#### **Logistic regression**

Genoveva Vargas-Solar French Council of Scientific Research, LIG genoveva.vargas@imag.fr

http://vargas-solar.com/data-centric-smart-everything/

## Logistic regression assumptions

- Dependent variable is binary
- Observations are independent of each other
- Little or no multicollinearity among the independent variables
- Linearity of independent variables and log odds

#### Logistic regression linear models (discriminative)

Dependent variable is binary Observations are independent of each other Little or no multicollinearity among the independent variables Linearity of independent variables and log odds

• Sigmoid function or logistic function:

$$orall z\in \mathbb{R}, \hspace{1em} g(z)=rac{1}{1+e^{-z}}\in ]0,1[$$

• Logistic regression: assume that  $y|x; \vartheta \sim \text{Bernoulli}(\varphi)$ 

$$\phi = p(y=1|x; heta) = rac{1}{1+\exp(- heta^T x)} = g( heta^T x)$$

Remark: there is no closed form solution for the case of logistic regressions.

• Odds ratio represents the constant effect of predictor X on the likelihood that on output will occur

### Logistic regression pipeline (1) linear models (discriminative)

- Reading data
- Basic explanatory data analysis (EDA)
  - Find non-numerical values / missing / null values
  - Descriptive Analysis: skewness, outliers, mean & median, correlation using pair plot
  - Pair plot many distributions each for every variable
- Model: select independent attributes, class variables, test size, seed repeatability of the code
- Train and test data splitting
- Accuracy report

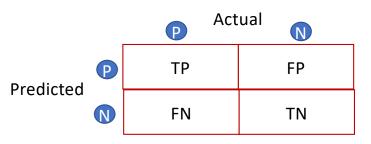
#### Logistic regression pipeline linear models (discriminative)

- Step 1: Classifying inputs to be in class 0/1
  - Compute the probability that an observation belongs to class 1 using a **logistic response function** 
    - Logit function  $P(y=1) = 1 / 1 + e^{-(\beta 0 + \beta 1 x i \dots + \beta n x n)} \beta_i$  selected to maximize the likelihood of predicting
    - Log odds Odds = P(y=1) / P(y=0) = the odds > 1 with high probability of y=1

the odds < 1 with high probability of y=0

```
Logit(P) = a+ bx
```

- Step 2: Defining the boundary for the odds (> 0,5)
  - p determines the FN FP to allow



Accuracy: how often is it correct Precision when P how often is it correct Recall when actually positive how often is it correctly predicted F1 harmonic mean AUC (receiver operating characteristic) TP rate sensitivity, FP rate specificity TN/TN+FP FPR 1- Specificity

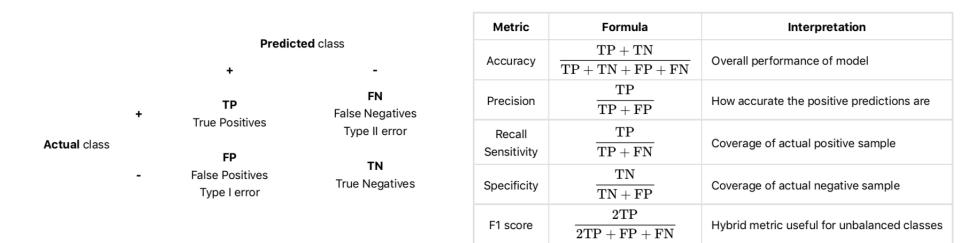
# Logistic regression & classification linear models (discriminative)

- Softmax regression: multiclass logistic regression
  - used to generalize logistic regression when there are more than 2 outcome classes.
  - By convention, we set  $\vartheta_k = 0$ , which makes the Bernoulli parameter  $\varphi_i$  of each class *i* equal to:

$$\phi_i = rac{\exp( heta_i^T x)}{\displaystyle\sum_{j=1}^K \exp( heta_j^T x)}$$

#### Classification metrics (1)

- In a context of a binary classification
- **Confusion matrix** used to have a more complete picture when assessing the performance of a model



#### Classification metrics (2)

• **Receiver operating curve (**ROC), is the plot of TPR versus FPR by varying the threshold

Metric	Formula	Equivalent
True Positive Rate TPR	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	Recall, sensitivity
False Positive Rate FPR	$\frac{\rm FP}{\rm TN+FP}$	1-specificity

• AUC/AUROC — area under the receiving operating curve is the area below the ROC <sup>1</sup>

