2. (10%) Suppose that only one of decision tree induction, k-nearest neighbors, and naïve Bayesian classifier can be applied on a data set. Which classification algorithm is preferred and why? Make necessary assumptions to justify your answer. 假設情況 運算效率：naïve Bayesian > decision tree > k-nearest

學習結果解讀：decision tree

3. (10%) The m-estimate of condition probability in naïve Bayesian classifier is represented as $P(x_i|y_j) = (r_{ij}+mp)/(r_j+m)$, where m is the equivalent sample size, and $p > 0$ is a user-specified parameter. What will be the problem in applying this m-estimate? Please justify your answer. 自行定義？$r_{ij} = r_j = 0$，$P(x_i|y_i) = \frac{mp}{m} = p$

p 值必須根據 accribute 自身狀態設定

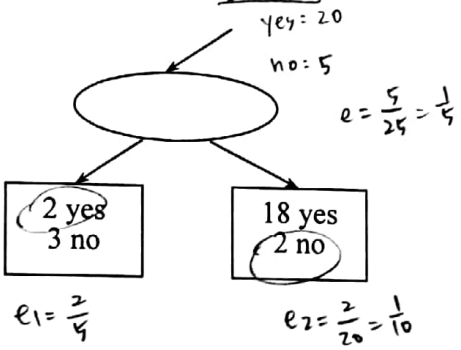
$\sum_i P(x_i|y_i) = k_i \cdot p = 1 \Rightarrow p = \frac{1}{k_i}$ ⇒ 正確！

0.235 not pruned

4. (10%) Let a subtree of a decision tree be as given below. Use the estimate of generalization error rate calculated by $\dfrac{e+\frac{z^2}{2N}+z\sqrt{\frac{e(1-e)}{N}+\frac{z^2}{4N^2}}}{1+\frac{z^2}{N}}$ for z = 0.69 to determine whether this subtree should be pruned.

$P(x_i|y_i) = \dfrac{r_{ij}+\frac{m}{k_i}}{r_j+m}$

0.261



$e = \frac{5}{25} = \frac{1}{5}$

$e_1 = \frac{2}{5}$　　$e_2 = \frac{2}{20} = \frac{1}{10}$

5. (10%) The testing results of two classification methods A and B on a data set by five-fold cross validation are summarized in the following table. We are interested in knowing whether the accuracies of the two algorithms are significantly different. Formulate the null hypothesis and calculate the test statistic by the independent-sample approach.

$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}$ = 20.689

Iteration	Number of testing instances	Number of correct instances for algorithm A	Number of correct instances for algorithm B
1	77	49 $\frac{49}{77} = 0.636$	46 $\frac{46}{77} = 0.597$
2	77	45 0.584	56 0.727
3	76	59 0.776	64 0.842
4	76	64 0.842	68 0.895
5	76	49 0.645	52 0.684

Hypothesis: $P_A - P_B = 0$ (accuracy)

先算各自的 accuracy，再求 $\bar{x}_A$, $\bar{x}_B$, s_A^2, s_B^2

$t = \dfrac{\bar{P}_A - \bar{P}_B}{\sqrt{\frac{s_A^2}{k} + \frac{s_B^2}{k}}}$　　accuracy ↑ sample mean

$\bar{x}_A = 0.697$，$s_A^2 = 0.0116$

$\bar{x}_B = 0.749$，$s_B^2 = 0.0144$

$t = \dfrac{0.697 - 0.749}{\sqrt{\frac{0.0116}{5} + \frac{0.0144}{5}}} = \dfrac{-0.052}{0.072} = -0.72$

Solution of Data Mining Midterm exam

1. (a) Since GainRatio(magnitude) = 0.1466/1.5656 = 0.0937 and GainRatio(visibility) = 0.0200/0.9710 = 0.0206, attribute "magnitude" is a better choice.

   (b) Let x represent instance <cloudy, tail, medium, good>. Since p(antomatic|x) = 9/15×4/12×6/11×5/12×7/11 = 0.0289 and p(manual|x) = 6/15×3/9×3/8×2/9×4/8 = 0.0056, the predicted class value is 'automatic'.

   (c) Let A, B, and C represent outlook, wind, and landing, respectively. Then H(A) = 1.5849, H(B) = 0.9970, H(C) = 0.9710, H(A, C) = 2.4729, and H(B, C) = 1.9329, and hence U(A, C) = 0.0650 and U(B, C) = 0.0354. Attribute "outlook" is more useful.

   (d) (1) The width of an interval is (3.05 – 0.05)/5 = 0.6, and hence the four splitting points are 0.65, 1.25, 1.85, and 2.45.

   (2) The number of observations in an interval is 15/5 = 3, and hence the four splitting points are 0.34, 0.81, 1.08, and 1.95.

   (3) The entropy values of three possible binary splitting points are entropy(0.3) = 0.7718, entropy(1.16) = 0.7219, and entropy(2.8) = 0.9195, and hence the best binary branching value is 1.16.

   (e) Let A, B, and C represent outlook, wind, and landing, respectively. Since we have H(A, B) = 2.4729 and U(A, B) = 0.0843, goodness({A, B}) = $(0.0650+0.0354)/\sqrt{1+0.0843+0.0843+1} = 0.0681$.

2. Depends on the categories of attributes, the data size, and the interpretation of learning results.

3. If p is not properly chosen, the value calculated by the expression will not be a probability.

4. Since e_{upper}(before branching) = 0.2606 and e_{upper}(after branching) = 0.2355, the subtree should not be pruned.

5. $\bar{x}_A = 0.6968$, $\bar{x}_B = 0.7491$, $s_A^2 = 0.011607$, $s_B^2 = 0.014390$, and t = -0.7261.

106學年 Data Mining Midterm Exam

1. Let the data shown in the following table be the instances that record whether the landing of a space shuttle is manual or automatic. We are interested in identifying the characteristics of shuttle landing.

A=9
N=6

outlook	wind	magnitude	visibility	landing	
sunny	head	low	bad	automatic	0.36
sunny	head	medium	good	automatic	0.93
sunny	head	medium	good	manual	1.22
sunny	head	low	good	manual	0.05
sunny	tail	low	bad	manual	0.28
cloudy	tail	strong	bad	manual	1.66
cloudy	tail	low	bad	automatic	0.69
cloudy	tail	medium	good	automatic	1.10
cloudy	head	strong	good	manual	2.24
cloudy	head	low	good	automatic	0.58
rainy	head	medium	good	automatic	1.06
rainy	tail	low	bad	automatic	0.32
rainy	tail	medium	good	automatic	0.98
rainy	head	strong	bad	manual	2.55
rainy	tail	strong	good	automatic	3.05

(a) (10%) Determine whether attribute "magnitude" or "visibility" is more appropriate for branching at the root of a decision tree by the gain ratio.

magnitude, 0.298

(b) (10%) Use the naïve Bayesian classifier with Laplace's estimate to classify a new instance <cloudy, tail, medium, good>.

P(A|x) = 0.0095, P(M|x) = 0.0045

(c) (10%) Use the symmetric uncertain to identify whether "outlook" or "wind" is more useful in determining the value of "landing".

$U(A,B) = 2 \times \frac{H(A)+H(B)-H(A,B)}{H(A)+H(B)}$

outlook → 0.065 / wind → 0.036

(d) Let the visibility of the 15 instances be 0.36, 0.93, 1.22, 0.05, 0.28, 1.66, 0.69, 1.10, 2.24, 0.58, 1.06, 0.32, 0.98, 2.55, and 3.05 kilometers, respectively, n=5

(1) (5%) Give the four splitting points for the equal-width discretization. 0.65 / 1.25 / 1.85 / 2.45

(2) (5%) Give the four splitting points for the equal-frequency discretization. 0.34 / 0.81 / 1.08 / 1.95

(3) (10%) Find the best binary splitting point in performing the entropy-based discretization for the 15 instances. 1.16

(e) (10%) Calculated the goodness of attribute subset {outlook, wind}. 0.047

$$\frac{\sum_{B \in S} U(A,C)}{\sum_{A \in S} \sum_{B \in S} U(A,B)}$$