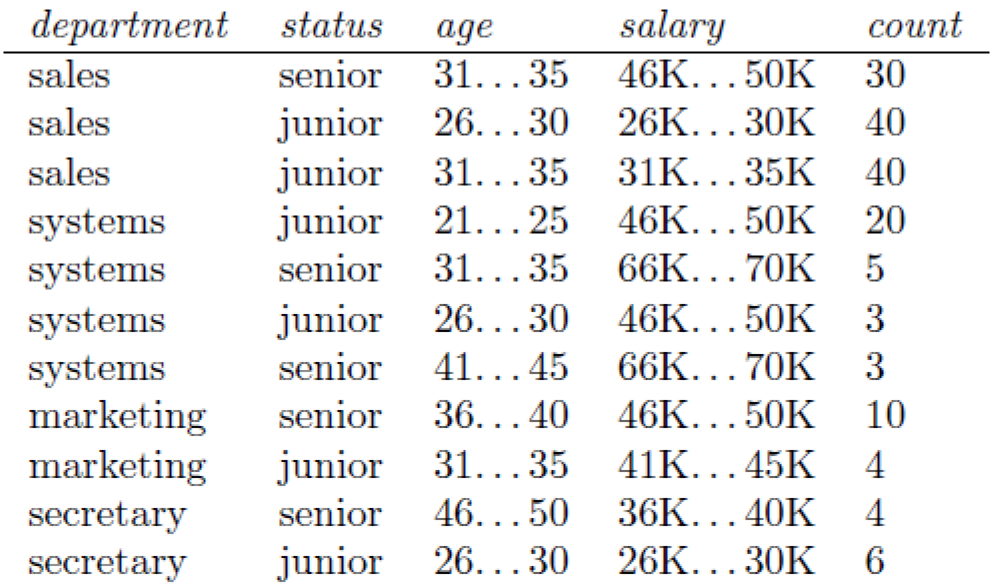
5369U Introduction to data mining – Spring 2010

Assignment 2 (100 points)

Due: 11:55pm, Mar 4, 10 (submit to TRACS)

1. (20 points) This question is about Naïve Bayes Classification. In your answer, show the calculation steps properly.

The following table consists of training data from an employee database. The data have been generalized. For example, "31 … 35" for *age* represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for *department*, *status*, *age*, and *salary* given in that row.



department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let *status* be the class label attribute. Given a data tuple having the values "systems", "26. . . 30", and "46K … 50K" for the attributes *department*, *age*, and *salary*, respectively, what would a naïve Bayesian classification of the *status* for the tuple be?

2. (15 points) This question is about data standardization. In your answer, show the calculation steps properly.

Given the following measurements for the variable age:

18, 22, 25, 42, 28, 43, 33, 35, 56, 28;

standardize the variable by the following:
(a) Compute the mean absolute deviation of age.
(b) Compute the z-score for the first four measurements.

3. (15 points) This question is about pairwise distance calculation. In your answer, show the calculation steps properly.

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
(a) Compute the *Euclidean distance* between the two objects.
(b) Compute the *Manhattan distance* between the two objects.
(c) Compute the *Minkowski distance* between the two objects, using $q = 3$.

4. (50 points) In this question, you are asked to program a simple *k*-means (any variant of your own choice) clustering algorithm, *kmeans*, on 2-dimensional numerical data.

You have the flexibility to choose a programming language. Your program should be executed as follows:

```
kmeans k input.txt
```

where input parameter $k > 1$ is an integer, specifying the number of clusters. *input.txt* is an input file containing many 2-dimensional data points in the following format,

274	119
317	144
267	164
233	137
272	99
297	116
268	142
522	286
468	308
441	263

Your program should output a txt file called output.txt, in the following format:

```
274     119     1
317     144     1
267     164     1
233     137     1
272     99      1
297     116     1
268     142     1
522     286     2
468     308     2
441     263     2
```

In output.txt, 1 and 2 are cluster labels. Each data point should be labeled using one of the labels from 1 to *k*. In the above example, there are 10 data points and *k* = 2.

For your convenience, a Windows data generator, gen.exe, is posted. You can use it to generate and visualize 2-dimensional data as well as clustering results.

Submission:

> For Questions 1, 2, and 3, use MS Word or Excel to type your answers. For Question 4, submit your source code, compiled executable, and a short note describing in what language and under what environment you implemented your program, and how to execute it. Zip everything in a single file and submit to TRACS.