

# *Principles of Statistical Machine Learning*

## *Measuring the Performance of a Binary Classifier The Receiver Operating Characteristic (ROC) curve*

*Ernest Fokoué*  
福尔特 教授

*School of Mathematical Sciences*  
*Rochester Institute of Technology*  
*Rochester, New York, USA*

*Principles of Statistical Machine Learning*  
*STAT 747-Autumn Semester 2017*

# Study of Diabetes among Pima Indian Women

*Motivating Example:* A study originally published by the National Institute of Diabetes and Digestive and Kidney Diseases sought to determine the relationship between the incidence of diabetes in Pima Indian Women and some specific medical and personal characteristics. A sample of women at least 21 years old and of Pima Indian heritage living near Phoenix were chosen and tested for diabetes.

---

npreg	Number of pregnancies
glu	Plasma glucose concentration
bp	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
bmi	Body mass index kg/m <sup>2</sup>
ped	Diabetes pedigree function
age	Age (years)

---

The response is **type** : **Yes** = diabetic; **No** = Non diabetic.

# Assessing the Predictive Performance of the Classifier

*For notational convenience, I re-coded the data set Pima.tr in such a way that the response variable is now appropriateness named Diabetes and is a binary indicator variable, i.e.*

$$\text{Diabetes} = \begin{cases} 0 & \text{Negative [Diabetes not found]} \\ 1 & \text{Positive [Diabetes was found]} \end{cases}$$

*This coding allows for the use of a standard terminology for assessing the performance of classifiers.*

# Assessing the Predictive Performance of the Classifier

- **[True Positives (TP)]**: The True Positives (TP) count is the number of **positives** correctly classified as **positives**. e.g.: **Diabetic** Pima Indian women correctly declared as **diabetic** by the classifier.
- **[False Positives (FP)]**: The False Positives (FP) count is means the number of **negatives** incorrectly classified as **positives**. e.g.: **Non diabetic** Pima Indian women incorrectly declared as **diabetic** by the classifier.
- **[True Negatives (TN)]**: The True Negatives (TN) count is the number of **negatives** correctly classified as **negatives**. e.g.: **Non diabetic** Pima Indian women correctly declared as non diabetic by the classifier.
- **[False Negatives (FN)]**: The False Negatives (FN) count is the number of **positives** incorrectly classified as **negatives**. e.g.: **Diabetic** Pima Indian women incorrectly declared as **non diabetic** by the classifier.

# Assessing the Predictive Performance of the Classifier

		<i>Prediction</i>	
		<i>Negative</i>	<i>Positive</i>
<i>Actual</i>	<i>Negative</i>	TN	FP
	<i>Positive</i>	FN	TP

**Table:** Counts of possible outcomes of binary classification

*With the 0 – 1 coding, our software delivers a confusion matrix of the form*

		<i>Prediction</i>	
		0	1
<i>Actual</i>	0	TN	FP
	1	FN	TP

# Elements of the Performance the Classifier

Let  $f$  denote our binary classifier. Then  $f$  is a mapping from the input space  $\mathcal{X}$  (usually a subset of  $\mathbb{R}^p$ ) to the indicator set  $\{0, 1\}$ . In other words, to each  $x \in \mathcal{X}$ , the function  $f$  assigns 0 or 1.

$$f : \mathcal{X} \rightarrow \{0, 1\}$$

The estimated accuracy of the classifier (estimated percentage of correct classification) is the most common measure of performance.

$$\widehat{\text{accuracy}}(f) = \Pr(\widehat{Y} = \widehat{f}(X)) = \frac{\text{number of correctly classified}}{\text{sample size}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The number of correctly classified is the sum of the diagonal elements of the confusion matrix. The False Positive Rate (FPR) and the True Positive Rate are also very important measures.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad \text{and} \quad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

# Elements of the Performance the Classifier

*For reasons that will be made clear later, the accuracy of a classifier alone does not tell enough of the story. Consider the following two measures, namely precision and recall. Precision in this context measures the proportion of estimated positives that are truly positives.*

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

*In other words, of all those that were classified as positive, which proportion was indeed positive? Now, a 100% precision says that the classifier all those declared positive where indeed positive. However, it does not necessarily account for all the positive that are there. For instance, a 100% precision on a diabetes classifier means that all the Pima indian women declared diabetic were indeed diabetic. But this does not mean all diabetic Pima Indian women were accounted for.*

# Elements of the Performance the Classifier

*On the other hand,*

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

*In other words, of all those that are actually positive, what proportion did the classifier succeed at classifying correctly. Here a 100% recall says that all positives are classified as positives. However, it does not account for the goodness of all the positive test results. For instance, a 100% recall on a diabetic classification means that all the Pima Indian women that are actually diabetic are declared diabetic by the classifier. This does not tell us anything about those positive tests that might have been on non diabetic Pima Indian women. Recall is sometimes referred to as sensitivity.*



# Elements of the Performance the Classifier

Another important measure is specificity defined by

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

The  $f$ -measure: Clearly, there is a dilemma between Precision and Recall. It seems therefore reasonable to solve this dilemma by combining precision and recall in such a way that a trade-off is achieved. Hence the use of a much more elaborate measure of goodness that combines both precision and recall. This measure is known as the so-called  $f$  measure defined simply by

$$f = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

The famous concept of Receiver Operating Characteristic (ROC) curve plots the The False Positive Rate (FPR) and the True Positive Rate (TPR) as a way to assess the quality of a classification technique.

# Confusion Matrix for Logistic Regression Model

```
Actual <- Pima$Diabetes  
confmat <- table(Actual, Predicted)
```

	Predicted	
Actual	0	1
0	116	16
1	30	38

Acc	FPR	TPR	FNR	TNR	F-measure	Precision	Recall
0.77	0.12	0.56	0.44	0.88	0.62	0.70	0.56

# Receiver Operating Characteristic (ROC) Curve

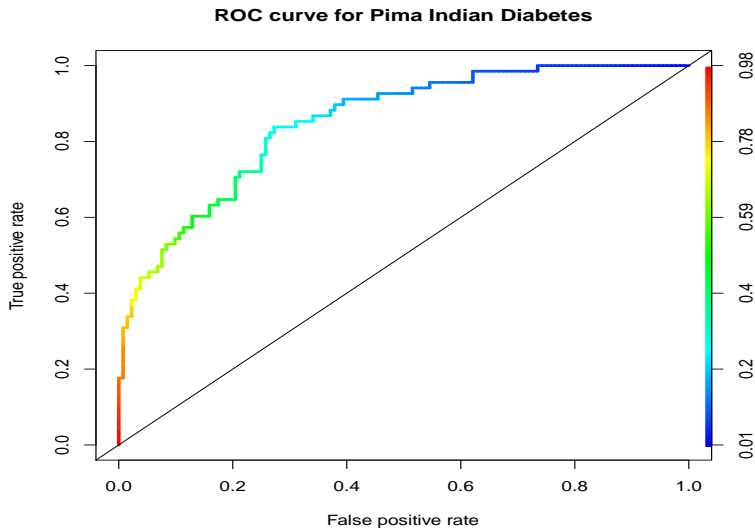
- The *R* package *ROCR* offers many functions for plotting various Receiver Operating Characteristic (ROC) Curves/Graphs.
  - ROC graphs
  - Specificity/Sensitivity curves
  - Lift Charts
  - True Positive Rate (TPR)/False Positive Rate (FPR) Curves
  - Precision/Recall plots

*ROC curves provides one of the excellent devices for assessing the performance of a classifier*

# Advantages of ROC Curves




- *ROCR is a flexible tool for creating cutoff-parametrized 2D performance curves by freely combining two from over 25 performance measures (new performance measures can be added using a standard interface).*
- *Curves from different cross-validation or bootstrapping runs can be averaged by different methods, and standard deviations, standard errors or box plots can be used to visualize the variability across the runs.*
- *The parametrization can be visualized by printing cutoff values at the corresponding curve positions, or by coloring the curve according to cutoff.*
- *All components of a performance plot can be quickly adjusted using a flexible parameter dispatching mechanism. Despite its flexibility, ROCR is easy to use, with only three commands and reasonable default values for all optional parameters.*

# ROC of Logistic Regression on Pima Indian Data



# Exercise session

- ➊ Consider the Crabs *Leptograpsus* dataset and perform a thorough logistic regression analysis on it
  - Comment on the goodness of fit
  - Provide an interpretation for each of the coefficient
- ➋ Consider the German credit dataset and perform a thorough logistic regression analysis on it
- ➌ Consider the Wisconsin breast cancer dataset and perform a thorough logistic regression analysis on it
- ➍ Provide your own dataset and perform a thorough logistic regression analysis on it
- ➎ With logistic regression, what is the increase in probability generated by an increase of  $\delta$  units in the predictor variable  $x_j$  assuming that the rest of variables remain fixed?

-  James, G, Witten, D, Hastie, T and Tibshirani, R (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, (e-ISBN: 978-1-4614-7138-7),(2013)
-  Clarke, B, Fokoué, E and Zhang, H (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer Verlag, New York, (ISBN: 978-0-387-98134-5), (2009)
-  J. J. Faraway(2002). *Practical Regression and ANOVA using R*. Lecture Notes contributed to the R project, (2002)