

Metrics for Binary Classification

Julien Genovese

Machine Learning Together Milan

14th December 2020



Table of Contents

- ① Classification problem basics
- ② Calibration plot
- ③ Robustness of class Probability
- ④ Metrics for classification

What is classification?

- **Classification is the problem of dividing an observation in a certain category according to its properties.**
- Some examples: Is this a triangle, a rectangle or a circle? What kind of character we have?

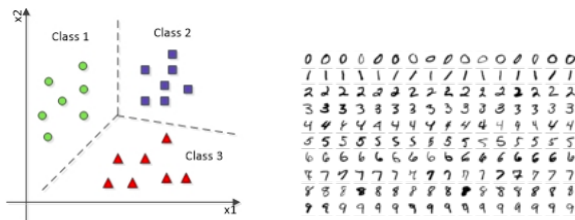


Figure: Two examples of classification

Types of classification problems

- We can have two main types of classification: **binary** and **multiclass**.
- The classification can be **balanced** and **imbalanced**:
 - Balanced: in this case all the classes have similar frequencies.
 - Imbalanced: some classes are more frequent than other ones.
Example: classification between rare and common events.

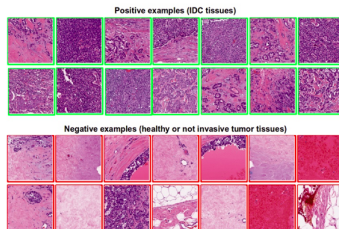


Figure: Disease vs not disease

Mathematical model for classification

- A classifier (algorithm to classify) gives us a **probability** to belong to a class and with a **threshold** that we decide we obtain the **label** associated to a class.
- Classification is a **supervised machine learning problem**, where we have an input of features X , and an output Y , and we want to learn the relationship f between them.
- The mathematical model is in the general form of:

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

where p is the **estimated probability** to belong to a certain class $\in Y$, X the **input**, ε an **error term** related to the stochasticity, f is the **relationship** between the input and the output and this is what the algorithm want to learn.

Binary classification

- In a binary problem we select one of the two class, that we call the "positive class" and p is the probability to belong to this class, $1 - p$ the probability for the other class, the "negative class".
- Example: the logistic regression is in the implicit form of:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and if we explicit p :

$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

Calibration of the Probabilities

- We desire that the estimated class probabilities are reflective of the true underlying probability of the sample, that is, **well-calibrated**.
- Example: if we predict that a mail is spam with a 0.2 probability, we want that if we take 5 similar mails, only one of them is spam.
- To do it we use a **calibration plot**.
- The next definition is **only for balanced problems**.

How to do a Calibration Plot

- First fix a set of bins like $[0, 10\%]$, $(10\%, 20\%]$, ..., $(90\%, 100\%]$.
- In each bin put all the observations you have according to the probability. Example: an observation with 0.15 probability will be assegnated to the bin $(10\%, 20\%]$.
- Compute the number of events with the class associated to the probability p , divide it by the number of observation in the bin: this is the **observed event rate**.
- If the probability reflects the true underlying probability of the sample, the observed event rate must stay on the diagonal. In this case we say that the probabilities are **well-calibrated**.

An example of a Calibration Plot

- In the figure we have the results of two model predictions. We see both an example of well-calibrated and not-well calibrated plot.

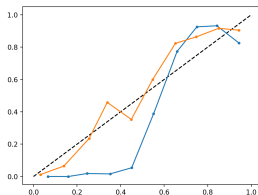


Figure: An example of well-calibrated in orange and not-well calibrated plot in blue

Presenting Class Probabilities

- For a binary classification problem we have a good tool to understand the robustness of a model, i.e. if our model is very sure of its prediction.
- An easy tool is to plot an histogram of the probabilities for each class.
- We want the probabilities for the positive class distributed on the right and the probabilities for the negative one on the left.
- In this way we have a tool to understand how the model is sure of what is saying.

Two examples of Class Probabilities

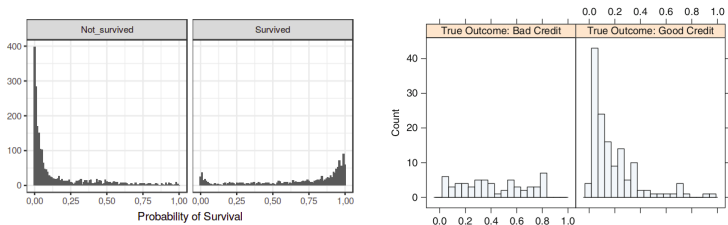


Figure: A robust classifier on the left and a not-robust one on the right

Confusion matrix

- To understand the validity of a classifier we want to understand how many events we have right predicted, how many mistakes have been done, and which ones.
- A classical tool is the **confusion matrix**.

Predicted	Observed	
	Event	Nonevent
Event	TP	FP
Nonevent	FN	TN

Figure: A general confusion matrix

Predicted	Observed	
	Bad	Good
Bad	24	10
Good	36	130

Figure: An example of confusion matrix

- With this matrix we can create different metrics.

Accuracy rate

- The accuracy rate is defined as:

$$\frac{\text{True predictions}}{\text{All the events}}$$

that we translate as:

$$\frac{TP + TN}{TP + TN + FP + FN}.$$

- This metric doesn't take into account:
 - **the errors that have been made.**
 - **the frequencies of each class.**

Ex: in an imbalanced problem the positive class could be 99%. In this case an algorithm that predict all the events as the most frequent class has an accuracy of 0.99 that seems good but it's not.

An example of accuracy computation

Let's suppose we have a confusion matrix like that:

Predicted — Reference	Positive	Negative
Positive +	300	60
Negative -	200	855

In this case the accuracy is:

$$\frac{300 + 855}{300 + 855 + 200 + 60} \cdot 100 = 82\%$$

Sensitivity (Recall) and Specificity

- **Sensitivity** = $\frac{\text{\#positive samples and predicted to be positives}}{\text{\#Positive samples}} = \frac{TP}{TP + FN}$.

This is the ability to **detect positive samples**.

- **Specificity** = $\frac{\text{\#negative samples and predicted as negatives}}{\text{\#Negative samples}} = \frac{TN}{TN+FP}$.

This is the ability to **detect negative samples**.

An example of Sensitivity and Specificity computation

Let's suppose we have a confusion matrix like that:

Predicted — Reference	Positive	Negative
Positive +	300	60
Negative -	200	855

In this case the sensitivity is:

$$\frac{300}{300 + 200} \cdot 100 = 60\%$$

and the specificity:

$$\frac{855}{855 + 60} \cdot 100 = 93\%$$

Probabilistic interpretation

- Sensitivity and specificity are two conditional probabilities.
- We use a medical explanation:
 - Sensitivity is defined as the probability of a positive test result given the presence of disease, written as:

$$P(\text{positive test}|\text{disease present}).$$

- Specificity is defined as the probability of a negative test result given the absence of disease, written as:

$$P(\text{negative test}|\text{disease absent}).$$

- These quantities doesn't depend on prevalence, that is the frequency of the positive class defined as:

$$\text{Prevalence} = \frac{\# \text{Positive samples}}{\# \text{All samples}} = \frac{TP + FN}{TP + FN + FP + TN}$$

because these probabilities only depends on the test.

Sensitivity-Specificity trade-off

- We know that changing the threshold we change the labels and so specificity and sensitivity.
- **Decreasing the threshold** we increase the number of true positives but the number of real positives is the same. So the **sensitivity increases**.
- **Decreasing the threshold** we also increase the number of false positives. So the **specificity decreases**.
- Increasing the threshold we have a similar discourse.
- We use the ROC curve to understand the Sensitivity-Specificity trade-off.

ROC curve

- We have $1 - \text{Specificity}$ on the x - axis and Sensitivity on the y -axis.

$$1 - \text{Specificity} = 1 - \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{FP}}{\text{TN} + \text{FP}}.$$

- The ROC curve is created changing the threshold for the model and predicting the associated label. After that we measure for each threshold/prediction step the Sensitivity and $1 - \text{Specificity}$.

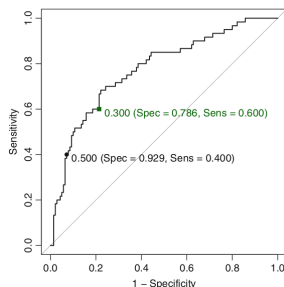


Figure: ROC curve of a classifier against random guess

Some observations on the ROC curve

- The bisector is the ROC curve for the random guessing.
- A perfect classifier has a ROC curve of the type $y = 1$.
- We can use the AUC of the ROC curve to select a model. This method is threshold-insensitive and insensitive to disparity in the class proportions (it's a function of sensitivity and specificity).
- Problems:
 - different curves could cross
 - maybe the threshold is an important discriminating factor.
 - when we will deal with very imbalanced dataset is not reliable.

Select a threshold using ROC curve

- When we are interested in the threshold selection we have to find a trade-off between specificity and sensitivity.
- We can use the *Youden's J index*:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

and finding the threshold that maximizes this function.

- Sometimes we can also use this indicator:

$$I = \text{Sensitivity} \cdot \text{Specificity}$$

PPV and NPV

We introduce two other important metrics:

- Positive Predicted Value (or precision) defined as:

$$\text{PPV} = \frac{\text{\#positive samples and predicted to be positives}}{\text{\#Samples predicted as positive}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Negative Predicted Value defined as:

$$\text{NPV} = \frac{\text{\#negative samples and predicted to be negative}}{\text{\#Samples predicted as negative}} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

- With the PPV and NPV we are seeing how a prediction is reliable.

An example of PPV and NPV computation

Let's suppose we have a confusion matrix like that:

Predicted — Reference	Positive	Negative
Positive +	300	60
Negative -	200	855

In this case the PPV is:

$$\frac{300}{300 + 60} \cdot 100 = 83\%$$

and the NPV is:

$$\frac{855}{855 + 200} \cdot 100 = 83\%$$

Probabilist interpretation

- PPV and NPV are two conditional probabilities too.
- We use a medical explanation:
 - PPV is defined as the probability of the presence of disease given a positive test result, i.e.,

$$P(\text{disease present}|\text{positive test}).$$

- NPV is defined as the probability of the absence of disease given a negative test result, i.e.,

$$P(\text{disease absent}|\text{negative test}).$$

- We can find a relationship with Sensitivity and Specificity using the Prevalence.

Understanding the role of prevalence in PPV and NPV

- We have seen that Sensitivity and Specificity depends only on the test/algorithm.
- The PPV and NPV cannot be independent from it.
- We first see some examples to better understand the idea and we see the mathematical proof of the fact.

Prevalence effect example

We take two medical examples where we use the same test with :

- Sensitivity = 90%
- Specificity = 90%
- 1000 people

and we change the prevalence from 0.5% to 20%.

Prevalence = 5%

- First case: Prevalence = 5% so we have 50 positive people and 950 negative ones. So the confusion matrix is:

Predicted — Reference	Disease	No Disease
Test +	45	95
Test -	5	855

Table: Prevalence = 5%

The PPV is:

$$PPV = \frac{45}{140} \cdot 100 = 32\%$$

that is very low.

Prevalence = 20%

- Second case: Prevalence = 20% so we have 200 positive people and 800 negative ones. So the confusion matrix is:

Predicted — Reference	Disease	No Disease
Test +	180	80
Test -	20	720

Table: Prevalence = 5%

The PPV is:

$$PPV = \frac{180}{260} \cdot 100 = 69\%$$

that is quite good.

Observation on the examples

- The PPV is dependent on the prevalence.
- The PPV can change from one population to another and the same test can be useful or not.

How PPV, specificity, sensitivity and prevalence are related?

We know that:

$$P(\text{disease}|\text{positive test}) = \frac{P(\text{positive test}|\text{disease}) \cdot P(\text{disease})}{P(\text{positive test})}$$

from Bayes theorem and:

$$P(\text{disease}) = \text{prevalence}$$

and

$$P(\text{positive test}) = P(\text{positive test}|\text{disease}) \cdot P(\text{disease}) \quad (1)$$

$$+ P(\text{positive test}|\text{not disease}) \cdot P(\text{not disease}) \quad (2)$$

and so we obtain:

$$\text{PPV} = \frac{\text{Sensitivity} \times \text{Prevalence}}{(\text{Sensitivity} \times \text{Prevalence}) + ((1 - \text{Specificity}) \times (1 - \text{Prevalence}))}$$

Some observations on the PPV

- The PPV is strongly dependent on the prevalence.
- This dependence is a problem. It's not easy to have an idea of the prevalence.

In a spam classifier we can have troubles because there are more schemes to invent spam.

In medicine the prevalence can also change according to the geographical position.

Precision-Recall Curve

- If the prevalence is known and fixed the PPV/precision can be useful.
- There is a trade-off between precision and recall related to the threshold chosen. If we increase the threshold we increase the precision but decrease the recall.

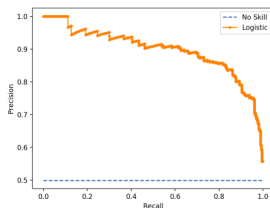


Figure: Precision-Recall curve

Select a threshold using precision-recall curve

- When we are interested in the threshold selection we have to find a trade-off between precision and recall.
- We can use the ***F1 score***:

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

and finding the threshold that maximizes this function.

- Sometimes we can also be more interested on precision or more on recall and we don't want to maximize the trade-off. It depends on the situation.

What curve is better?

- It can be difficult to choose between the ROC and precision-recall curve to select the best threshold.
- In a very imbalanced dataset (1:100), if the prevalence is known and constant, it's better to use the precision-recall curve because the ROC is optimistic.
- This is related to the fact that the false positive rate is:

$$FPR = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

and FP can be very different in order of magnitude with respect to TN . So we can have a FP number high w.r.t. TP but low w.r.t. TN . Therefore we can have a very good recall and sensitivity and low precision.

Example of ROC problem

Let's see an example of the previous problem.

Predicted — Reference	Negative	Positive
Negative	9.4e+04	4.4e+03
Positive	10	1.6e+02

In this case we have:

- Sensitivity: 0.94
- Specificity: 0.95
- Precision: 0.035



Thank you for your attention!!