Zurich University
of Applied Sciences

School of
Engineering

Pa.03 – Lab Assignment

1. Confusion matrices – small example

You are given the following data of loan repayments. These are all customers that were given a loan, so we know whether they were able to repay it or not (“actual repayment” where 1 means “repaid” and 0 means “defaulted”). A classifier has been trained to try to predict the actual repayment with a point prediction (i.e., predicting $\hat{Y}$, not a score) (“predicted repayment” where 1 means “predicted to repay” and 0 means “predicted to default”).

Customer ID	Actual repayment (AR)	Predicted repayment (PR)
#16829	1	1
#22975	1	0
#39753	1	1
#49853	1	1
#51290	1	1
#68923	0	0
#70237	0	0
#81390	0	1
#91234	0	1
#10592	0	0

Calculate the confusion matrix of the actual and the predicted repayment, both with numbers (how many individuals) and with probabilities.

	AR=0	AR=1
PR=0		
PR=1		

2. Confusion matrix – large dataset

Resources for this task

- Dataset ("confusion_matrix.csv", see Moodle)

You are given a dataset of 10,000 individuals. For each individual, you know their true probability p=P[Y=1] and their actual outcome Y.

Apply a decision rule where only individuals with probabilities above 0.7 get D=1. Calculate the resulting confusion matrix.

	Y=0	Y=1
D=0		
D=1		

3. Calculating conditional probabilities

You are given the following confusion matrix:

Confusion matrix	Y=0	Y=1
D=0	0.2	0.3
D=1	0.1	0.4

Calculate the following metrics using the definition of conditional probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$:

(a) True positive rate: P(D=1|Y=1)

(b) False positive rate: P(D=1|Y=0)

(c) True negative rate: P(D=0|Y=0)

(d) False negative rate: P(D=0|Y=1)

(e) Positive predictive value: P(Y=1|D=1)

(f) False discovery rate: P(Y=0|D=1)

(g) Negative predictive value: P(Y=0|D=0)

(h) False omission rate: P(Y=1|D=0)

4. The COMPAS case

Resources for this task

- Preprocessed COMPAS dataset (“compas_preprocessed.csv”, see Moodle)

Download the preprocessed dataset from Moodle. This dataset was gathered by ProPublica for their “Machine Bias” investigation of the COMPAS tool and has then been preprocessed by us in the same way that ProPublica preprocessed the data for their analysis. In the dataset, you will find the risk scores that COMPAS assigned to defendants (column ‘decile_score’). The risk score is a number between 1 and 10. 10 means that COMPAS expects the highest risk of the person being rearrested within 2 years while 1 indicates the lowest predicted probability for being rearrested. According to ProPublica “Scores 1 to 4 were labeled by COMPAS as ‘Low’; 5 to 7 were labeled ‘Medium’; and 8 to 10 were labeled ‘High.’” To make the analysis easier, ProPublica split these scores into “high risk” (labels ‘medium’ and ‘high’, so scores 5 to 10) and “low risk” (label ‘low’, so scores 1 to 4). Therefore, you will also find a “high_risk” feature in the data. 1 indicates a high risk and 0 a low risk. Using the preprocessed data, answer the following questions:

(a) Plot a histogram that shows the distribution of the COMPAS risk scores (x-axis: risk scores, y-axis: frequency among Black/white defendants) but do this separately for Black and white defendants (race = “African-American” vs race = “Caucasian”). Describe your findings.

(b) What is the share of Black and white defendants that are rearrested within 2 years (column ‘two_year_recid’) ? This quantity is called “base rate” and may be different for different subpopulations. Formally, it corresponds to the probability P[Y=1].

(c) Now compare the actual rearrest rates to COMPAS’s predictions from (b) and compare them to the share of Black and white defendants that receive a “high risk” prediction. We might expect that the actual rearrest rates (for the whole dataset, evaluated separately for Black and white defendants) correspond to the share of individuals that received a “high risk” prediction. Is this the case ? And if not, what do we learn from the data about how the discretized prediction model (i.e., the prediction of “high risk”/“low risk”, according to ProPublica’s categorization) works differently for Blacks and Whites ?

(d) Let's go back to the risk scores from the COMPAS algorithm. Is this prediction model (i.e., the COMPAS prediction model) calibrated on the full population? To answer this question, create a calibration plot (x-axis: predicted rearrest rate, y-axis: actual rearrest rate), for all values (1,..,10) of the risk score (column 'decile_score').

(e) Check the calibration of the model separately for Black and white defendants. Can we state that calibration-between-groups is achieved?

(f) Calculate the false positive rate for Black and white defendants (i.e., the share of defendants who are incorrectly labeled as 'high risk' among all the people who are not rearrested). Interpret this difference: What do we learn about how the prediction algorithm works differently for Black and White defendants? Explain this in words, such that a user of COMPAS who is not familiar with technology (imagine a judge as an example) would understand that and how the prediction model works differently.

(g) Calculate the false negative rate for Black and white defendants (i.e., the share of defendants who are incorrectly labeled as 'low risk' among all the people who go on to get rearrested). Interpret your findings as you did with the false positive rate.

5. Conditional probabilities for breast cancer detection

A hospital wants to introduce 3D mammography on top of the existing 2D mammography to improve the early detection of breast cancer. The mammography diagnosis is a prediction model, whose input is an x-ray photo, and whose output is a point prediction $\hat{Y}$ ($\hat{Y} = 1$ for a positive diagnosis, $\hat{Y} = 0$ for a negative diagnosis).

The hospital tests combining these two procedures on 1000 patients to compare it to 2D mammography alone which is how they have tested patients for many years. In this population, 10% of patients actually had breast cancer ($Y = 1$).

In the population of patients who were screened with 2D mammography alone, 89.7% of all patients correctly received a negative diagnosis while 7.6% of the patients correctly received a positive diagnosis.

Combining 2D mammography with 3D mammography had the following effect on the 1000 patients: Among the 100 patients with breast cancer, 81 have been correctly diagnosed with breast cancer. In total, 83 patients were diagnosed with breast cancer.

Fill in the confusion matrix for the old procedure (2D mammography) and for the combination of the new procedure (2D and 3D mammography).

2D mammography	$Y = 0$	$Y = 1$
$\hat{Y} = 0$		
$\hat{Y} = 1$		

2D+3D mammography	Y=0	Y=1
D=0		
D=1		

For the following values, write down what conditional probability they correspond (i.e., the name, e.g., false negative rate), the conditional probability ($P[\widehat{Y} = 0|Y = 1]$) to and calculate its value based on the confusion matrices.

2D mammography:

- Among the patients with breast cancer, what is the share of those who were correctly diagnosed with breast cancer?
- Among patients diagnosed with breast cancer, what is the share of those who actually had breast cancer?
- Among patients with a negative test, what is the share of those who should have received a positive test because they had breast cancer?
- Among patients with a negative test, what share received the correct result?

2D+3D mammography:

- Among the patients with a negative test, what share received the wrong diagnosis?
- Among patients who did not have breast cancer, what share was incorrectly diagnosed with breast cancer?
- Among patients with a positive result, what share was actually breast-cancer-free?