Logistic Regression

ML2: AI Concepts and Algorithms (SS2025)
Faculty of Computer Science and Applied Mathematics
University of Applied Sciences Technikum Wien

Lecturer: Rosana de Oliveira Gomes
Author: M. Blaickner, B. Knapp, S. Rezagholi, R.O. Gomes

Classification
KNN classification
Classification trees
Ensembles & boosting
Random Forest
Logistic regression
Naive Bayes
Support vector machines
Neural networks

Regression
KNN regression
Regression trees
Linear regression
Multiple regression
Ridge and Lasso regression
Neural networks

Supervised learning

AI

Data handling
EDA
Data cleaning
Feature selection
Class balancing
etc

Non-supervised learning

Clustering
k-means
Hierachical clustering
DB-scan

Dimensionality reduction
PCA / SVD
tSNE
MDS

Reinforcement learning
Covered in a separate lecture.

FH University of Applied Sciences TECHNIKUM WIEN

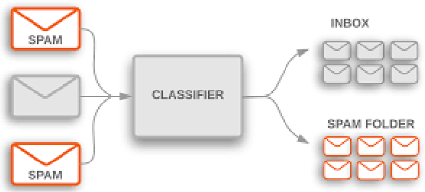
Classification Problems

- Problems in which an ML model predicts to which **category** an input belongs to.
  ML uses training observations to build a classifier.

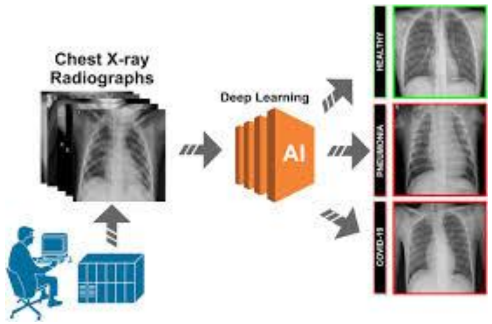
- Logistic regression is a **classification** method: applicable to different data types (tabular, text, image, etc) and number of categories.

- Examples:

Spam detection



Medical diagnosis



Fraud detection

FH University of Applied Sciences TECHNIKUM WIEN

Classification and Regression

- **Classifiers** model the **probability of belonging to a category** (similar to regression).

- Binary Classification: categorical response variable can be determined by a linear regression with a dummy variable approach:

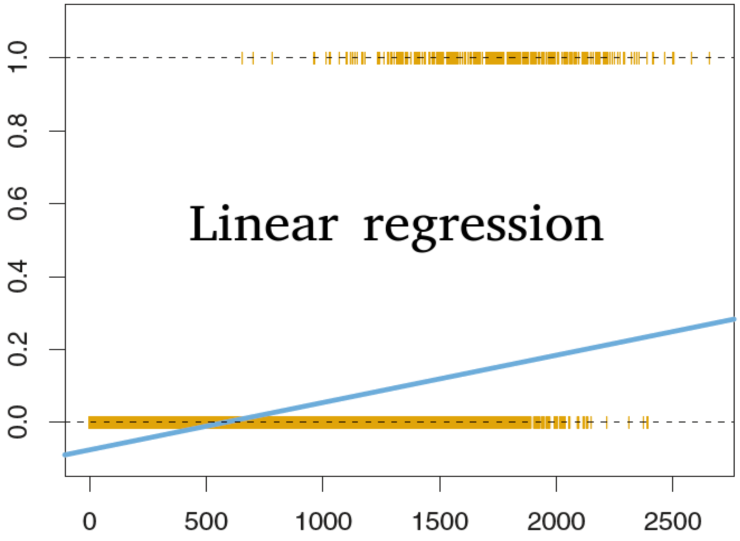
$$y_i = \begin{cases} 1 \text{ if patient } i \text{ was sent to ICU,} \\ 0 \text{ otherwise.} \end{cases}$$

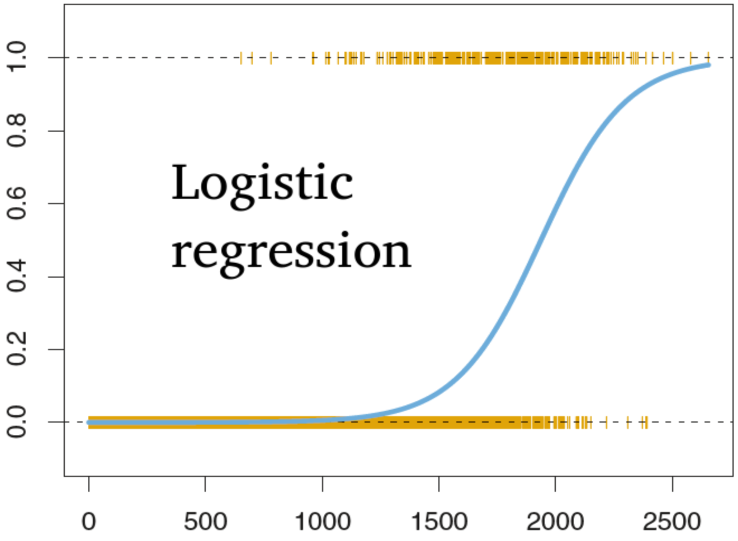
- **If the predicted value is larger than 0.5** then predict *'ICU'* and otherwise predict *'non-ICU'*.

Linear Regression for Binary Classification?

Problem:

The use of linear regression **can yield values outside the interval [0, 1]** of probability-values.





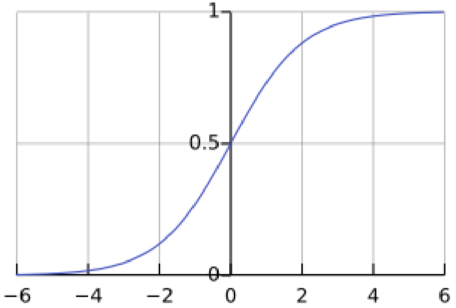
Modified from [1]

Logistic Model

- A **sigmoid function (logistic function):** squashes a real number onto the interval (0,1).

$\sigma : \mathbb{R} \rightarrow (0,1)$ is defined as follows:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



As $z \rightarrow \infty$, $\sigma(z) \rightarrow 1$.

As $z \rightarrow -\infty$, $\sigma(z) \rightarrow 0$.

If $\sigma(z) > 0.5$, predict **Class** 1.

If $\sigma(z) < 0.5$, predict **Class** 0.

- First **apply a function *z* to the feature vector** and then **squash the result onto the unit interval**.
- A possible classification threshold could be 0.5. If the output of the logistic function is bigger than 0.5 the patient is classified as ICU. More conservative classifications correspond to lower thresholds.

FH University of Applied Sciences TECHNIKUM WIEN

Logistic Model

Interpretation:

Standard Logistic Regression (Linear z)

$$z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

- As in regression **training data** is used to **fit** or train the model: find optimal values for the parameters ω_0, ω_1, …, ω_p .

Example: for the ICU problem with only one variable x_1 representing age, inserting x_1 into the equation above (with optimized ω_0 and ω_1) yields the **probability** of a **patient with age x_1** being **admitted to ICU.**

FH University of Applied Sciences TECHNIKUM WIEN

Multiple Logistic Regression

- Going back to multiple logistic regression.

$$x \mapsto \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}$$

- Just as in linear regression variables can be correlated. We can i**ncorporate interaction terms** as in linear regression.

$$x \mapsto \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \beta_{1,2} x_1 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \beta_{1,2} x_1 x_2)}$$

FH University of Applied Sciences TECHNIKUM WIEN

Multiple-Class Logistic Regression

Example:

Classification into more than two classes is a common problem. Example: 'non-hospital', 'hospital but no ICU', and 'ICU'.

- Multiple-class extensions of logistic regression are possible and some solutions are available, but in practice they are not used often.
- One can **split the problem into many binary classification problems** and predict the class for which the respective logistic model gives the largest probability. Attention: This approach is dangerous because the confidence intervals may be different for the models. Careful analysis is necessary!
- For multiclass classification other approaches are preferred (kNN, decision tress, neural networks, …).

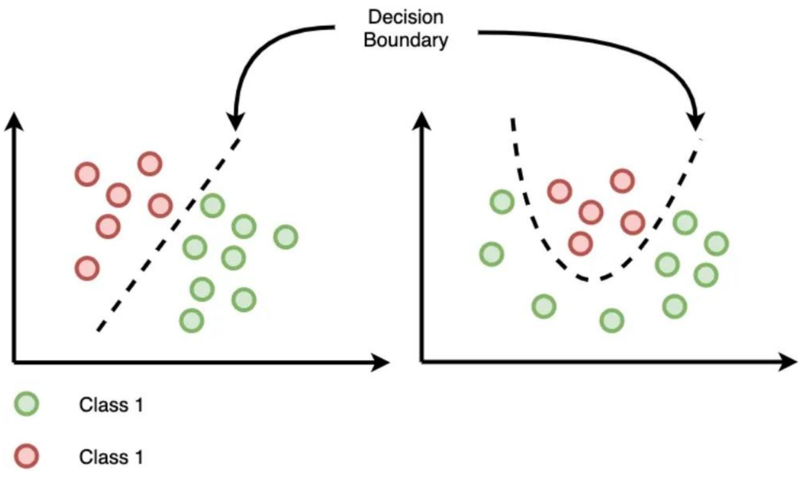
FH University of Applied Sciences TECHNIKUM WIEN

Decision Boundary

A **decision boundary** is a surface (line, curve, or higher-dimensional hyperplane) that **separates different classes** in a classification problem.

It defines the threshold where the model switches from predicting one class to another.

Set of points where the predicted probability is exactly **sigma=0.5, which corresponds to *z=0***



Standard Logistic Regression (Linear z)

$$z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

Polynomial Logistic Regression (Non-linear z)

$$z = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots$$

FH University of Applied Sciences TECHNIKUM WIEN

Quiz

You are using logistic regression to predict whether a customer will purchase a product (1) or not (0) based on their income. The model outputs a probability $P(y=1|x)=0.8$ for a customer with an income of $50,000. Which of the following is a correct interpretations of this probability?

A) The customer will buy the product 80% of the time.
B) The odds of the customer buying the product are 5:1.
C) The customer is more likely to buy than not, but the model does not guarantee it.
D) The purchase decision is determined solely by the sigmoid function.

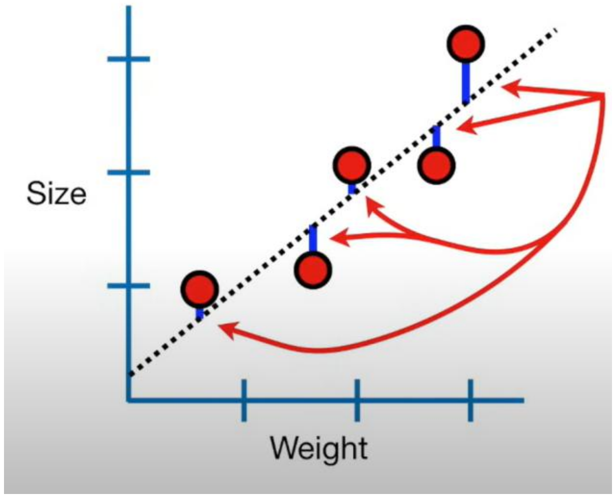
You are given a dataset with binary labels (0 and 1). After training a logistic regression model, you observe that the decision boundary is a straight line in the feature space. What can you conclude?

A) Logistic regression is always a linear classifier, regardless of data distribution.
B) The dataset is perfectly linearly separable.
C) The sigmoid function forces the decision boundary to be linear.
D) The model is underfitting.

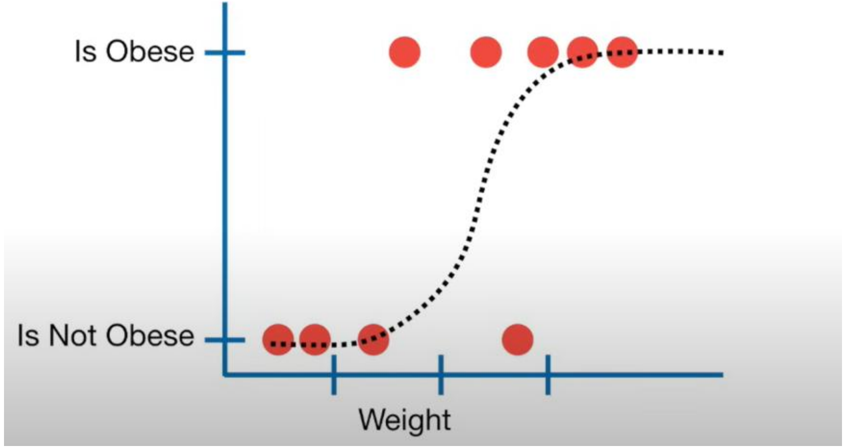
FH University of Applied Sciences TECHNIKUM WIEN

Evaluation Metrics

Least squares: curve that minimizes the sum of the square of the residuals (Linear Regression)



Logistic regression does not have the concept of a residual (binary classification).



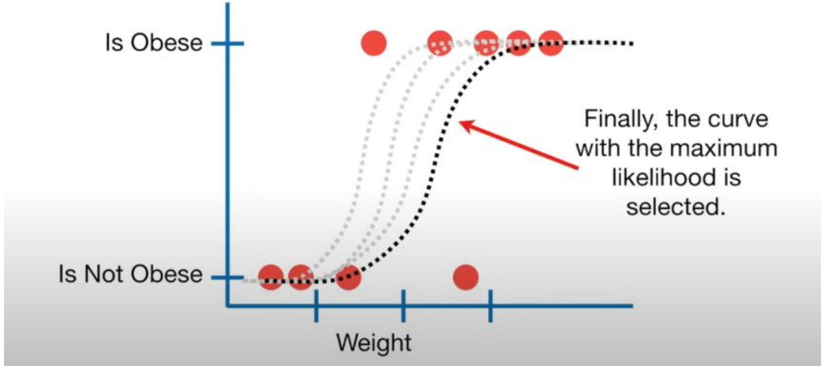
FH University of Applied Sciences TECHNIKUM WIEN

Maximum Likelihood for Logistic Regression

"Maximum likelihood estimation is a method of estimating the parameters of a statistical model, given observations. The method obtains the parameter estimates by finding the parameter values that maximize the likelihood function"

https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

How to obtain the maximum likelihood estimator for a logistic curve?



FH University of Applied Sciences TECHNIKUM WIEN

Maximum Likelihood for Logistic Regression

Likelihood of data given a function is the multiplication of the likelihood of each data point.

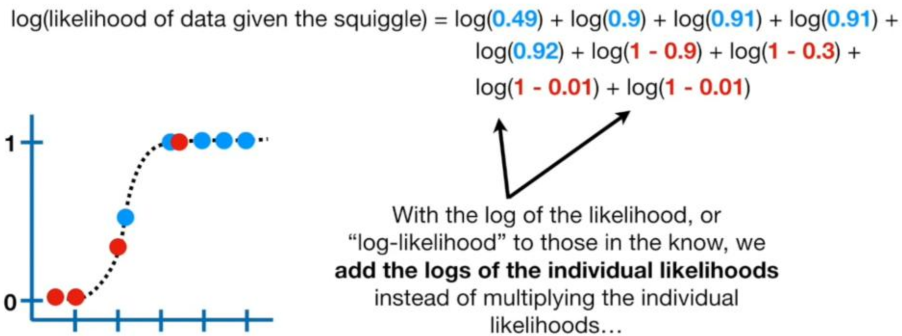
(Odds of being in **Class 0**) X (Odds of being in **Class 1**)

The Maximum Likelihood Estimation (MLE) for Logistic Regression is defined as:

$$L(\beta_0, \ldots, \beta_p) = \prod_{i:y_i=1} \text{model}(x_i) \prod_{i:y_i=0} \Big(1 - \text{model}(x_i)\Big)$$

The parameters $b_0, b_1, \ldots, b_p$ are chosen to **maximize this likelihood** function.

To make the likelihood symmetric, the natural logarithm of the likelihood is taken:



P.S.: Mathematical details of the maximum likelihood method are not part of this lecture. See more here .

Evaluation Metrics

Confusion matrix: summarizes the performance of a classification algorithm by showing the number of correct and incorrect predictions.

- True Positives (TP):
Correctly predicted positive instances

- True Negatives (TN):
Correctly predicted negative instances

- False Positives (FP):
Incorrectly predicted as positive (Type I error)

- False Negatives (FN):
Incorrectly predicted as negative (Type II error)

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP)**	**False negative (FN)**
	Negative (N)	**False positive (FP)**	**True negative (TN)**

https://en.wikipedia.org/wiki/Confusion_matrix

Evaluation Metrics

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Accuracy: proportion of correctly classified instances (both true positives and true negatives) out of all instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: proportion of *positive* predictions that are *actually correct*.

$$Precision = \frac{TP}{TP + FP}$$

Recall: proportion of *actual positives* that are *correctly identified*.

$$Recall = \frac{TP}{TP + FN}$$

ROC-AUC: area under the *Receiver Operating Characteristic (ROC) curve*. Measures the model's ability to **distinguish between classes** at different thresholds.

F-score: harmonic mean of precision and recall. It balances the trade-off between precision and recall, especially useful when there is class imbalance.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

FH University of Applied Sciences TECHNIKUM WIEN

Evaluation Metrics

Metric	High Value Meaning	Low Value Meaning
Accuracy	The model correctly classifies most instances overall. Works well if the dataset is balanced.	The model makes frequent misclassifications. May indicate underfitting or an imbalanced dataset.
Precision	When the model predicts a positive class, it is usually correct. Few false positives.	Many false positives; the model incorrectly labels negative instances as positive.
Recall	The model correctly identifies most actual positive cases. Few false negatives.	Many false negatives; the model fails to detect a significant number of actual positives.
F1-score	The model balances precision and recall well. Useful when both false positives and false negatives are important.	Either precision or recall (or both) are low, indicating poor overall classification performance.
AUC-ROC	The model effectively distinguishes between classes at all classification thresholds.	The model is close to random guessing (AUC ≈ 0.5), indicating poor discriminatory power.

FH University of Applied Sciences TECHNIKUM WIEN

Challenges & Mitigation

Challenges:

- **Imbalanced datasets:** lead to biased predictions towards the majority class and result in poor model performance for the minority class.

- **Multicollinearity:** (when two or more predictor variables are highly correlated) makes it difficult to determine the individual effect of each predictor and make the model unstable.

- **Outliers:** disproportionately influence the model, affecting coefficients and potentially leading to erroneous conclusions.

- **Overfitting:** can happen if the number of observations is lesser than the number of features, it may lead to overfitting.

Challenges & Mitigation

Challenges:

- **Imbalanced datasets:** lead to biased predictions towards the majority class and result in poor model performance for the minority class.

- **Multicollinearity:** (when two or more predictor variables are highly correlated) makes it difficult to determine the individual effect of each predictor and make the model unstable.

- **Outliers:** disproportionately influence the model, affecting coefficients and potentially leading to erroneous conclusions.

- **Overfitting:** can happen if the number of observations is lesser than the number of features, it may lead to overfitting.

Mitigation Strategies:

- Alternative metrics specific for classification (precision, recall, F-score)
- Feature Scaling
- Resampling
- Feature Selection
- Regularization (L1 and L2)

Quiz

You are working on a medical diagnosis problem where you use logistic regression to classify patients as having a disease (1) or not (0). Your model's precision is 0.9 and recall is 0.6. What does this tell you about the model's performance?

A) The model detects most of the patients who have the disease.
B) The model misclassifies many non-diseased patients as diseased.
C) The model detects some of the diseased patients but misses many.
D) The model is perfectly balanced.

A logistic regression model achieves 99% accuracy on an email spam detection task but performs poorly in real-world predictions. What is the most likely issue?

A) The model is overfitting.
B) The dataset is imbalanced.
C) Logistic regression is not suitable for binary classification.
D) The decision boundary is nonlinear.

FH University of Applied Sciences TECHNIKUM WIEN

Takeaway

Logistic Regression

- Method used to solve problems involving **classification (category prediction)**
- It uses a sigmoid function that transforms values into a (0,1) interval.
- A **decision boundary** will define a threshold curve or hyperplane to which the categories are separated in the space of parameters
- The **maximum likelihood estimation (MLE)** is a common metric to evaluate classification problems. **Other evaluation metrics** are needed to understand the behavior of a model:

Metric	High Value	Low Value
Accuracy	Correctly classifies most cases	May be misleading for imbalanced data
Precision	Few false positives	Many false positives
Recall	Few false negatives	Many false negatives
F1-score	Good balance of precision & recall	Either precision or recall is low
AUC-ROC	Good class separation	Model is close to random guessing

FH University of Applied Sciences TECHNIKUM WIEN

Assignment: Logistic Regression

a) Explain logistic regression as pseudo-code or via visualizations.

Use self-made images or even hand drawings (of which you take a photo).

Use self written explanations.

Do not copy from the lecture slides or the internet (neither text nor images).

b) Use sklearn.linear_model.LogisticRegression and compare the results with a tree and kNN model.

Generate 5 datasets using sklearn.datasets.make_classification()

Go through the whole ML workflow: (1) 10-fold cross validation, (2) scaling, (3) hyperparameter training, etc. on all 5 datasets.

Which of the 3 algorithms performs best on average (mean +/- standardDeviation) on each of the 5 datasets and which algorithm is the overall winner?

Which characteristics of a dataset advantage a certain algorithm? Interpret the results.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

FH University of Applied Sciences TECHNIKUM WIEN