czechitas

6th Lecture: Classification Part I

Justina Ivanauskaite, Pavel Fišer, Anna Štrobová

30.5.2023

Your team today

Justina

Data Science Lead,

MSD Animal Health

Pavel

Data Scientist,

MSD Animal Health

Anna

Data Scientist,

Merck Research Labs IT

czechitas

Today's structure

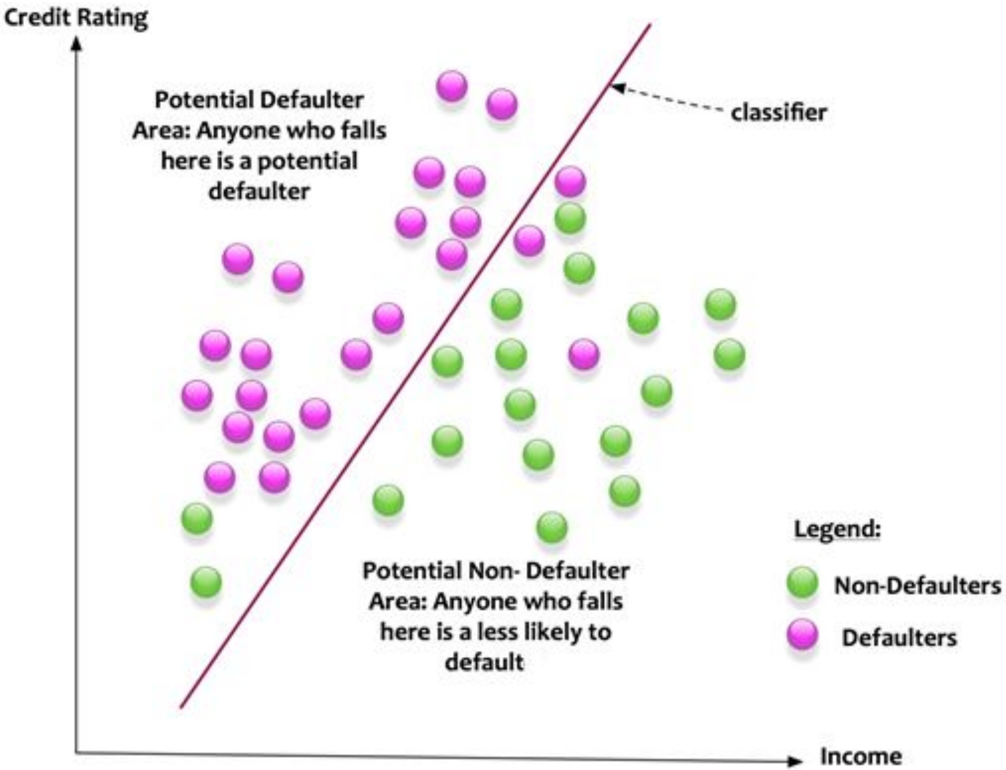
1. Why we need classification?
   When to use classification?
2. Overview of most common methods
3. How to choose between different methods?
4. Logistic regression: intro
5. Logistic regression: hypothesis testing
6. Logistic regression: prediction

Regression vs. Classification

What's the difference in these examples?

1. Email spam detection
2. Predict temperature based on various characteristics (humidity, wind speed)
3. Client risk prediction for loans
4. Estimate of your apartment price when selling it

Classification



Credit Default = a binary variable!

Classification

CustomerID	Income	Education	Age	Default
2343	50 000	17	35	No
1213	35 000	15	32	Yes
4533	40 000	15	53	No
4563	100 000	19	51	No
7554	50 000	18	28	No
6465	27 500	13	25	Yes
7453	34 000	13	32	No
6775	72 000	18	43	No
4643	50 000	19	47	No
6886	48 000	19	37	?
8668	62 500	21	39	?
8765	78 000	23	46	?
9797	23 000	12	29	?

Labeled Data

Unlabeled Data

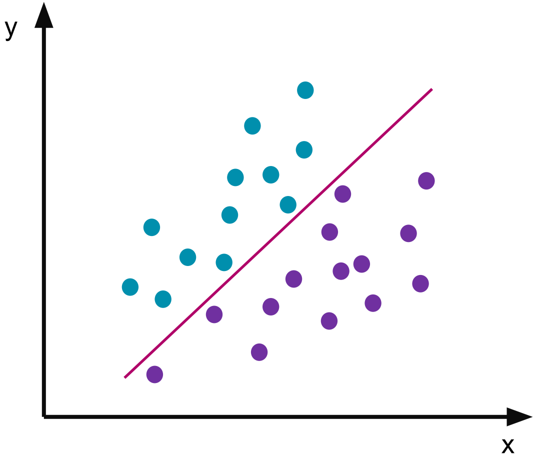
czechitas

Classification

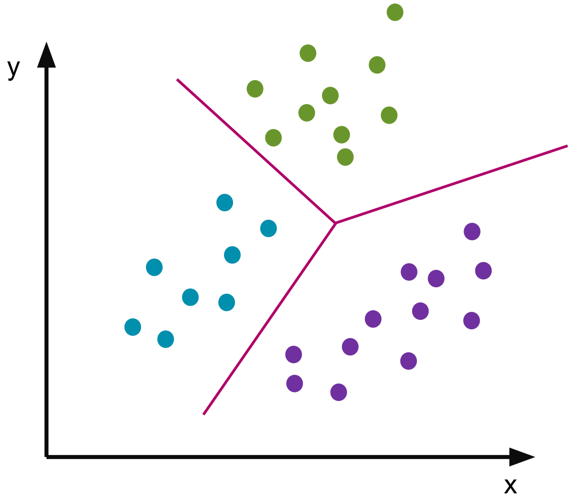
- Discrete response (instead of continuous)
- Model evaluation - accuracy, F1 score, sensitivity, etc. (instead of R^2)
- Dependent variable can be binary or multi-class (with special case of multi-label)
- Supervised learning
- Structured or unstructured data

Types of Classifications

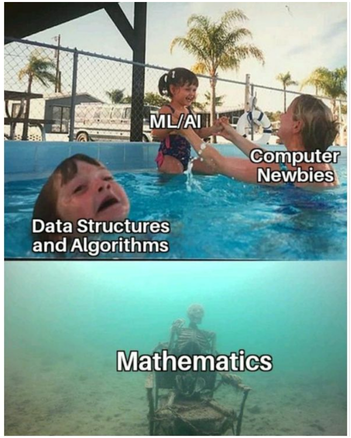
Binary



Multi-Class



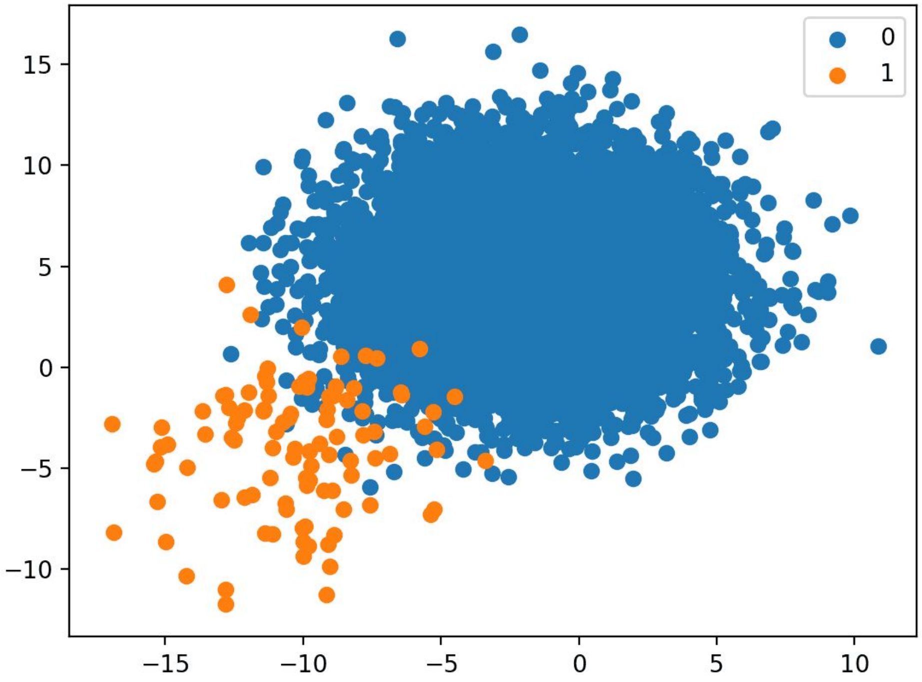
Multi-Label



Labels:

- ☑ Child
- ☒ Snow
- ☑ Water
- ☑ Tree
- ☒ Bike

Imbalanced Classifications



Always check
the distribution
Cannot be ignored!

Binary

Multi-Class

Multi-Label

Examples of classification problems

- Email spam detection (spam or not) Binary
- Client risk prediction (risky or not) Binary
- Risk assessment of audit outcomes (high or low risk) Binary
- Negative comment classification (threat, toxic, obscene, insult..) Multi-Label
- Face classification Multi-Label
- Animal species classification Multi-Class

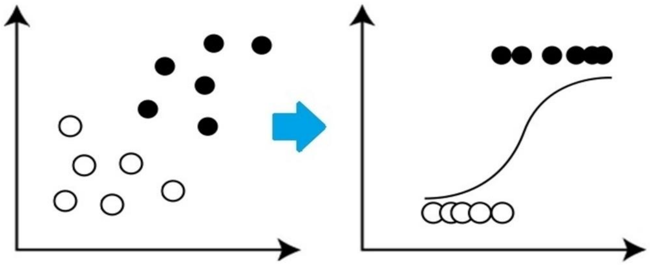
czechitas

Today's structure

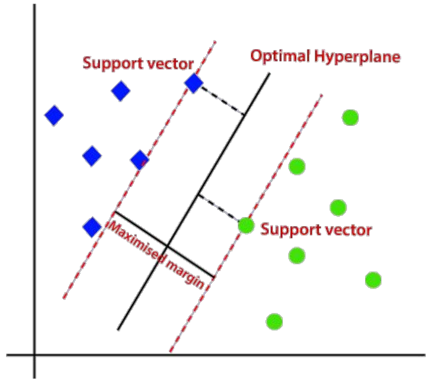
1. Why we need classification?
   When to use classification?
2. Overview of most common methods
3. How to choose between different methods?
4. Logistic regression: intro
5. Logistic regression: hypothesis testing
6. Logistic regression: prediction

Classification methods

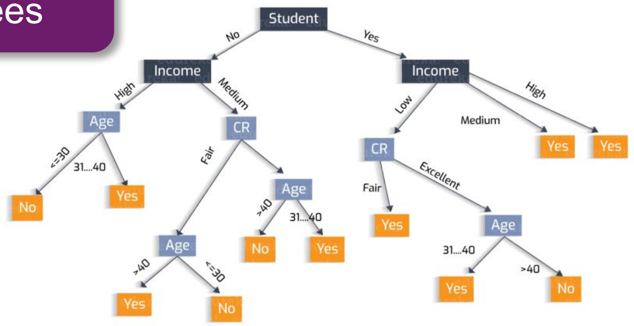
Logistic regression



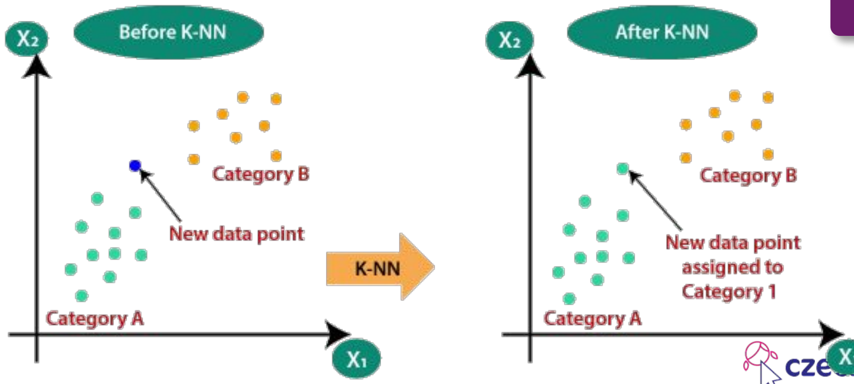
Support Vector Machines



Decision Trees



k-NN



czechitas

Today's structure

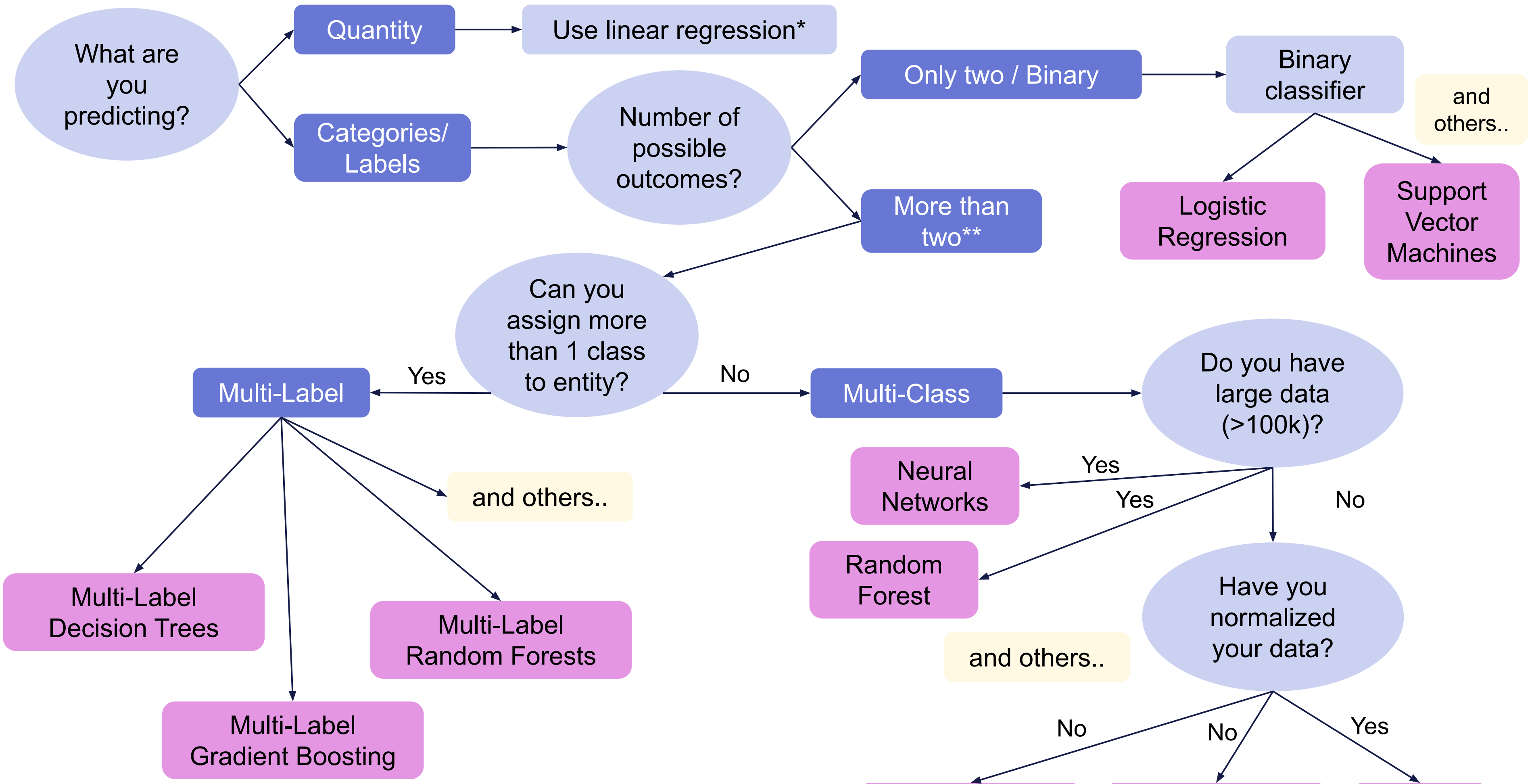
1. Why we need classification?
   When to use classification?
2. Overview of most common methods
3. How to choose between different methods?
4. Logistic regression: intro
5. Logistic regression: hypothesis testing
6. Logistic regression: prediction

How to choose between different methods?

- What are you predicting? *(continuous or categorical response)*
- How many possible outcomes are there? *(two or more)*
- Can you assign more than 1 class to entity? *(multi-class vs. multi-level)*
- Do you have large or small data? *(>100k)*
- Have you normalized your data?
- *Do you have missing values in the data? Are the parameters independent and identically distributed? Is there multicollinearity among the independent variables? ...*

Check algorithm assumptions before applying it!

czechitas



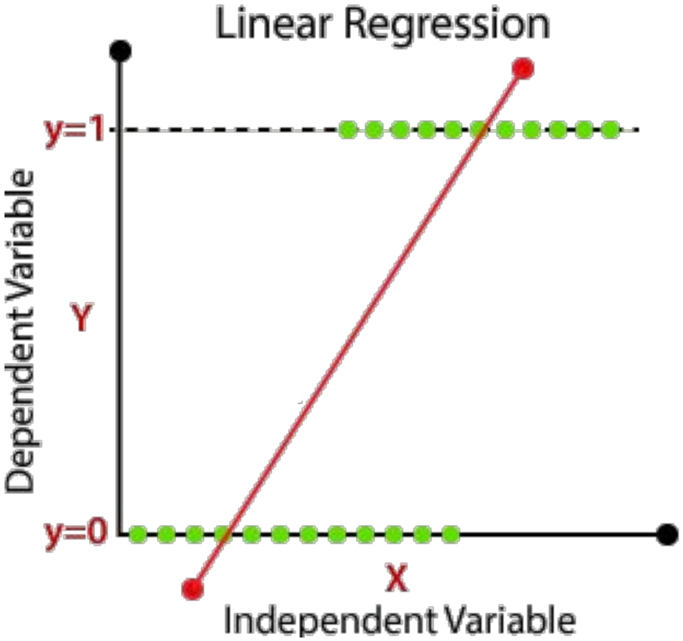
** or other methods for continuous dependent variables (incl. some mentioned as the multi-class m.)*
*** the multi-class methods can be applied on binary class. as well*

Today's structure

1. Why we need classification?
   When to use classification?
2. Overview of most common methods
3. How to choose between different methods?
4. Logistic regression: intro
5. Logistic regression: hypothesis testing
6. Logistic regression: prediction

Linear (OLS) vs Logistic regression in classification problem

Quiz: What do you think is the problem of using linear regression model on classification problem? For example Client risk prediction



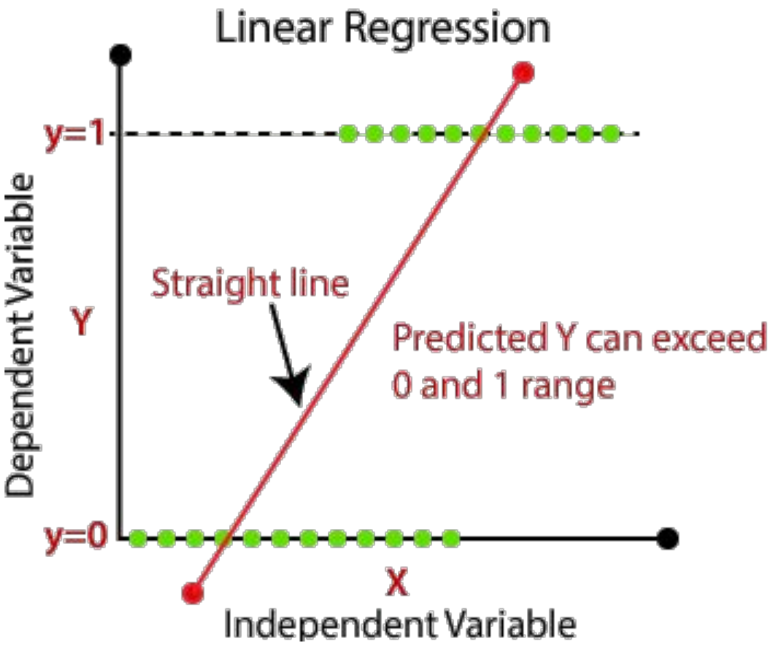
czechitas

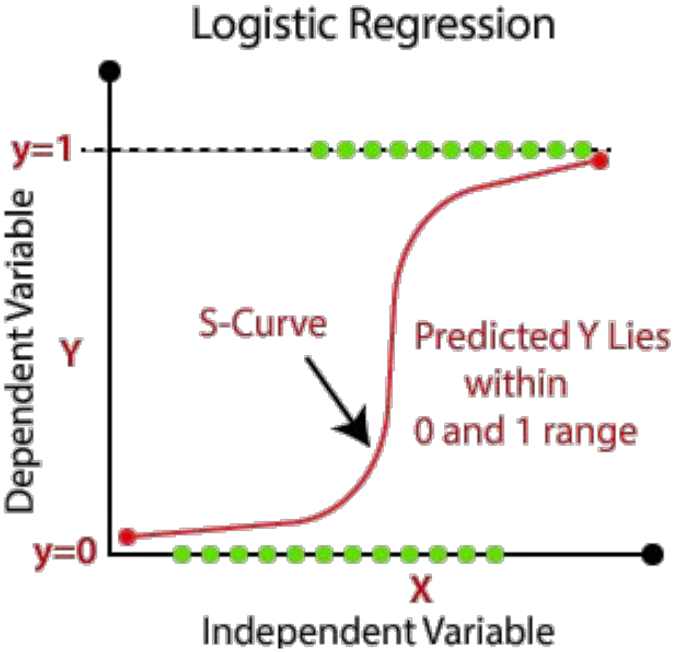
Source: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Linear (OLS) vs Logistic regression in classification problem

Linear relationship
Continous dependent variable

S curve relationship (limits at 0 and 1)
Binary dependent variable
Model can predict probability between 0 and 1

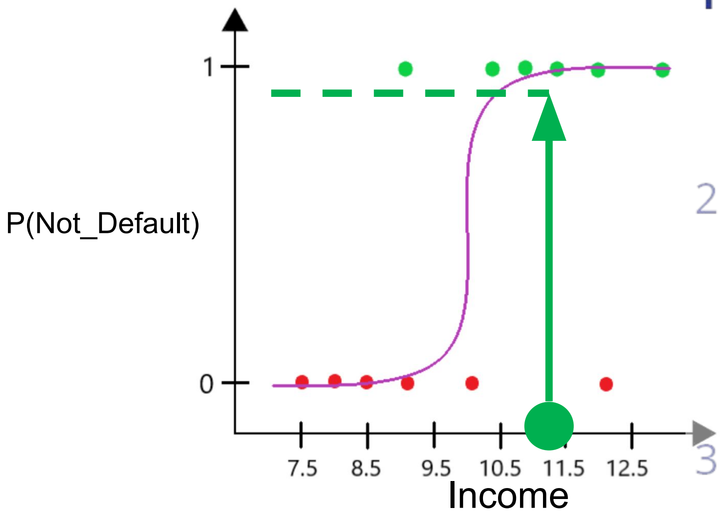




czechitas

Source: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Logistic regression specification



1. **In general, logistic regression provides probability of outcome y**

$$\boldsymbol{p(y) = F(x_1, x_2)}$$

2. To get p(y) probability between 0, 1 we need to do non-linear transformation

$$p(y) = \frac{e^{\beta_0 + \beta_1 * x_1 + \beta_2 * x_2}}{1 + e^{\beta_0 + \beta_1 * x_1 + \beta_2 * x_2}}$$

3. We introduce concept of odds ratio

$$Odds\ Ratio = \frac{p(y)}{1 - p(y)} = e^{\beta_0 + \beta_1 * x_1 + \beta_2 * x_2}$$

4. **Final logistic regression coefficients are estimated:**

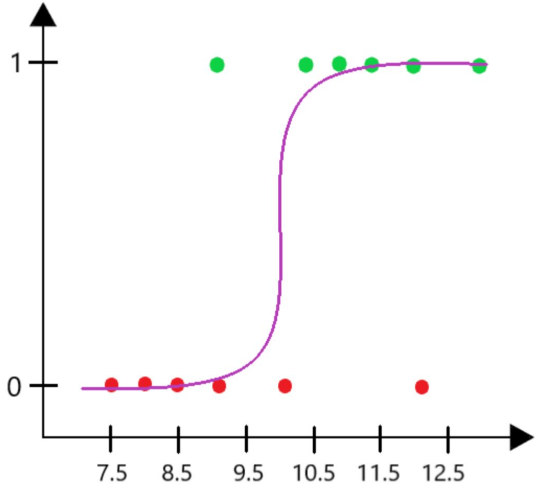
$$\boldsymbol{\ln\left(\frac{p(y)}{1 - p(y)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

czechitas

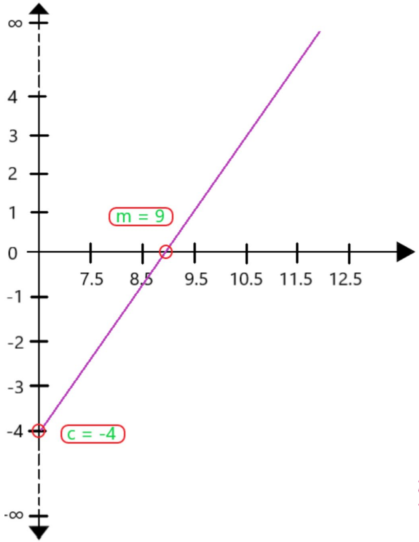
Why we are using Odds and probabilities?

Probability - Change in 1 year have varying impact
Odds - Change in 1 year have constant impact
=> Odds have better properties then probability

Probability



Odds



czechitas

What is the difference between odds and probability?

$$\text{Odds ratio} = \frac{p}{1 \ - p}$$

p → probability (odds) of success
$1 - p$ → probability (odds) of failure

Quiz:

1. What is the $\text{Odds Ratio} = ?$ if probability $p = \frac{1}{2}$

Answer: $\text{Odds Ratio} = 1$

2. What is the probability $p = ?$ if $\text{Odds Ratio} = \frac{1}{5}$

Answer: $p = \frac{1}{6}$

czechitas

What is the difference between odds and probability?

$$\text{Odds ratio} = \frac{p}{1-p}$$

p → probability (odds) of success

$1-p$ → probability (odds) of failure

Dice roll example

Quiz: What is the probability and odds ratio of rolling 6?

1	2	3	4	5	6

$$Probability\ p = \frac{1}{6} \neq Odds\ Ratio = \frac{1}{5}$$

Logistic regression coefficients are estimated using Odds ratio + log transformation:

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

Probability can be derived from Odds ratio:

$$p(y) = \big(1 - p(y)\big)e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$$

czechitas

Today’s structure

1. Why we need classification?
   When to use classification?
2. Overview of most common methods
3. How to choose between different methods?
4. Logistic regression: intro
5. Logistic regression: hypothesis testing
6. Logistic regression: prediction

Titanic survival analysis - example

Goal of the analysis:
analyze which people characteristics mattered and by how much whether people would be saved from sinking ship
Key hypothesis:
Does sex, *age* and *ticket class* significantly impact the probability of surviving the accident?

$$\mathrm{odds}(survival) \sim sex + age + class$$

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Data source:
https://CRAN.R-project.org/package=titanic and https://www.kaggle.com/c/titanic/data

Logistic regression model

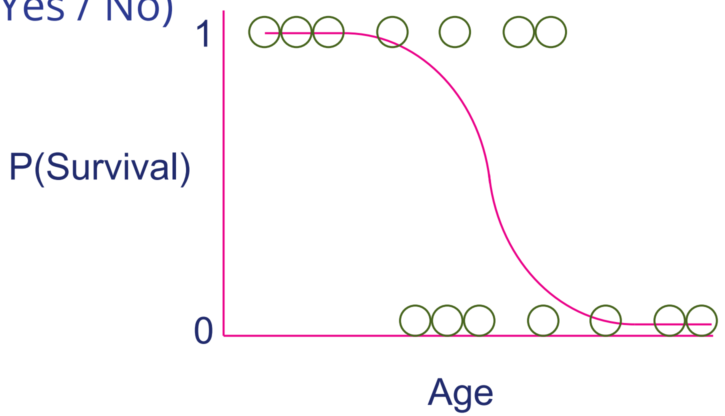
Logistic regression formula

$$\ln\left(\frac{p(survival)}{1 - p(survival)}\right) = \beta_0 + \beta_1 * Age + \beta_2 Sex + \beta_3 pClass$$

Our interest

Probability of survival (between 0 and 1) based on variables

Prediction of survival (Yes / No)



czechitas

Logistic regression model - Quiz

Logistic regression formula

$$\ln\left(\frac{p(survival)}{1 - p(survival)}\right) = \beta_0 + \beta_1 * Age + \beta_2 Sex + \beta_3 pClass$$

Quiz

How would you interpret β_2?

What sign of $\beta_1, \beta_2, \beta_3$ coefficients would you expect? $\frac{+}{-}$

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	

Interpretation of results

Logistic regression formula

$$\ln\left(\frac{p(survival)}{1-p(survival)}\right) = \beta_0 + \beta_1 * Age + \beta_2 Sex + \beta_3 pClass$$

To get impact on odds we need to **exponentiate the coefficients!**

$$odds\ ratio = \frac{p(survival)}{1-p(survival)} = e^{\beta_i}$$

Remember:
Odds ratio < 1 ⇒ relative probability is decreasing
Odds ratio > 1 ⇒ relative probability is increasing

```
                           Logit Regression Results
==============================================================================
Dep. Variable:               Survived   No. Observations:                  620
Model:                          Logit   Df Residuals:                      616
Method:                           MLE   Df Model:                            3
Date:                Tue, 30 May 2023   Pseudo R-squ.:                  0.3386
Time:                        07:30:26   Log-Likelihood:                -273.07
converged:                       True   LL-Null:                       -412.87
Covariance Type:            nonrobust   LLR p-value:                 2.571e-60
==============================================================================
```

	coef	std err	z	P>|z|	[0.025	0.975]
constant	5.3899	0.578	9.330	0.000	4.258	6.522
Age	-0.0412	0.009	-4.725	0.000	-0.058	-0.024
Sex_male	-2.6892	0.231	-11.622	0.000	-3.143	-2.236
Pclass	-1.3266	0.154	-8.618	0.000	-1.628	-1.025

Interpretation of results

1. What is the impact of **one additional year of age** on the odds of survival?

$e^{\beta_1} = e^{-0.041} = 0.96$

Each additional year decreases the odds by 4 %

Each additional year multiplies the probability of survival by 0.96

2. What is the impact of **different sex** on the odds of survival?

$e^{\beta_2} = e^{-2.689} = 0.07$

Males have 93% lower odds of surviving than women

Males has 0.07 times the odds of women to survive

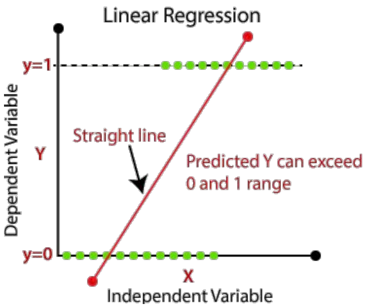
Logit Regression Results

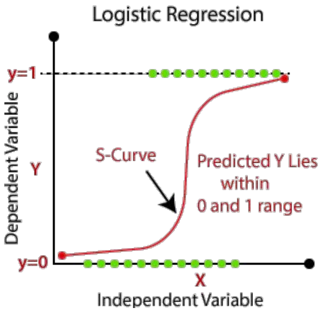
Dep. Variable:	Survived	No. Observations:	620
Model:	Logit	Df Residuals:	616
Method:	MLE	Df Model:	3
Date:	Tue, 30 May 2023	Pseudo R-squ.:	0.3386
Time:	07:30:26	Log-Likelihood:	-273.07
converged:	True	LL-Null:	-412.87
Covariance Type:	nonrobust	LLR p-value:	2.571e-60

	coef	std err	z	P>|z|	[0.025	0.975]
constant	5.3899	0.578	9.330	0.000	4.258	6.522
Age	-0.0412	0.009	-4.725	0.000	-0.058	-0.024
Sex_male	-2.6892	0.231	-11.622	0.000	-3.143	-2.236
Pclass	-1.3266	0.154	-8.618	0.000	-1.628	-1.025

czechitas

Summary





Linear regression is not appropriate model when dependent variables is binary

Logistic regression is used for **classification problems/prediction** of binary outcome

$$Probability\ p = \frac{1}{6} \quad \neq \quad Odds\ Ratio = \frac{1}{5}$$

Logistic regression model use **odds ratio** (not probabilities)

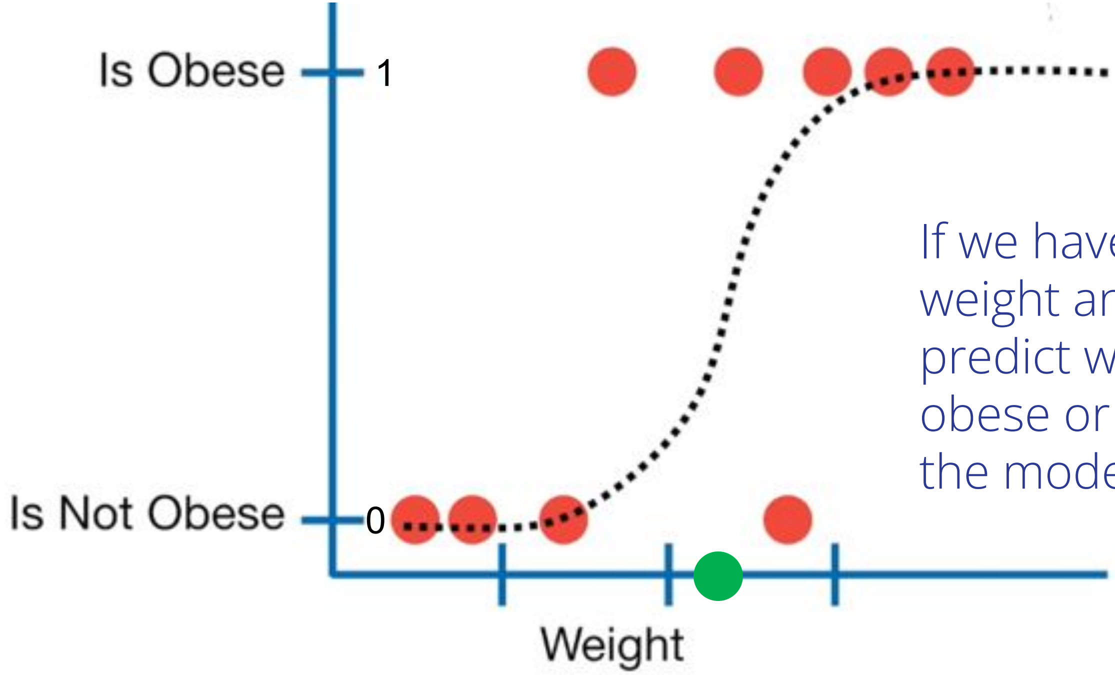
$$odds\ ratio = \frac{p(survival)}{1 - p(survival)} = e^{\beta_i}$$

When testing hypothesis (interpreting coefficients) always remember **to exponentiate the coefficients**

Today’s structure

1. Why we need classification?
   When to use classification?
2. Overview of most common methods
3. How to choose between different methods?
4. Logistic regression: intro
5. Logistic regression: hypothesis testing
6. Logistic regression: prediction

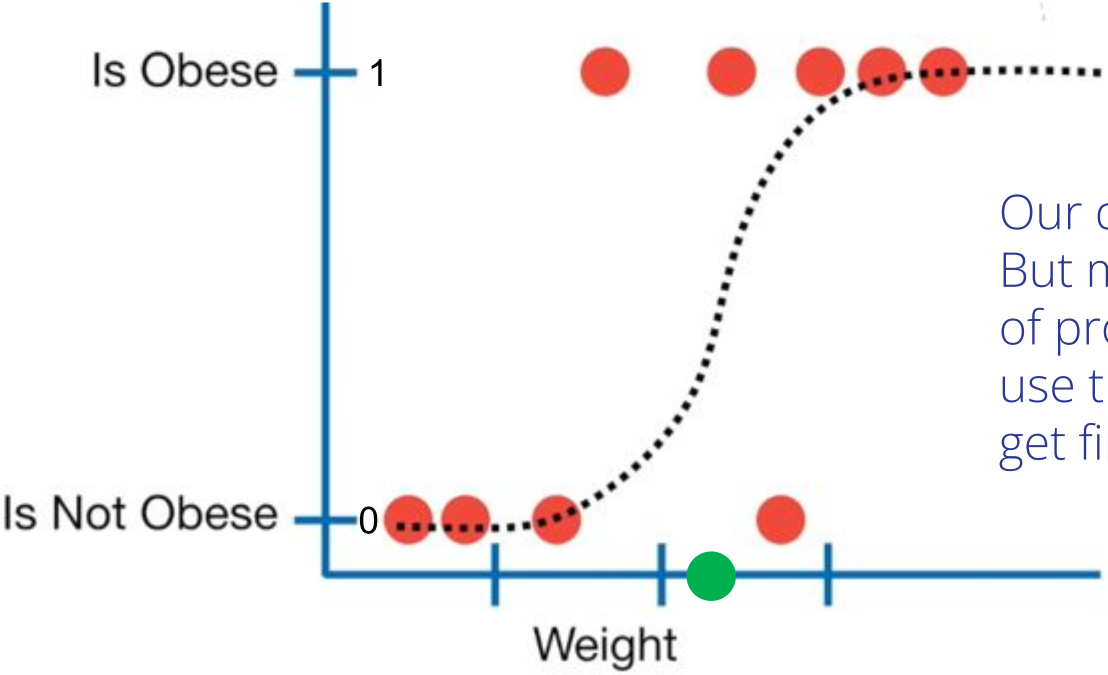
How to use logistic regression model for prediction?



If we have a new person weight and we need to predict whether they are obese or not how do we use the model?

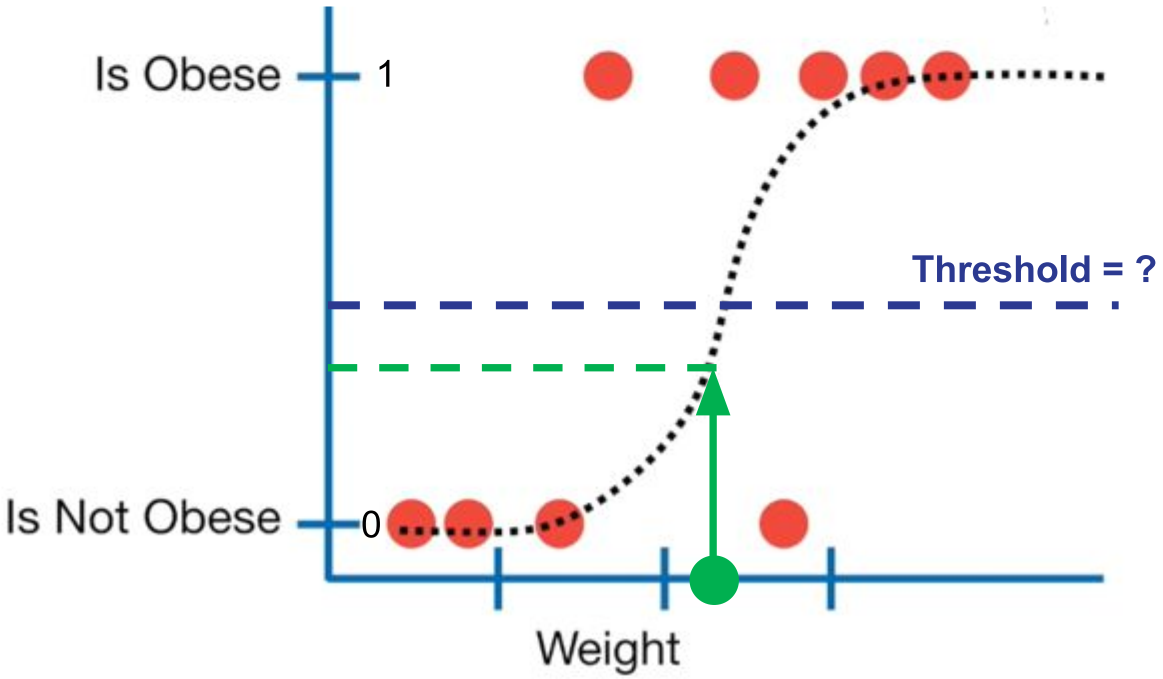
czechitas

How to use logistic regression model for prediction?

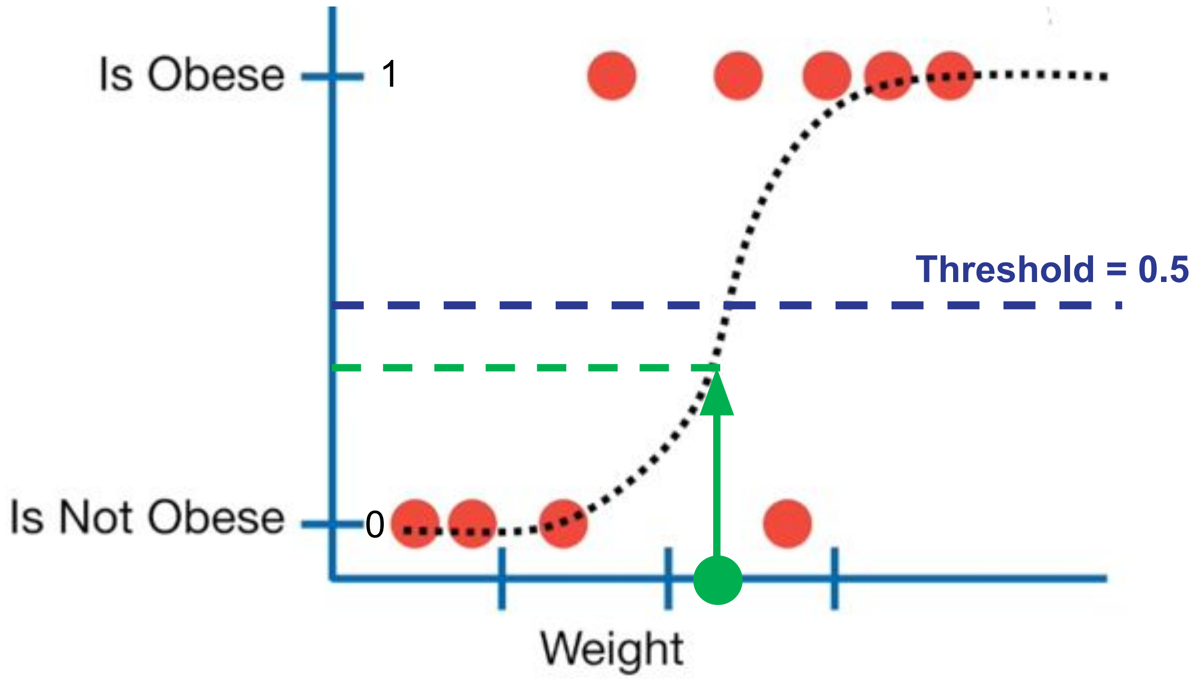


Our data is binary 0 or 1
But model gives us a curve of probabilities how do we use those probabilities to get final binary answer?

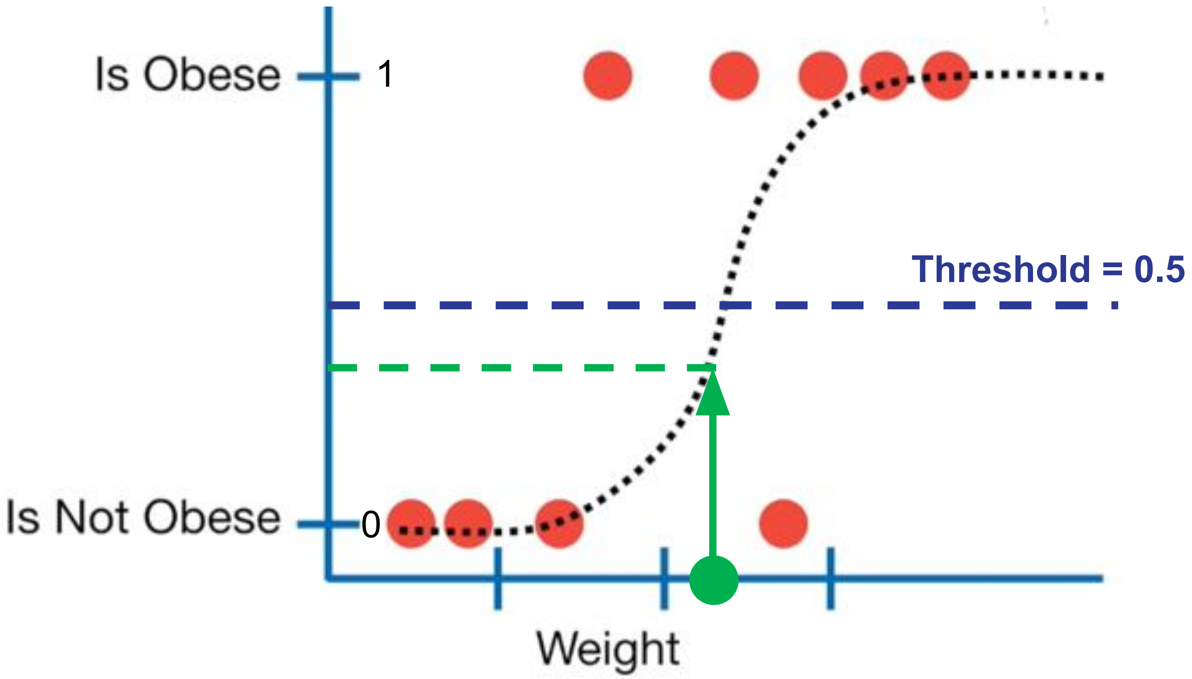
We need to choose probability threshold



We need to choose probability threshold



How to choose optimal threshold given our data and goals of the analysis?



czechitas

To choose threshold we need to know which performance measure we need to optimize

Accuracy or error measures are always related to the risk that business case would suffer if error happens

- One must understand the goal of business case you are working on
- One must consult with business owner/subject matter expert

Performance measures that are relevant for business case

czechitas

CLASSIFICATION PERFORMANCE: measures

	Predicted Yes	Predicted No
Observed Yes	TRUE POSITIVES	FALSE NEGATIVES
Observed No	FALSE POSITIVES	TRUE NEGATIVES

2 types of errors can be made with binary classification

- False Positive – predict Yes when observed is NO (person is obese when in reality person is not, person will be successfully treated by medicine when in reality person will not be successfully treated by medicine)
- False Negative – predict No when observed is Yes (model predicts person is not obese but in reality is, model predicts person will not be cured by new medicine but in reality it is cured with new medicine)

czechitas

CLASSIFICATION PERFORMANCE: measures

	Predicted Yes	Predicted No
Observed Yes	TRUE POSITIVES	FALSE NEGATIVES
Observed No	FALSE POSITIVES	TRUE NEGATIVES

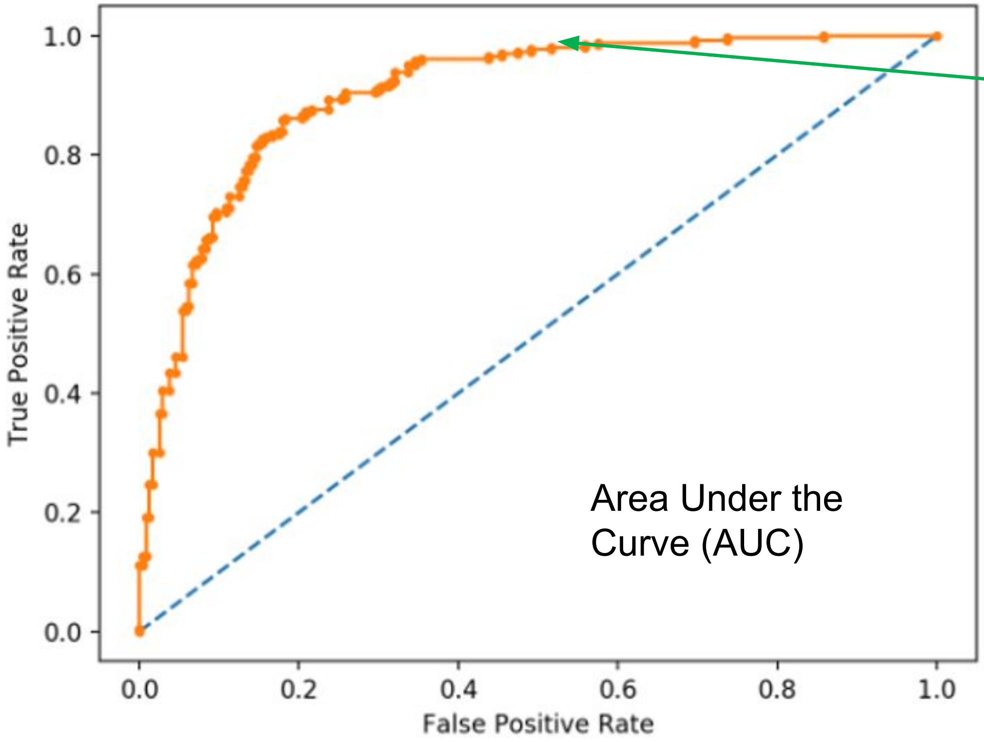
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

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

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN}$$

czechitas

ROC* curve summarize trade-off between true positive and false positive rate



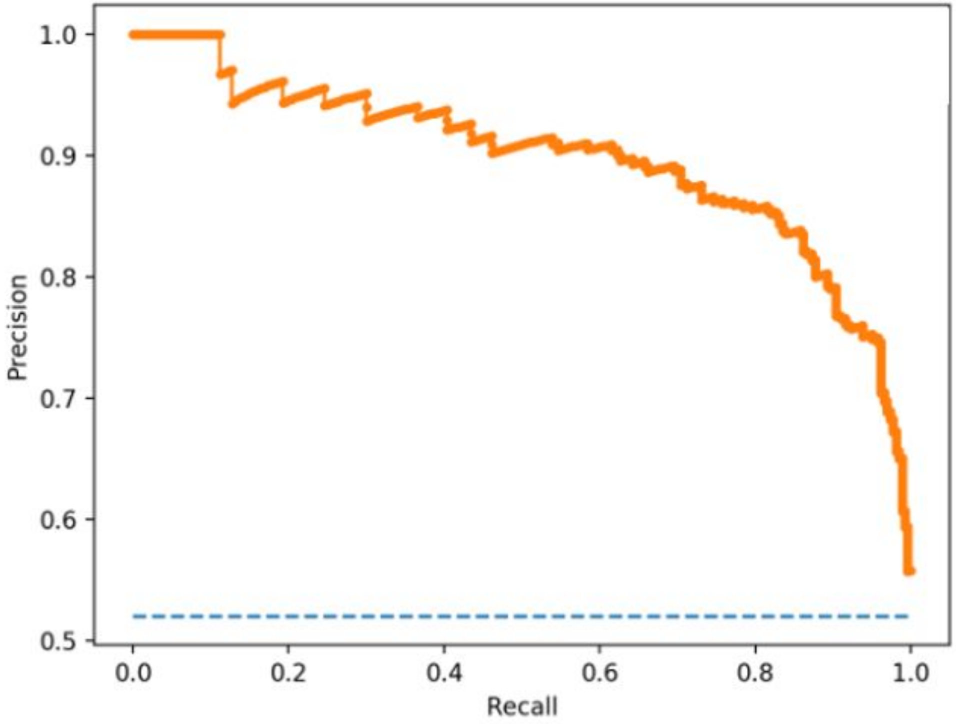
Each point represents threshold

But what if we have unbalanced sample? ROC curves are appropriate if only samples are balanced

*Receiver Operating Characteristic wiki

czechitas

Precision recall curve – summarize trade-off between true positive rate and the positive predictive value

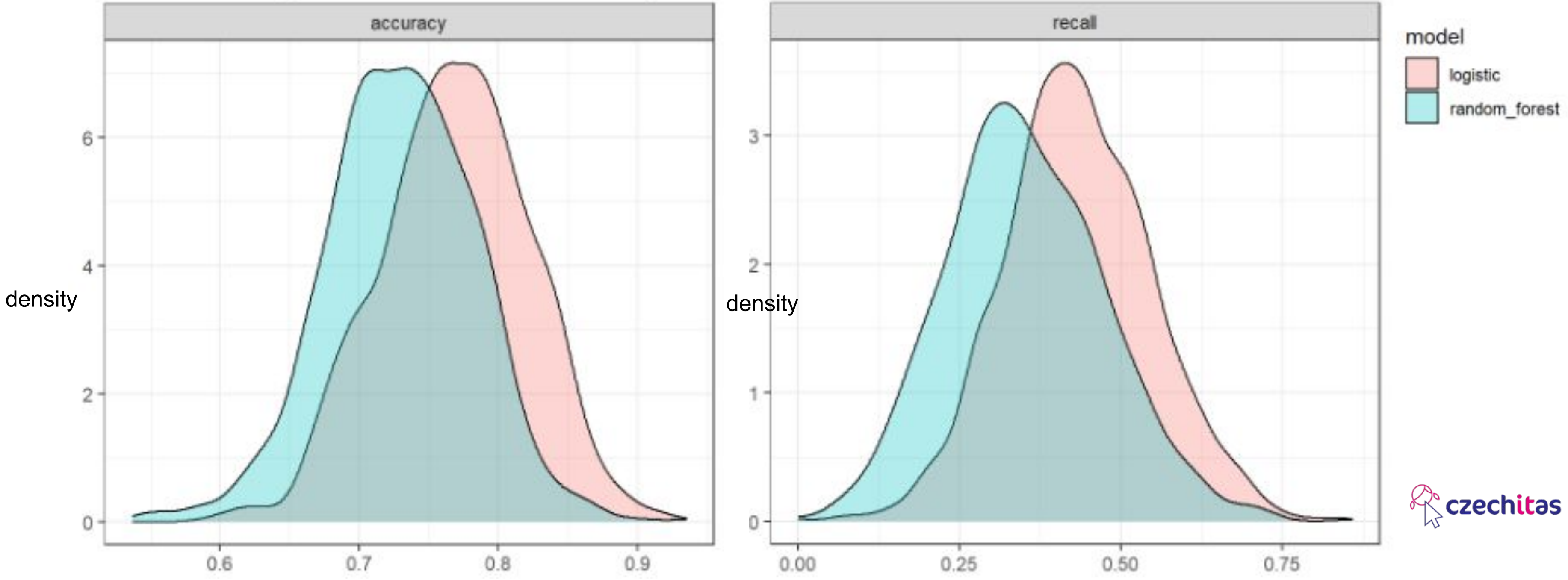


A very useful reading
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-

Cross-validation

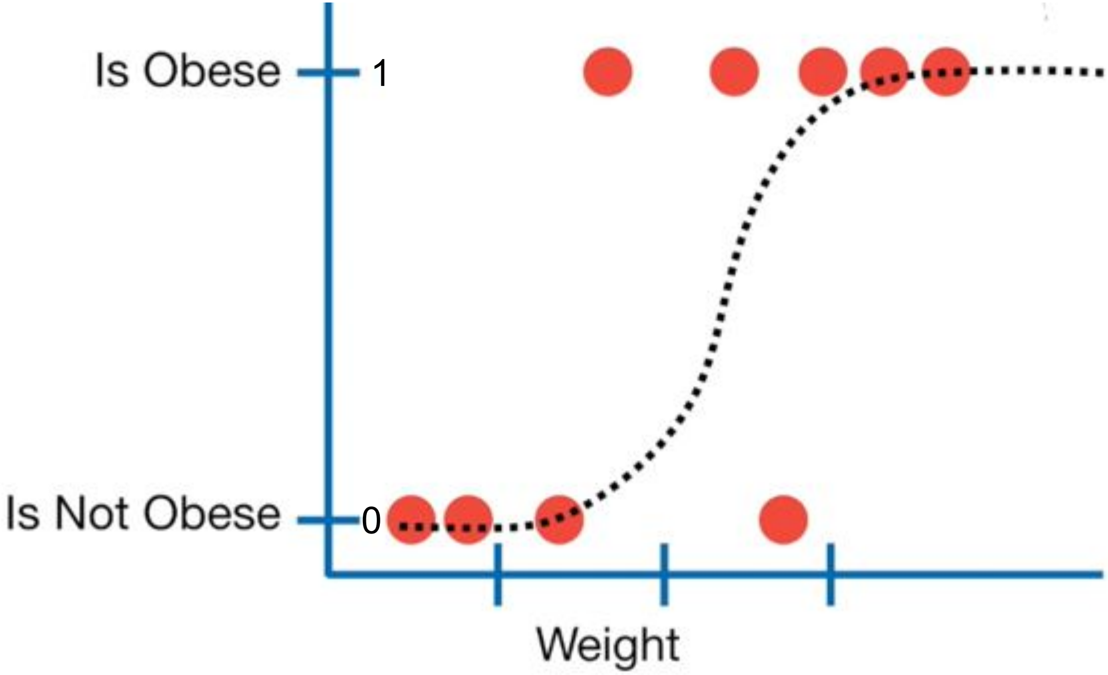
- Divide your data to train and test samples
- Calculate performance metrics that are relevant for your business case
- Repeat multiple times to get distribution of performance errors

1000 random draws of train & test sample combinations

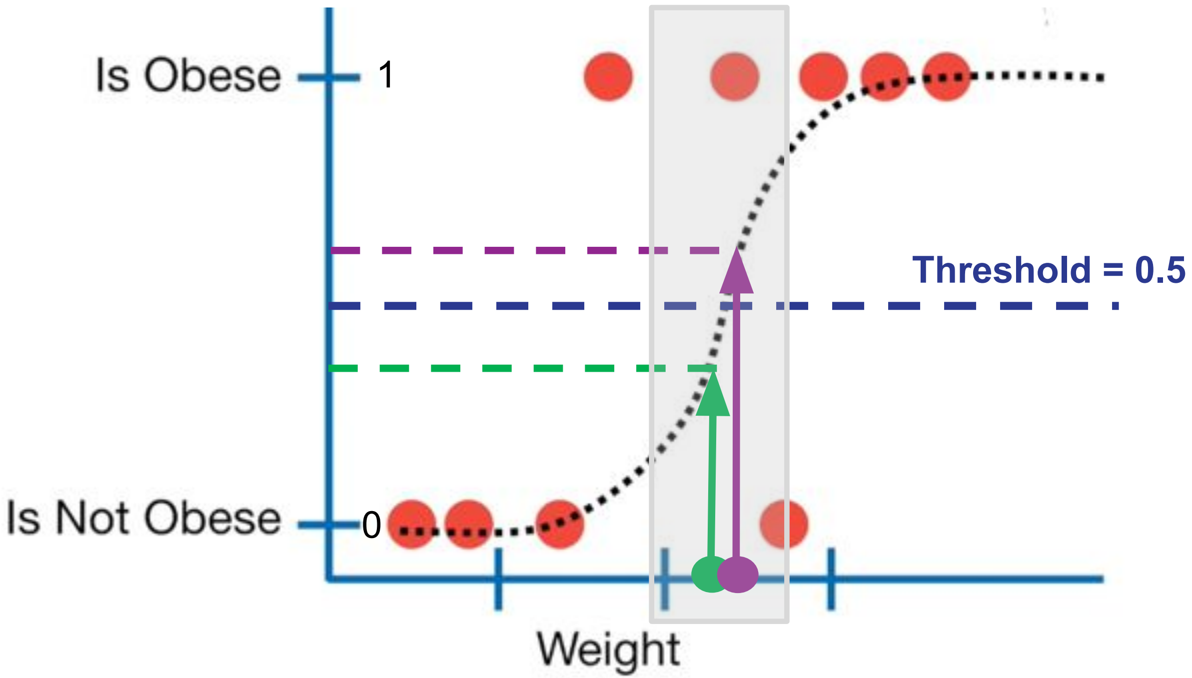


czechitas

Which part of the graph do you think can produce most errors in the prediction?



Grey zone – where most errors happen

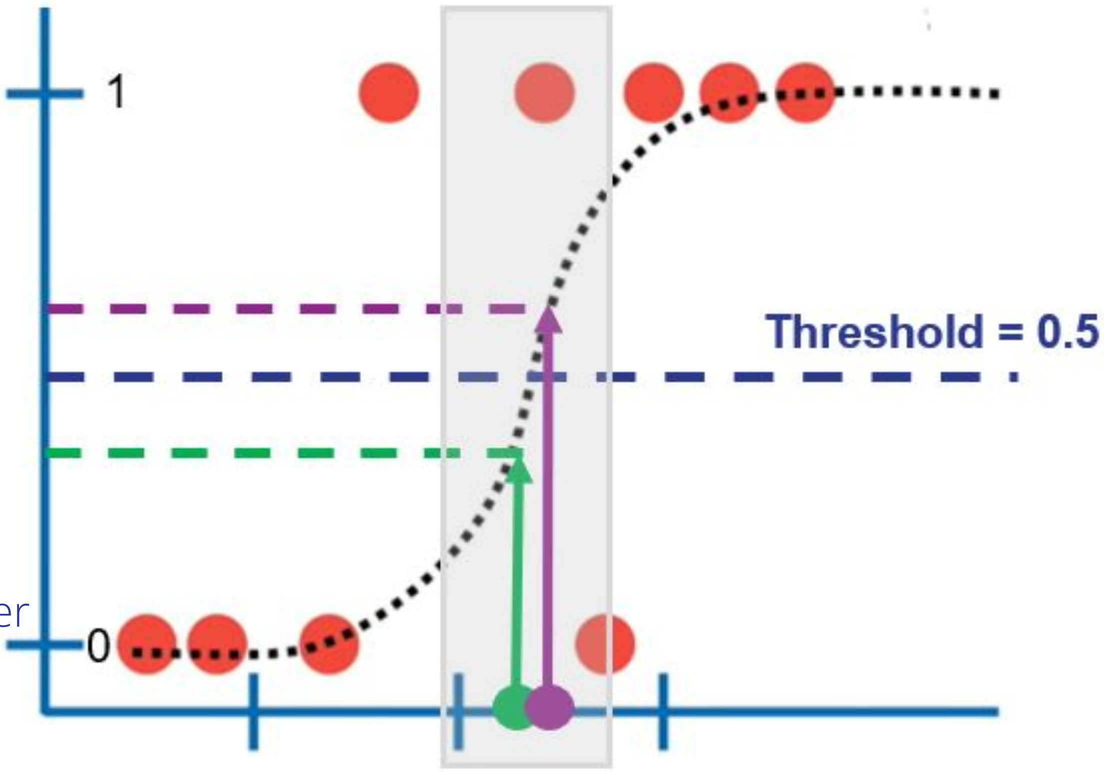


czechitas

Grey zone – should we trust algorithm for all cases?

New drug will treat cancer in patient

New drug will not treat cancer in patient



Imagine we use this model to help decision making – whether a person should be treated or not with expensive new cancer drug which can give side effects

Maybe in grey zone humans should be still making decision about treatments and out of grey zone would be left to automate for algorithms

How much this system reduces the error compared with human judgement?

czechitas

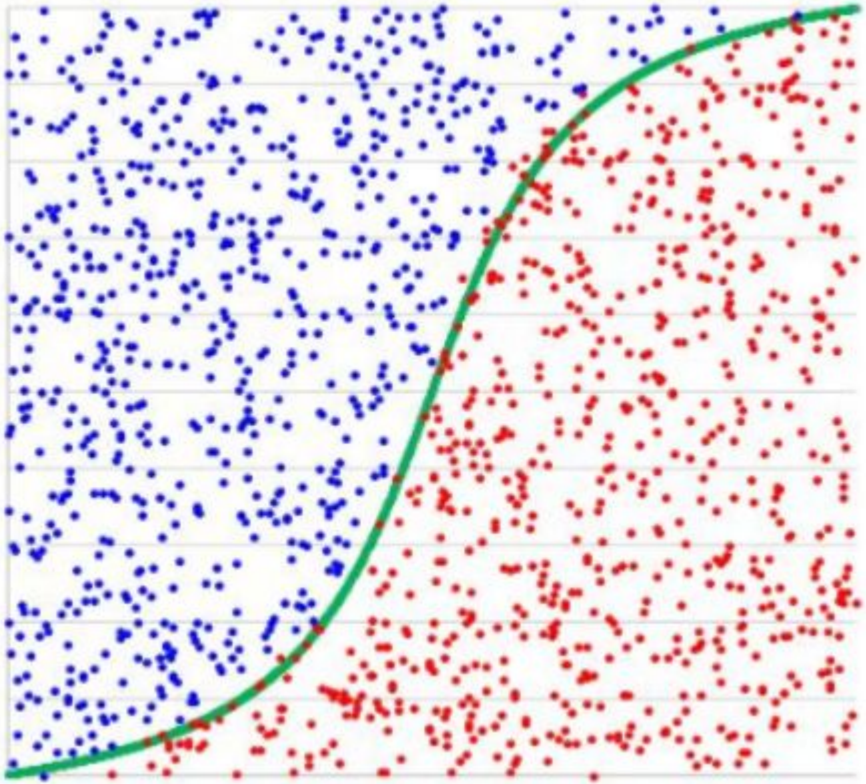
Is logistic model best for prediction?

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

- Logistic model still assumes linear relationship between variables – using more complex function could potentially give better results
- But logistic model would still give easiest explainable results

czechitas

Summary



Logistic regression is a good model for prediction when you want to be sure exactly how each prediction value was calculated (no black box)

You must choose performance measures of logistic regression based on understanding of business case and risks

It is fair to say that some parts of prediction can be risky of high errors (grey zone) and that model should not be used there

What we learned today and what can you expect next time

Linear regression is not appropriate model when dependent variables is binary

There are specific methods used for **classification problems/prediction** of binary outcome

Always check the distribution of observations across classes

When testing hypothesis (interpreting coefficients) always remember **to exponentiate the coefficients**

It is crucial to understand your **business case** to pick correct performance measure for your binary classifier

czechitas

Thank you for your attention.

We are looking forward to the next lecture!