I would like to describe in Chinese. But after months of training, I find English more comfortable.

This time let's work together with heart and soul, to solve this problem and win the happiness of Emirati. You are the best.

1. Load training data chunk 1–5 into MATLAB. After this step, they should be cell array.

2. Missing categorical values. They should be represented as "NaN" in MATLAB. Replace missing **categorical** values with **0**.

3. Missing numeric values. They should be represented as NaN in MATLAB. Replace them with mean of that column. (I suppose it would be more clear to describe in Algorithms. The rest is done with algorithms.)

```
for each column do
    Check if there is $NaN$.
    Add a new row. If $Column(j)\ has\ NaN$, $row(i,j) = 1$. Else mark 0.
end for
```

Figure 1: Missing values

```
for Missing categorical values do
    Replace $NaN$ in missing categorical values with 0.
    if elements in categorical values$\equiv NaN$
    $NaN --> 0$
end for
```

Figure 2: Missing categorical values

```
Compute each column's mean. $mean(Column)$.
Replace $NaN$ in numerical columns with each column's mean.
```

Figure 3: Missing numeric values

I will describe this in Chinese, in a separate file.

Figure 4: Recode categorical values

```
use $abs(mas())$ to calculate each column's max value.
Normalize each column. For each value i in column j, replace value i with
    $i/max\ value\ in\ that\ column$. Namely, scale the value to [0–1].
```

Figure 5: Normalize

Sample code is given as follows. I will describe in a separate file in Chinese.

```
matlab> SPECTF = csvread('SPECTF.train');
 % read a csv file
matlab> labels = SPECTF(:, 1);
 % labels from the 1st column
matlab> features = SPECTF(:, 2:end);
matlab> features_sparse = sparse(features);
% features must be in a sparse matrix
matlab> libsvmwrite('SPECTFlibsvm.train', labels, features_sparse);
```

Figure 6: Write to csv