# Assignment Project Exam Help COMP9318 Tutorial 2: Classification

https://powcoder.com

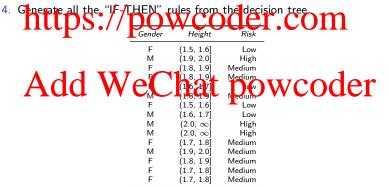
Add WeChat powcoder

Consider the following training dataset and the original decision tree induction algorithm (ID3).

Risk is the class label attribute. The Height values have been already discretized into disjoint ranges.

## ASSIGNATION PAIR TO THE IS thosen a XI BETTA LITTLE P 2. Calculate the information gain if Height is chosen as the test attribute.

- 2. Calculate the information gain if Height is chosen as the test attribute.
- 3. Draw the final decision tree (without any pruning) for the training dataset.



Consider applying the SPRINT algorithm on the following training dataset

	0 -			J	
	23	family	High	]	
Assignme	nt³ F	ratec	High High	xam	Help
	68	family	Low		1
	32	truck	Low		
1	, 20	family	→ High		
Answeight Distribution					

CarType

Risk

- 1. Write down the attribute lists for attribute Age and CarType, respectively.
- 2. Assume the first split criterion is Age < 27.5. Write down the attribute list for the left had node (i.e. corresponding to the partition whose Age (c.f.)
- 3. Assume that the two attribute lists for the root node are stored in relational tables name  $AL\_Age$  and  $AL\_CarType$ , respectively. We can in fact generate the attribute lists for the child nodes using standard SQL statements. Write down the SQL statements which will generate the attribute lists for the left child node for the split criterion Age < 27.5.
- 4. Write down the final decision tree constructed by the SPRINT algorithm.

Consider a (simplified) email classification example. Assume the training dataset contains 1000 emails in total, 100 of which are spams.

1. Calculate the class prior probability distribution. How would you classify a SSISIMPHETIT PROJECT Exam Help 2. Affilend of you suggests that whether the email contains a \$ char is a

Afriend of you suggests that whether the email contains a \$ char is a good feature to detect spam emails. You look into the training dataset and obtain the following statistics (\$ means emails containing a \$ and \$ are those ports on thin in a ny \$ } ,

antity south powcoder.com

Described (naive Bees Classification and ill Victor of the Control of the Control

"evidence". How would this classifier predict the class label for a new incoming email that contains a \$ character?

3. Another friend of you suggest looking into the feature of whether the email's length is longer than a fixed threshold (e.g., 500 bytes). You obtain the following results (this feature denoted as  $L(\bar{L})$ ).

SPAM 40 60

https://powcoder.com How would a naive Bayes classifier predict the class label for a new incoming email that contains a \$ character and is shorter than the threshold?

### Add WeChat powcoder

Based on the data in the following table,

1. estimate a Bernoulli Naive Bayes classifer (using the add-one smoothing)

#### Signature of the Especial Examem Help 3. estimate a multinomial Naive Bayes classifier (using the add-one smoothing)

4. apply the classifier to the test document
You do not be a sylinal plannave of a doct tree following the

Λ	class = China?		
training e	1	Vaire Taiwa (U) W C	DAGI
_	2	Macao Taiwan Shanghai	Yes
	3	Japan Sapporo	No
	4	Sapporo Osaka Taiwan	No
test set	5	Taiwan Taiwan Taiwan Sapporo Bangkok	?

- 1. First, we randomly obtained 47 training examples among which we have 22 negative instances (denoted as "-"), and 25 positive instances (denoted as "-"), and 25 positive instances (denoted as "-") training examples among which we have 22 negative instances (denoted as "-") and 25 positive instances (denoted as "-") training examples among which we have 22 negative instances (denoted as "-") and 25 positive instances (denoted as "-") training examples among which we have 22 negative instances (denoted as "-") and 25 positive instances (denoted as "-") training examples among which we have 25 negative instances (denoted as "-") and 25 positive instances (denoted as "-") training examples among which we have 25 negative instances (denoted as "-") and 25 positive instance belongs to the positive class?
- 2. We then identify a feature x, and rearrange the 47 training examples based on their x wives the identities shown in the table below er

Assignment Pro	oject Exam Help
3	+   6
https://pow	vcoder.com

count

### Add We Chat powcoder

For each of the group of training examples with the same x value, compute its probability  $p_i$  and  $logit(p) := log \frac{p}{1-p}$ .

- 3. What is your estimate of the probability that a novel test instance belongs to the positive class if its x value is 1?
- 4. We can run a linear regression on the (x, logit) pairs from each group. Will this be the same as what Logistic Regression does?

- ▶ Represent the vectors in the non-orthogonal bases  $\mathcal{B} = \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}$ .
- Let  $Z_p$  be a vector Z represented in the polar coordinate:  $(p, \theta)$ . What if white  $Z_p$  is a vector Z represented in the polar coordinate:  $(p, \theta)$ . What if  $Z_p$  is  $Z_p$  is  $Z_p$  in  $Z_p$  in
- Can you construct a matrix M such that its impact on vectors represented in polar coordinates exhibit "linearality"? i.e., M(x+y) = Mx + My? Add WeCnat powcoder

 $\mathbf{X}_{n\times d} = \begin{pmatrix} \mathbf{o}_1 \\ \mathbf{o}_2 \\ \vdots \\ \mathbf{o}_d \end{pmatrix}$ . Now we consider a linear projection  $\mathbf{A}_{d\times m}$  of all the points to a m-dimensional space (m) which  $\mathbf{a}_{d\times m}$  is mapled on the points to  $\pi(o_i) = \mathbf{A}^{\top} \mathbf{o}_i$ .

 $\begin{array}{c} \text{Computer } \mathit{r} \coloneqq \frac{\|\pi(o_i)\|^2}{\|o_i\|^2}. \text{ Can you guess what will be the maximum and} \\ \text{Maximum values of } e Chat powcoder \\ \end{array}$