COMP9318 Tutorial 2: Classification

Wei Wang @ UNSW

Q1 I

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.

1. Calculate the information gain if *Gender* 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.
4. Generate all the "IF-THEN" rules from the decision tree.

Gender	*Height*	*Risk*
F	(1.5, 1.6]	Low
M	(1.9, 2.0]	High
F	(1.8, 1.9]	Medium
F	(1.8, 1.9]	Medium
F	(1.6, 1.7]	Low
M	(1.8, 1.9]	Medium
F	(1.5, 1.6]	Low
M	(1.6, 1.7]	Low
M	(2.0, ∞]	High
M	(2.0, ∞]	High
F	(1.7, 1.8]	Medium
M	(1.9, 2.0]	Medium
F	(1.8, 1.9]	Medium
F	(1.7, 1.8]	Medium
F	(1.7, 1.8]	Medium

Q2 I

Consider applying the SPRINT algorithm on the following training dataset

Age	*CarType*	*Risk*
23	family	High
17	sports	High
43	sports	High
68	family	Low
32	truck	Low
20	family	High

Answer the following questions:

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 lists for the left child node (i.e., corresponding to the partition whose $Age < 27.5$).
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.

Q3 I

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 new incoming email?
2. A friend 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 $\bar{\$}$ are those not containing any \$).

Class	\$	$\bar{\$}$
SPAM	91	9
NOSPAM	63	837

Describe the (naive) Bayes Classifier you can build on this new piece of "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 ($\bar{L}$)).

Q3 II

Class	L	$\bar{L}$
SPAM	40	60
NOSPAM	400	500

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?

Q4 I

Based on the data in the following table,

1. estimate a Bernoulli Naive Bayes classifer (using the add-one smoothing)
2. apply the classifier to the test document
3. estimate a multinomial Naive Bayes classifier (using the add-one smoothing)
4. apply the classifier to the test document

You do not need to estimate parameters that you don't need for classifying the test document.

	docID	words in document	class = China?
training set	1	Taipei Taiwan	Yes
	2	Macao Taiwan Shanghai	Yes
	3	Japan Sapporo	No
	4	Sapporo Osaka Taiwan	No
test set	5	Taiwan Taiwan Taiwan Sapporo Bangkok	?

Q5 I

Consider a binary classification problem.

1. First, we randomly obtained 47 training examples among which we have 22 negative instances (denoted as "-"), and 25 positive instances (denoted as "+").
   What is your estimate of the probability that a novel test instance belongs to the positive class?
2. We then identify a feature x, and rearrange the 47 training examples based on their x values. The result is shown in the table below.

Q5 II

x	y	count
1	-	6
1	+	2
2	-	5
2	+	2
3	-	7
3	+	6
4	-	3
4	+	7
5	-	1
5	+	8

Table: Training Data

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?

Q6 I

Consider two-dimensional vectors $\mathbf{A} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ and $\mathbf{C} = \mathbf{A} + \mathbf{B}$.

- Represent the vectors in the non-orthogonal bases $\mathcal{B} = \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}$.
- Let $\mathbf{Z}_p$ be a vector $\mathbf{Z}$ represented in the polar coordinate: (ρ, θ). What if we still do $\mathbf{Z}_p = \mathbf{A}_p + \mathbf{B}_p$ in the old "linear" way? Will $\mathbf{Z}_p$ be the same as $\mathbf{C}_p$?
- Can you construct a matrix $\mathbf{M}$ such that its impact on vectors represented in polar coordinates exhibit "linearality"? i.e., $\mathbf{M}(\mathbf{x} + \mathbf{y}) = \mathbf{M}\mathbf{x} + \mathbf{M}\mathbf{y}$?

Q7 I

Consider a set of d-dimensional points arranged in a *data matrix* $\mathbf{X}_{n\times d} = \begin{pmatrix} \mathbf{o}_1 \\ \mathbf{o}_2 \\ \vdots \\ \mathbf{o}_n \end{pmatrix}$. Now we consider a linear projection $\mathbf{A}_{d\times m}$ of all the points to a m-dimensional space ($m < d$). Specifically each $\mathbf{o}_i$ is mapped to a new vector $\pi(o_i) = \mathbf{A}^\top \mathbf{o}_i$.

- Computer $r := \frac{\|\pi(\mathbf{o}_i)\|^2}{\|\mathbf{o}_i\|^2}$. Can you guess what will be the maximum and minimum values of r?