Data Engineering COMP2031/8031

- Topic Coordinator: Dr Mehwish Nasim
- Office: 3.13 Tonsley

Flinders UNIVERSITY

Supervised learning

- Uses a training set to teach models to yield the desired output

Supervised vs. unsupervised learning

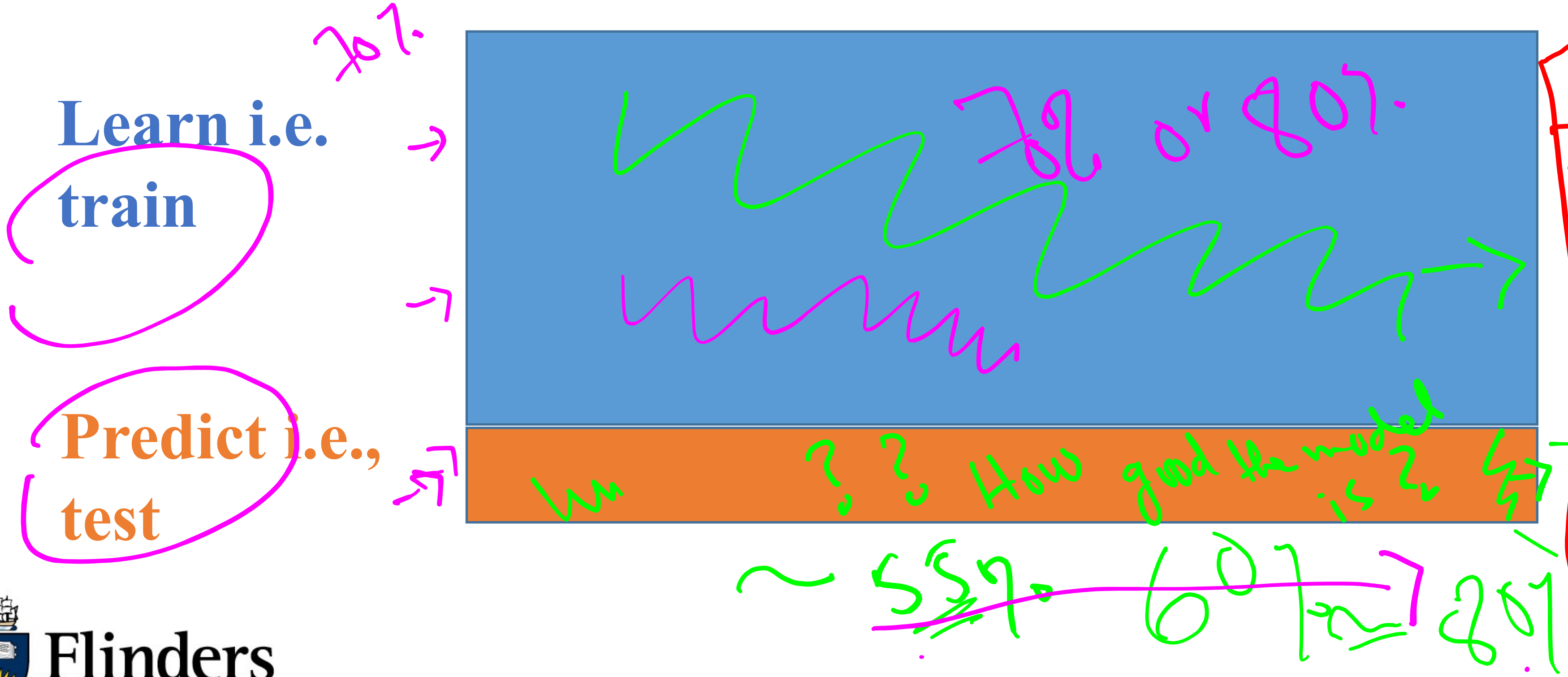
Cross validation 10%

labels → Pass/Fail, Yes/NO

- Labeled datasets!
- supervised learning, the algorithm "learns" from the training dataset by iteratively making predictions on the data.

output

O_2 Hbeat BP



	y	x_1	x_2	x_3
1	Yes	92	–	–
2	NO	98	–	–
3	NO	94	–	–
⋮	Yes	85	–	–
	Yes	70	–	–
	NO	60	–	–
100	??			

?	x_1 x_2 x_3

Regression Vs. Classification

- Regression: understand the relationship between dependent and independent variables. It is commonly used to make projections
- Classification: accurately assign test data into specific categories

Supervised learning algorithms

- Linear Regression

LDA

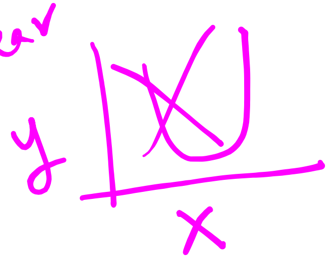
- Logistic Regression (used to solve binary classification problems)

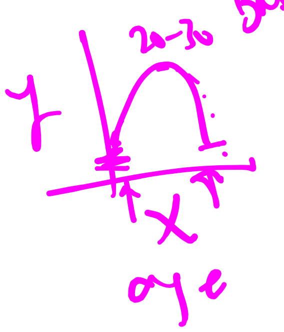
y ~ x linear

y ~ x + linear

- Näive Bayes
- Support Vector Machines (SVM)
- Neural Networks
- K nearest neighbors (kNN)
- Random forests

and more…





Logistic regression

Flinders UNIVERSITY

Logistic regression

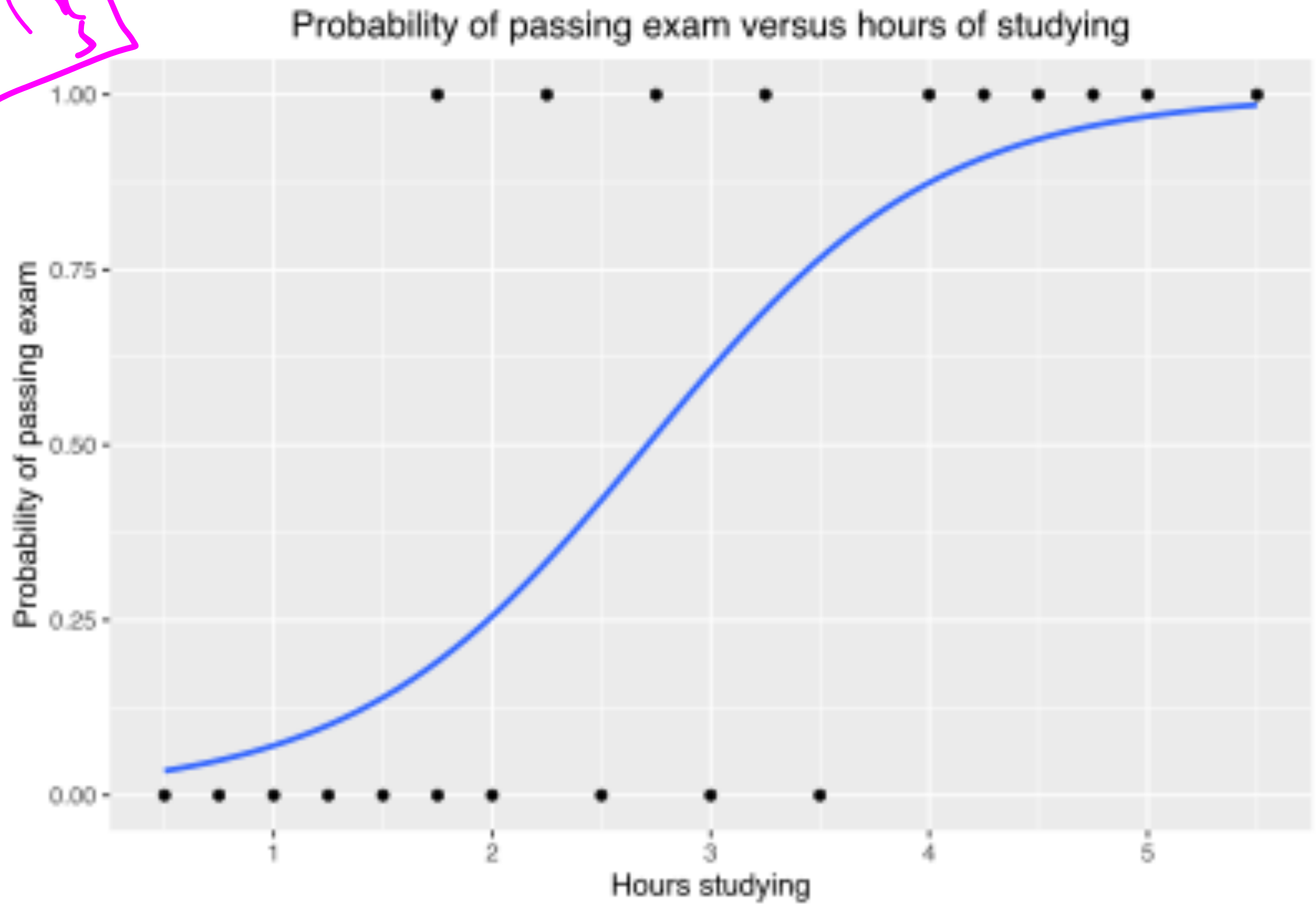
- logistic model is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more **independent variables** ("predictors")

- Independent variables: Independent variables, do not depend on any other variable in the scope of the experiment in question.

some common independent variables are time, space, density, mass, and previous values of some observed value of interest (e.g. human population size) to predict future values (the dependent variable). E.g., $y = 2x+1$,

Here x is an independent variable and y is a dependent variable

Example (wikipedia)



Example

- The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).
- The x variable is called the "explanatory variable", and the y variable is called the "categorical variable" consisting of two categories: "pass" or "fail" corresponding to the categorical values 1 and 0 respectively.

Hours (x_k)	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Example

$(1-p)$

Hours (x_k)	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

100% 0.9 0.3 0.2 0.

1

what is the probability to pass given the students have studied for a certain number of hours?

estimated probability of passing the exam for several values of hours studying

Hours of study (x)	Passing exam		
	Log-odds (t)	Odds (e^t)	Probability (p)
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
$\mu \approx 2.7$	0	1	$\frac{1}{2}$ = 0.50
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

p-value

	Coefficient	Std. Error	*z*-value	*p*-value (Wald)
Intercept (β_0)	−4.1	1.8	−2.3	0.021
Hours (β_1)	1.5	0.6	2.4	0.017

Hours

$p < 0.05$

Fitting the regression line

$$y = \alpha + \beta x$$

Logistic regression uses a method called maximum likelihood estimation to find an equation of the form:

$$\log[\frac{p(X)}{(1-p(X)]} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Flinders UNIVERSITY

Example in R

- #default: Indicates whether or not an individual defaulted.
- #student: Indicates whether or not an individual is a student.
- #balance: Average balance carried by an individual.
- #income: Income of the individual.

Console Terminal Jobs

R 4.1.0 · ~/OneDrive - Flinders/DataEngineering-Material/Quiz4/

```
> summary(model)

Call:
glm(formula = default ~ student + balance + income, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5586  -0.1353  -0.0519  -0.0177   3.7973

Coefficients:
                 Estimate    Std. Error z value
(Intercept) -11.478101194   0.623409555 -18.412
studentYes   -0.493292438   0.285735949  -1.726
balance       0.005988059   0.000293765  20.384
income        0.000007857   0.000009965   0.788
                       Pr(>|z|)
(Intercept) <0.0000000000000002 ***
studentYes               0.0843 .
balance     <0.0000000000000002 ***
income                   0.4304
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2021.1  on 6963  degrees of freedom
Residual deviance: 1065.4  on 6960  degrees of freedom
AIC: 1073.4

Number of Fisher Scoring iterations: 8
```

Flinders UNIVERSITY

Examples of supervised learning

- Image recognition
- Predictive analytics
- Sentiment analysis
- Spam detection

Challenges

- Expertise
- Time
- Human error
- Labeled data

Supervised vs. unsupervised

- Goals
  - Supervised: predict outcomes for **new data**
  - Unsupervised: get insights
- Applications
  - Supervised: spam detection, sentiment analysis, weather forecasting
  - Unsupervised: anomaly detection, recommendation engines
- Complexity
  - Supervised: Simpler
  - Unsupervised: need powerful tools for working with large amounts of unclassified data
- Drawbacks
  - Supervised: Time consuming to train
  - Unsupervised: can have inaccurate results and may require human intervention

Which is best for you?

- **Evaluate your input data:** Is it labeled or unlabeled data? Do you have experts that can support additional labeling?
- **Define your goals:** Do you have a recurring, well-defined problem to solve? Or will the algorithm need to predict new problems?
- **Review your options for algorithms:** Are there algorithms with the same dimensionality you need (number of features, attributes or characteristics)? Can they support your data volume and structure?