorização);
  2. Treinamento do modelo (Aprendizado do Modelo);
  3. Validação (Inferência).

Universidade de São Paulo

Resumo

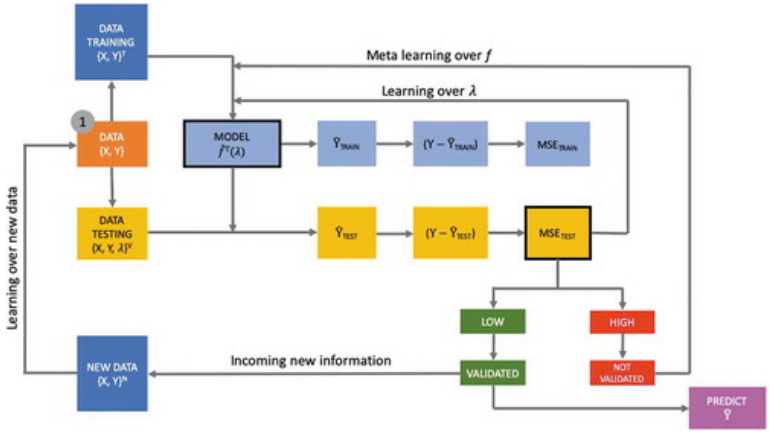


Figure: Cerulli (2023)

Universidade de São Paulo

Resultados, Preditores e o Erro

$$Y = f(X) + \varepsilon$$

- *Dependent Variable*: A variável ou desfecho que queremos **prever**
- f = Modelo estatístico (*Learner*)
- *Predictors=Features*: Variáveis explicativas
- *Error*: Variação Estocástica

Universidade de São Paulo

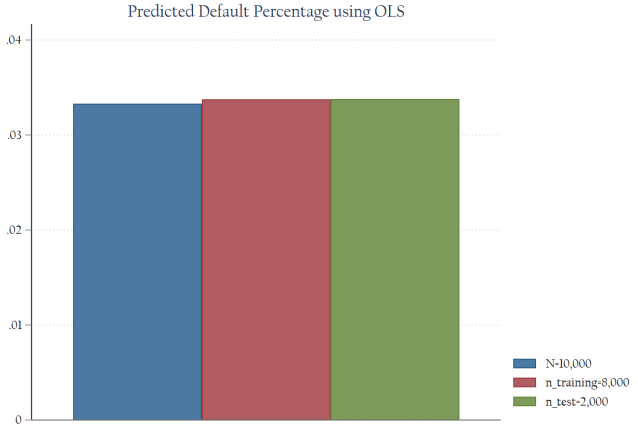
Conceitos básicos: Training e Test Data Sets *Y*: Exemplo Default

Default	Training Data Set	Test Data Set	Total
No	7,739	1,928	9,667
	80.06	19.94	100
	96.74	96.40	96.67
Yes	261	72	333
	78.38	21.62	100
	3.26	3.60	3.33
Total	8,000	2,000	10,000

Universidade de São Paulo

Conceitos básicos: Valores Preditos de *Y*: Exemplo Default

Learner Model: Default= *f*(Student, Balance, Income)



Universidade de São Paulo

Conceitos básicos: Modelo de Aprendizado de OLS *Y*: Exemplo Default

Table: OLS Regression table

	(1) default_all	(2) default_training
income	0.000000199	4.70e-08
	(0.000000192)	(0.000000214)
balance	0.000133	0.000134
	(0.00000355)	(0.00000398)
student_dummy	-0.0103	-0.0136
	(0.00566)	(0.00635)
Constant	-0.0708	-0.0619
	(0.0131)	(0.0146)
N	10000	8000
r2	0.124	0.125
rmse	0.168	0.169

Standard errors in parentheses

Conceitos básicos: Mean Square Error (MSE)

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\
&= [f(X) - \hat{f}(X)]^2 + Var(\varepsilon)
\end{aligned}
\tag{1}
$$

Universidade de São Paulo

Reducible and Irreducible Error

$$= E(Y - \hat{Y})^2 = E[(f(x) + \epsilon - \hat{f}(X)]^2$$
$$= E[(f(x) - \hat{f}(X)] + Var(\epsilon)$$
= Reducible error + Irreducible error

A precisão (*accuracy*) de $\hat{Y}$ como uma previsão para Y depende de ambos tipos de erro, e em aprendizado de maquina procuramos reduzir *reducible error*.

This error is reducible because we can potentially improve the accuracy of $\hat{f}$ *by using the most appropriate statistical learning technique to estimate f.*

Universidade de São Paulo

MSE *Y* do Modelo Linear: Exemplo Default

Default	Training Data Set	Test Data Set	Entire Sample
Predicted % of Default	3.375	3.390	3.333
Mean Squared Error	0.0285	0.0268	0.0282
N	8,000	2,000	10,000

Universidade de São Paulo

Erro

- *Training error rate*: Erro que obtemos ao aplicar o modelo (aprendizado) aos dados da amostra
- *Test error rate*: Erro de usar o modelo para novos dados

Universidade de São Paulo

Default regression with train/test predictions

Learner Model: Default= f($Student, $Balance, $Income)

Learner: Least Squares regression

Dataset information

Target variable = "default_dummy"	Number of features = 3
N. of training units = 8000	N. of testing units = 2000
N. of used training units = 8000	N. of used testing units = 2000

Validation results

MSE = mean squared error	MAPE = mean absolute percentage error
Training MSE = .02854126	Testing MSE = .02684024
Training MAPE % = 84.5667	Testing MAPE % = 84.45597

Universidade de São Paulo

Default regression with train/test predictions

Learner Model: Default$= f(\$Student, Balance)$

Learner: Least Squares regression

Dataset information

Target variable = "default_dummy"	Number of features = 2
N. of training units = 10000	N. of testing units = 2000
N. of used training units = 10000	N. of used testing units = 2000

Validation results

MSE = mean squared error	MAPE = mean absolute percentage error
Training MSE = .02820106	Testing MSE = .02684149
Training MAPE % = 84.545751	Testing MAPE % = 84.609244

Universidade de São Paulo

5-fold Cross-validation

- Um método para estimar o MSE de teste usando os dados de treinamento.
- Cada fase, 4 sub-amostras de treinamento e 1 sub-amostra para validação.
- Cinco fases.

Train	Train	Train	Train	Validate

Default regression with train/test predictions and cross-validation

Learner Model: Default= f($Student, $Balance), 5-fold

Learner: Least Squares regression

Dataset information

Target variable = "default_dummy"	Number of features = 2
N. of training units = 12000	N. of testing units = 2000
N. of used training units = 12000	N. of used testing units = 2000

Validation results

MSE = mean squared error	MAPE = mean absolute percentage error
Training MSE = .02797446	Testing MSE = .0268391
Training MAPE % = 84.555853	Testing MAPE % = 84.704467

Universidade de São Paulo

Regressão Linear e Logística

- Regressão Linear: $P(Default) = \beta_0 + \beta_1 Income + \beta_2 Student$
- Modelo Logit: $P(Default) = \frac{e^{\beta_0 + \beta_1 Income + \beta_2 Student}}{1 + e^{\beta_0 + \beta_1 Income + \beta_2 Student}}$

Universidade de São Paulo

Regressão Linear e Logística

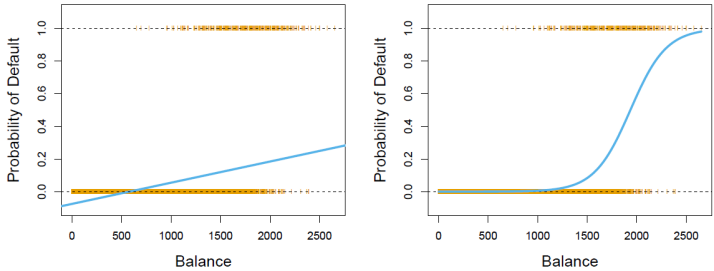


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Universidade de São Paulo

Regressão Logística

Remember the **Odds Ratio (OR)** is a ratio, so it takes a value between 0 and $+\infty$ (it cannot be negative).

- **OR>1.0** represents a **positive logistic association;**
- **OR<1.0** represents a **negative logistic association;**
- **OR=1.0** represents a **lack of association.**

Universidade de São Paulo

Regressão Logística

Table: Log Odds-Ratio Regression table

	(1) default_dummy
income	1.000 (0.37)
balance	1.006*** (24.74)
student_dummy	0.524** (-2.74)
N	10000

Exponentiated coefficients; *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Universidade de São Paulo

Regressão Logística: Classsification Results

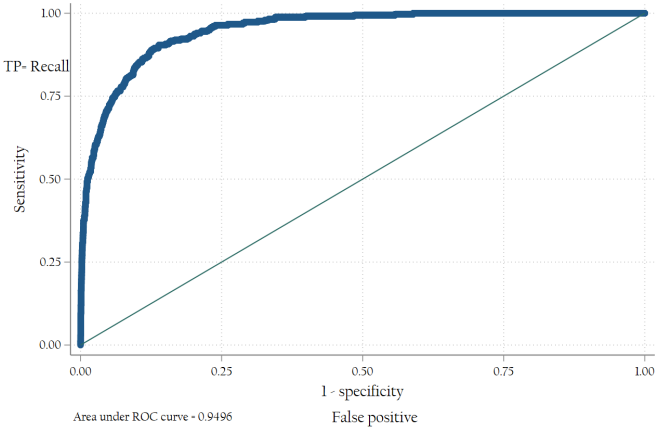
Logistic model for default_dummy

	True		
Classified	D	~D	Total
+	**105**	**40**	**145**
-	**228**	**9627**	**9855**
Total	**333**	**9667**	**10000**

Classified + if predicted Pr(D) >= .5
True D defined as default_dummy != 0

Sensitivity	Pr(+| D)	**31.53%**
Specificity	Pr(-|~D)	**99.59%**
Positive predictive value	Pr(D| +)	**72.41%**
Negative predictive value	Pr(~D| -)	**97.69%**
False + rate for true ~D	Pr(+|~D)	**0.41%**
False - rate for true D	Pr(-| D)	**68.47%**
False + rate for classified +	Pr(~D| +)	**27.59%**
False - rate for classified -	Pr(D| -)	**2.31%**
Correctly classified		**97.32%**

Regressão Logística:ROC



Universidade de São Paulo

Regressão Logística

Table: Logit Odds-Ratio Regression table

	(1) default	(2) default$_{train}$
income	1.000	1.000
	(0.37)	(-0.34)
balance	1.006***	1.006***
	(24.74)	(22.14)
student_dummy	0.524**	0.458**
	(-2.74)	(-3.03)
N	10000	8000

Exponentiated coefficients; *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Universidade de São Paulo

Regressão Logística: Classsification Results for Train

Logistic model for default_dummy

	True		
Classified	D	~D	Total
+	**80**	**34**	**114**
-	**190**	**7696**	**7886**
Total	**270**	**7730**	**8000**

Classified + if predicted Pr(D) >= .5
True D defined as default_dummy != 0

Sensitivity	Pr(+| D)	**29.63%**
Specificity	Pr(-|~D)	**99.56%**
Positive predictive value	Pr(D| +)	**70.18%**
Negative predictive value	Pr(~D| -)	**97.59%**
False + rate for true ~D	Pr(+|~D)	**0.44%**
False - rate for true D	Pr(-| D)	**70.37%**
False + rate for classified +	Pr(~D| +)	**29.82%**
False - rate for classified -	Pr(D| -)	**2.41%**
Correctly classified		**97.20%**

Universidade de São Paulo

Conceitos básicos: CER *Y*: Exemplo Default com Modelo Logit

Default	Training Data Set	Test Data Set	Entire Sample
Predicted % of Default	3.375	3.369	3.333
Classification Error Rate	0.0280	0.0215	0.0268
N	8,000	2,000	10,000

Universidade de São Paulo

Aprendizado com Modelo Logit: Resultados de Exemplo de Default

```
Learner: Multinomial classification

Dataset information

Target variable = "default_dummy"          Number of features  = 2
N. of training units = 8000                N. of testing units = 2000
N. of used training units = 8000           N. of used testing units = 2000


Validation results

CER = classification error rate            Training CER = .028
Testing CER = .0215
```

Universidade de São Paulo

Aprendizado com Modelo Logit e Cross-Validicação: Resultados de Exemplo de Default

Learner: Multinomial classification

Dataset information

Target variable = "default_dummy"	Number of features = 2
N. of training units = 8000	N. of testing units = 2000
N. of used training units = 8000	N. of used testing units = 2000

Cross-validation results

Accuracy measure = rate correct matches	Number of folds = 5
Training accuracy = .96625	Testing accuracy = .96625
Std. err. test accuracy = 1.110e-16	

Validation results

CER = classification error rate	Training CER = .03375
Testing CER = .0315	

Universidade de São Paulo

Avaliação do Modelo

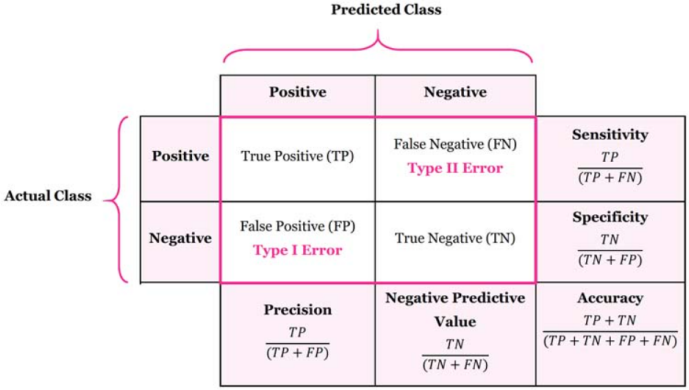


Table 1. Confusion matrix with advanced classification metrics

Universidade de São Paulo

Laboratório 9

Universidade de São Paulo