LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Machine Learning

Evaluation: ROC Analysis 1

Department of Statistics – LMU Munich

IMBALANCED BINARY LABELS

- Consider a binary classifier for diagnosing a serious medical condition
- Here, label distribution is often **imbalanced**, i.e, not many people have the disease
- Evaluating with error rate for imbalanced labels is often inappropriate
- Assume that only 0.5 % of 1000 patients have the disease
- Always returning "no disease" has an error rate of 0.5 %, which sounds good
- However, this sends all sick patients home, which is the worst possible system, even classifying everyone as "sick" might be better, depending on what happens next
- This problem is sometimes known as the **accuracy paradox**

BINARY CLASSIFIERS AND COSTS

- Another point of view is imbalanced costs
- In our example, classifying a sick patient as healthy, should incur a much higher loss then classifying a healthy patient as sick
- The costs depend a lot on what happens next: We can likely assume that our system is some type of screening filter, often the next step after labeling someone as "sick" might be a more invasive, expensive, but more reliable test for the disease
- Erroneously subjecting someone to that second step is not good (psychologically, economically, or because the second test might introduce medical risks), but sending someone home to get worse or die seems much worse
- Such a situation not only arises under label imbalance, but also when labels are maybe balanced but costs differ
- We could see this as imbalanced costs of misclassification, rather than imbalanced labels; both situations are tightly connected

BINARY CLASSIFIERS AND COSTS

- Problem is: If we could specify costs precisely, we could evaluate against them, we might even optimize our model for them
- This important subfield of ML is called **cost-sensitive learning**, which we will not cover in this lecture unit
- Unfortunately, users often have a notoriously hard time to come up with precise cost numbers in imbalanced scenarios
- Evaluating "from different perspectives", with multiple metrics, often helps, especially to get a first impression of the quality of the system

ROC ANALYSIS

ROC Analysis – which stands for "receiver operating characteristics – is a subfield of ML which studies the evaluation of binary prediction systems.

Quoting Wikipdia:

"The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields and was soon introduced to psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, biometrics, forecasting of natural hazards, meteorology, model performance assessment, and other areas for many decades and is increasingly used in machine learning and data mining research"

CONFUSION MATRIX AND ROC METRICS

- We call one class "positive" ($+$) and the other "negative" ($-$).
- Their respective class sizes are denoted by n_+ and n_-.
- The positive class is the more important, often smaller one.
- We represent all predictions in a confusion matrix and count correct and incorrect class assignments
- **False Positive**: We assigned "positive", but were wrong.

		True Class y	
		$+$	$-$
Pred. $\hat{y}$	$+$	**True Positive (TP)**	**False Positive (FP)**
	$-$	**False Negative (FN)**	**True Negative (TN)**

CONFUSION MATRIX AND ROC METRICS

By normalizing the rows and columns of the confusion matrix, we can derive several metrics that help in assessing the performance in imbalanced or cost-sensitive settings (the best possible value for those metrics is 1):

- **TPR**: How many of the true $+$ did we predict as $+$?
- **TNR**: How many of the true $-$ did we predict as $-$?
- **PPV**: If we predict $+$ how likely is it a true $+$?
- **NPV**: If we predict $-$ how likely is it a true $-$?

		True Class y		
		$+$	$-$	
Pred. $\hat{y}$	$+$	**True Positive (TP)**	**False Positive (FP)**	Positive Predictive Value (PPV) = $\frac{\text{TP}}{\text{TP+FP}}$
	$-$	**False Negative (FN)**	**True Negative (TN)**	Negative Predictive Value (NPV) = $\frac{\text{TN}}{\text{FN+TN}}$
		True Positive Rate (TPR) = $\frac{\text{TP}}{\text{TP+FN}}$	True Negative Rate (TNR) = $\frac{\text{TN}}{\text{FP+TN}}$	Accuracy = $\frac{\text{TP+TN}}{\text{TOTAL}}$

EXAMPLE

Suppose 2030 observations with 30 positives, 2000 negatives, and a confusion matrix of a classifier with 91 % accuracy:

		True Class y		
		$+$	$-$	
Pred. $\hat{y}$	$+$	**True Positive (TP)** TP = 20	**False Positive (FP)** FP = 180	PPV = $\frac{20}{200} = 0.1$
	$-$	**False Negative (FN)** FN = 10	**True Negative (TN)** TN = 1820	NPV = $\frac{1820}{1830} = 0.995$
		TPR = $\frac{20}{30} \approx 0.67$	TNR = $\frac{1820}{2000} = 0.91$	Acc.= $\frac{1840}{2030} \approx 0.91$

- An accuracy of 0.91 seems fine.
- However, a PPV of 0.1 is really bad.

MORE METRICS AND ALTERNATIVE TERMINOLOGY

Unfortunately, for many concepts in ROC, 2-3 different terms exist.

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= \frac{1}{\frac{1}{2}\left(\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right)}$

▸ Clickable version/picture source

▸ Interactive diagram

F1-MEASURE

- It is difficult to achieve a high *Positive Predictive Value* and high *True Positive Rate* simultaneously.
- A classifier that predicts more "positives" will tend to be more sensitive (higher TPR), but it will also tend to give more false positives (lower TNR, lower PPV).
- A classifier that predicts more "negatives" will tend to be more precise (higher PPV), but it will also produce more false negatives (lower TPR).

A measure that balances the two conflicting goals is the F_1-measure, which is the harmonic mean of PPV and TPR:

$$F_1 = 2\frac{PPV \cdot TPR}{PPV + TPR}$$

Note that this measure still doesn't account for the number of true negatives.

F1-MEASURE

- A model with $TPR = 0$ (no one predicted as positive from the positive class) or $PPV = 0$ (no real positives among the predicted) has an F1 of 0
- Predicting always "neg" has $F_1 = 0$
- Predicting always "pos" has $F_1 = 2PPV/(PPV + 1) = 2n_+/(n_+ + n)$, which will be rather small, if the size of the positive class n_+ is small.

Tabulated F1-Score for different TPR (rows) and PPV (cols) combinations. We see that the harmonic means tends more towards the lower of the 2 combined values.

```
##     0.0  0.2  0.4  0.6  0.8  1.0
## 0.0   0 0.00 0.00 0.00 0.00 0.00
## 0.2   0 0.20 0.27 0.30 0.32 0.33
## 0.4   0 0.27 0.40 0.48 0.53 0.57
## 0.6   0 0.30 0.48 0.60 0.69 0.75
## 0.8   0 0.32 0.53 0.69 0.80 0.89
## 1.0   0 0.33 0.57 0.75 0.89 1.00
```